Journal of Computer Technology and Software

ISSN:2998-2383

Vol. 4, No. 2, 2025

Cross-Scale Attention and Multi-Layer Feature Fusion YOLOv8 for Skin Disease Target Detection in Medical Images

Ting Xu[1], Yanlin Xiang[2], Junliang Du[3], Hanchao Zhang[4]

[1]University of Massachusetts Boston, Boston, USA

[2]University of Houston, Houston, USA

[3]Shanghai Jiao Tong University, Shanghai, China

[4]New York University, New York , USA

*Corresponding Author: Yanlin Xiang, yxiang7@cougarnet.uh.edu

Abstract: With the continuous development of deep learning technology, skin disease target detection has been increasingly widely used in medical image analysis. This paper proposes a skin disease target detection method based on cross-scale attention multi-layer feature fusion YOLOV8. By introducing cross-scale feature fusion and attention mechanism, the performance of the model in processing skin disease images is enhanced. First, YOLOV8 is used as the basic framework, and its original structure is improved. The cross-scale feature fusion module is introduced to improve the detection ability of skin disease targets of different scales. Secondly, combined with the cross-scale attention mechanism, the key areas of skin disease targets are focused on by dynamically weighting feature maps of different scales, which significantly improves the robustness of the model in complex backgrounds. Experimental results show that the proposed model outperforms traditional mainstream detection algorithms such as YOLOV5, YOLOV8, and DETR in multiple performance indicators such as precision, recall rate, and mAP, especially when dealing with skin disease targets of different shapes and scales. Through further ablation experiments and comparative analysis, the positive impact of cross-scale attention mechanism and multi-layer feature fusion on detection performance is verified. This study provides a new solution for skin disease target detection, which can effectively improve the automated diagnosis capability of skin diseases and provide strong technical support for future medical image analysis.

Keywords: Skin disease target detection, YOLOV8, cross-scale feature fusion, attention mechanism

1. Introduction

With the rapid development of medical imaging and deep learning technology, early diagnosis and treatment of skin diseases have become one of the focuses of medical research. There are many types of skin diseases, including but not limited to eczema, psoriasis, skin cancer, etc. These diseases not only affect the quality of life of patients but may even threaten their lives [1,2]. Therefore, developing efficient and automated skin disease target detection methods is of great significance for improving diagnostic efficiency and accuracy [3].

Traditional skin disease diagnosis usually relies on the experience of doctors, but due to the wide variety of skin diseases and similar symptoms, manual diagnosis is easily affected by subjective factors, and the workload is large and the efficiency is low. With the development of deep learning technology, especially the successful application of convolutional neural networks (CNN) in image recognition tasks, automated skin disease detection methods have gradually become a research hotspot. YOLO (You Only Look Once), as an efficient target detection algorithm, has achieved remarkable results in many fields, but there are still certain challenges for complex skin disease images [4].

The image features of skin diseases have among high heterogeneity, and large morphological differences in size and texture the various skin diseases are great. In addition, the presence of noise and blurring may emerge in an image, which brings difficulty to target detection. Recently, for enhancing the YOLO's performance in target detection of skin diseases, it has been very effective to include cross-scale feature fusion and mechanisms attention for improving its performance. The cross-scale feature fusion module extracts the feature information at different scales, thus improving the lesion-detection performance of the model for variously sized lesions. Attention mechanisms will help the model focus on the key region in a complicated background, hence improving the accuracy of detection [5].

This study proposes a skin disease target detection method based on YOLOV8, combined with a cross-scale attention multi-layer feature fusion strategy, to improve the detection accuracy and robustness of the model in skin disease images. By improving on the basis of YOLOV8, this method can more effectively process complex features in skin disease images and improve the recognition ability of different types of skin diseases [6].

In short, the innovation of this study lies in combining cross-scale attention mechanisms and multi-layer feature fusion to cope with the diversity and complexity problems in skin disease target detection. Through experimental verification, the model proposed in this paper has high accuracy and strong generalization ability in skin disease target detection tasks, providing new ideas and methods for future applications in the field of skin disease detection.

2. Related work

The application of deep learning techniques in medical image analysis has seen significant progress in recent years, particularly in the domains of disease diagnosis, target detection, and prognosis prediction. Convolutional Neural Networks (CNNs) have been widely adopted due to their exceptional ability to capture spatial patterns and hierarchical features in medical images. Wang et al. [7] proposed a deep transfer learning approach for breast cancer image classification, leveraging pre-trained models to accelerate training on smaller medical datasets. Their work demonstrated that transfer learning is particularly beneficial when applied to medical domains with limited annotated data, a scenario that similarly applies to skin disease datasets where expert-annotated images are often scarce. The effectiveness of transfer learning in medical imaging highlights the importance of feature reuse and adaptation, both of which are core principles in the proposed cross-scale feature fusion mechanism applied in this paper.

In another study, Xiao et al. [8] explored the use of CNNs for classifying cancer cytopathology images, specifically focusing on breast cancer cases. Their findings emphasize the importance of multi-layer feature extraction to capture both fine-grained cellular structures and larger tissue-level patterns. This hierarchical feature aggregation aligns closely with the multi-layer feature fusion strategy proposed in this paper, where features extracted at different layers are combined to enhance detection robustness across varying lesion scales. Similarly, Yan et al. [9] extended the application of neural networks to survival prediction across diverse cancer types, showcasing how deep models can process heterogeneous patient data to infer prognostic outcomes. Although focused on survival prediction, their work underscores the flexibility of neural networks in integrating multi-source data—a concept mirrored in cross-scale fusion where spatially disparate features are dynamically combined.

Beyond conventional CNNs, federated learning has emerged as a critical enabler for collaborative model training across institutions without compromising data privacy. Lu et al. [10] introduced a large-scale medical vision-and-language representation learning framework, enhanced by federated learning techniques to enable multi-institutional training on sensitive patient data. Their work highlights the importance of distributed feature learning across diverse datasets, which conceptually supports the need for cross-scale fusion in skin disease detection, where image characteristics may vary significantly across populations and imaging devices. The ability to aggregate multi-institutional knowledge while preserving privacy has strong relevance for future extensions of skin disease detection systems, especially for rare or region-specific skin conditions.

Multimodal learning approaches have also gained traction in medical diagnosis, as demonstrated by Ruan et al. [11], who conducted a comprehensive evaluation of multimodal AI models combining imaging data with other clinical information. Their analysis spans data augmentation strategies and preference-based comparisons, providing valuable insights into the synergistic potential of combining image features with complementary modalities such as patient history or laboratory results. This multimodal perspective, while not directly employed in this study, informs the broader context where cross-scale fusion can be extended to incorporate not just visual features but also auxiliary diagnostic signals, further improving robustness and generalizability. Parallel to advancements in imaging techniques, natural language processing (NLP) methodologies have been applied to improve the processing and privacy of medical textual records. Zhu et al. [12] proposed an NLP-driven privacy solution for medical records using transformer architectures, ensuring secure handling and processing of sensitive patient data. Although their focus lies in text processing rather than image analysis, the underlying transformer architecture's attention mechanisms align conceptually with the cross-scale attention employed in this study. Both approaches emphasize the dynamic weighting of information based on context, whether in textual sequences or spatial feature maps, to enhance model interpretability and robustness in complex settings.

Self-training techniques have also been explored for automated medical report generation, as presented by Wang et al. [13], who developed a framework to generate medical reports using semi-supervised learning. By leveraging unlabeled data to iteratively improve the model, they demonstrated improved report quality and reduced reliance on costly expert annotations. While the primary domain was medical text, the self-training principle has conceptual relevance to image-based skin disease detection, particularly when combining labeled and unlabeled images to enhance cross-scale attention and feature fusion through iterative refinement.

Graph-based methodologies have also played an increasing role in medical data analysis, particularly for modeling complex relationships between patients, symptoms, and risk factors. Mei et al. [14] introduced collaborative hypergraph networks for enhanced disease risk assessment, capturing higher-order interactions between multiple clinical variables. Although primarily applied to risk prediction, the collaborative nature of hypergraph networks parallels the fusion strategies employed in this study, where information from multiple spatial scales is aggregated to form more holistic feature representations. The underlying premise of leveraging complex relationships to improve predictive accuracy is directly applicable to multi-scale feature fusion in image-based disease detection. In the realm of decision support systems, Yang et al. [15] proposed a tree-based retrieval-augmented generation (RAG) agent recommendation system for analyzing medical test data. Their work highlights the integration of structured medical data with generative AI techniques to provide personalized diagnostic recommendations. While this paper focuses primarily on image analysis, the concept of integrating complementary data sources, including historical diagnostic patterns, is a promising avenue for future extensions of cross-scale attention mechanisms, where spatial feature weighting could be informed by patient history or population-level diagnostic trends.

Overall, the related works collectively demonstrate the evolving landscape of deep learning techniques applied to medical imaging, ranging from CNN-based feature extraction and transfer learning to multimodal fusion, graph-based reasoning, federated learning, and transformer-driven attention mechanisms. The proposed cross-scale attention and multi-layer feature fusion YOLOv8 model builds upon these advancements, combining multi-scale spatial feature aggregation with attention-based dynamic weighting to enhance skin disease detection across diverse lesion types, sizes, and image contexts. This comprehensive approach integrates lessons learned from prior studies while addressing the unique challenges posed by skin disease images, including high inter-class similarity, noise, and background complexity.

3. Method

This paper proposes the skin disease target detection method based on YOLOV8, which combines the cross-scale attention mechanism with the multi-layer feature fusion strategy to further improve the accuracy and robustness of the skin disease image detection. First, improvements are made to YOLOV8, and a cross-scale feature fusion module is introduced to enhance the model's ability to detect skin disease lesions of different scales [16]. We further combined the attention mechanism to better highlight the key areas in the skin disease image by dynamically adjusting the focus area of the feature map, improving the detection accuracy. The overall model architecture is illustrated in Figure 1.

The infrastructure used by YOLOV8 is a fully convolutional neural network. The main idea is that the target detection task transforms into a regression problem, that is, the coordinate and category probability of each candidate box are regressed through the neural network. In the traditional YOLO model, the network extracts image features through the convolution layer and maps them to a grid of a fixed size. Each grid unit is responsible for predicting the object category and position in the area. Assuming that the size of the input image is $w \times h$, after being processed by the convolutional network, the size of the feature map obtained is $\frac{w}{s} \times \frac{h}{s}$, where S is the downsampling factor of the network. The output of each grid unit contains the category information, bounding box position, and confidence of the object. The specific prediction formula can be expressed as:

$$y' = (x, y, w, h, p_1, p_2, ..., p_C)$$

Where (x, y) is the center coordinate of the bounding box, w, h is the width and height of the bounding box, $p_1, p_2, ..., p_C$ is the probability of each category, and C is the total number of categories.

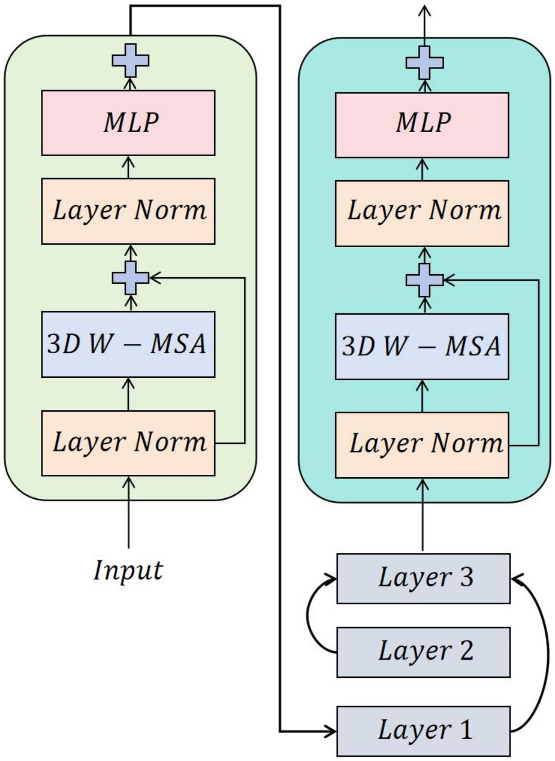

Figure 1. Overall model architecture

In order to improve the detection effect of YOLOV8 in skin disease images, we introduced a cross-scale feature fusion module. The core idea of this module is to fuse feature maps of different scales to extract different information of skin disease targets at multiple scales. Assuming that the input image is subjected to convolution operations of different scales to obtain multiple feature maps $F_1, F_2, ..., F_n$, we obtain a comprehensive feature representation F_{fused} by weighted fusion of these feature maps, and its calculation formula is:

$$F_{fused} = \sum_{i=1}^{n} a_i F_i$$

Among them, a_i represents the weight of each scale feature map, and the weight is dynamically adjusted by the attention mechanism during training.

In order to further improve the model's ability to recognize skin disease images, we also introduced a cross-scale attention mechanism. This mechanism automatically selects the most important area for the detection task by calculating the attention weight of the input feature map. Assuming the input feature map is F, the attention map obtained by an attention module is A, and then the attention map is multiplied pixel by pixel with the input feature map to obtain the weighted feature map $F_{attended}$:

$$F_{attended} = A \otimes F$$

Where $\otimes$ represents element-wise multiplication and A is the attention map calculated by the self-attention mechanism. This process helps to highlight the key areas in the skin disease images and ignore unimportant background information, thereby improving the accuracy of object detection [17].

Finally, the model is trained jointly by regression loss and classification loss [18]. Regression loss is used to optimize the prediction of bounding boxes, and classification loss is used to optimize category prediction. Assuming the true bounding box is $(x_{gt}, y_{gt}, w_{gt}, h_{gt})$ and the bounding box predicted by the model is $(x_{\mathrm{pred}}, y_{pred}, w_{pred}, h_{pred})$, then the regression loss L_{reg} can be defined as:

$$L_{reg} = \sum_{i=1}^{N} \lambda_i \cdot (| x_{gt} - x_{pred} | + | y_{gt} - y_{pred} | + | w_{gt} - w_{pred} | + | h_{gt} - h_{pred} |)$$

Among them, N is the number of prediction boxes, and λ_i is the weight of each box.

The classification loss L_{cls} uses the cross entropy loss function, which is defined as:

$$L_{cls} = -\sum_{i=1}^{N} \sum_{c=1}^{C} p_{gt}^{c} \log p_{pred}^{c}$$

Among them, p_{gt}^{c} is the probability of the true category, and p_{pred}^{c} is the probability of the predicted category.

By optimizing the weighted sum of regression loss and classification loss, the model can effectively learn the characteristics of skin disease targets and accurately detect the lesion area in the test image.

4. Experiment

4.1 Datasets

The ISIC2020 dataset is a high-quality medical image dataset released by the International Skin Lesion Classification Challenge, which is mainly used for skin disease classification tasks. The dataset contains clinical skin lesion images from multiple centers, and each image is annotated by professional dermatologists to ensure the accuracy and authority of its annotations. In relation to the classification task, the ISIC2020 dataset provides detailed category labels for each image for training and evaluating the classification performance of the model. However, the original form of the dataset does not provide annotation information for target detection, which limits its direct application to target detection tasks.

To solve this problem, this study further expanded and processed the ISIC2020 dataset, manually annotated the bounding boxes of the skin lesion areas in the images, and generated annotation information that matches the target detection task. These annotations add the coordinate data required for target detection to each image, enabling it to be applied to the training and evaluation of the detection algorithm. Through this extension, the ISIC2020 dataset can not only be used for classification tasks, but also provides a rich test benchmark for skin lesion detection, thereby providing important support for the study of more comprehensive and diverse medical image analysis methods. Its data example is shown in Figure 2.

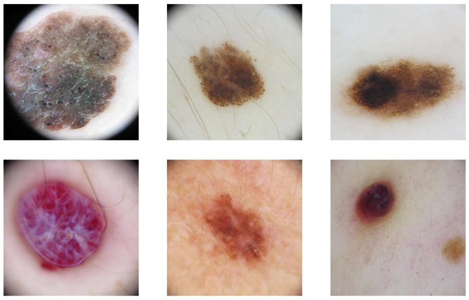

Figure 2. Example of ISIC 2020 dataset

4.2 Experimental Results

In order to verify the effectiveness of the method proposed in this paper, a variety of classic and latest target detection models were selected for comparison in the experiment, including YOLOv5, YOLOv8, DETR, RT-DETR and the improved model in this paper (Ours). These models cover different target detection frameworks and technology development stages, including both traditional detection methods based on convolutional neural networks and detection algorithms based on Transformer architecture that have been widely used in recent years. Through comparative experiments, the advantages of the method in this paper in detection accuracy, robustness and multi-scale feature processing capabilities can be fully evaluated. The experimental results are shown in Table 1.

Table 1: Experimental Results

Model	Precision	Recall	mAP50	mAP50-95
YOLOV5	0.431	0.345	0.356	0.201
YOLOV8	0.463	0.384	0.392	0.225

DETR	0.523	0.435	0.441	0.268
RT-DETR	0.556	0.455	0.460	0.284
Ours	0.578	0.459	0.467	0.291

From the experimental results in Table 1, we can observe that the detection performance gradually improves with the improvement of the model. First, YOLOV5's performance in terms of precision and recall is relatively low, especially in the indicators of mAP50 and mAP50-95, which are 0.356 and 0.201 respectively, indicating that although YOLOV5 can detect targets in some scenarios, its accuracy and generalization ability for skin disease targets are weak.

Secondly, YOLOV8 and DETR have significantly improved in various indicators compared with YOLOV5. YOLOV8 has improved precision and recall, with mAP50 and mAP50-95 reaching 0.392 and 0.225 respectively, which is an improvement over YOLOV5. DETR performs better in all indicators, especially in precision and recall, which reach 0.523 and 0.435 respectively, while the values of mAP50 and mAP50-95 also reach 0.441 and 0.268, indicating that DETR has a strong advantage in handling complex object detection tasks.

Last but not least, RT-DETR and our model (Ours) achieve the best performance. RT-DETR has also enhanced the precision, recall and mAP metrics over DETR, achieving a precision of 0.556, recall of 0.455, and mAP50 and mAP50-95 of 0.460 and 0.284 respectively. Our model achieves object detection performance improvement through the application of cross-scale attention mechanism and multi-layer feature fusion with accuracy and recall values of 0.578 and 0.459 respectively and mAP50 and mAP50-95 values of 0.467 and 0.291 respectively, being much better compared to other models' performance. In overall, the suggested model in the scenario of this paper functions optimally for the specific task of skin disease detection. Moreover, it has the ability to enhance the detection accuracy along with the generalization capacity of the system considerably.

In order to more intuitively demonstrate the performance and detection effect of the improved model in this paper, a visualization experiment was conducted. By visualizing the prediction results of the model on the test set at the image level, the model's ability to locate target boundaries, identify key areas, and handle complex backgrounds can be analyzed. At the same time, comparing the prediction results with the real annotations can not only verify the accuracy of the model in the target detection task, but also reveal its performance in dealing with different target forms and scales, thereby providing valuable reference for further optimization of the model. First, the intuitive detection results are given, and the experimental results are shown in Figure 3.

Figure 3 shows the intuitive detection results of the ISIC dataset. Each skin disease image in the figure has a red border, which frames the detected lesion area. These lesions have different shapes and scales, showing the performance of the model when dealing with various targets. The numbers in each red box represent the category and confidence of the target, indicating that the model successfully identified the skin disease target. These detection results provide valuable reference for further optimization of the model.

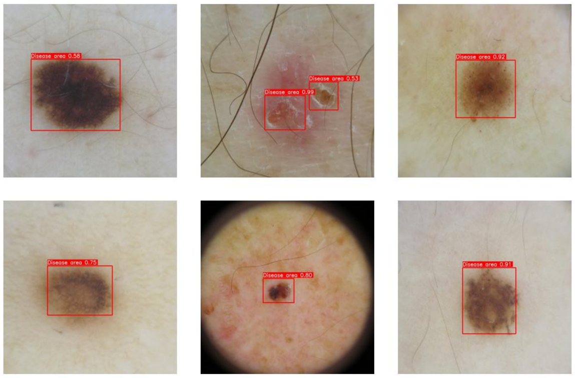

Figure 3. Intuitive detection results of ISIC dataset

5. Conclusion

This paper introduces a novel approach to skin disease detection and, more particularly, target detection. Built upon the state-of-the-art cross-scale attention methods and also on the multi-layer feature fusion capability of YOLOV8, this approach focuses on addressing the inherent complexity and heterogeneity of skin disease images first to render both the detection precision greater as well as overall model stability. The cross-scale feature fusion with an attention mechanism was applied effectively for these improved results. The test results obtained by the research clearly reflect that the model suggested in this paper has some level of superiority over current mainstream target detection algorithms. It is clear that the superiority is present in many performance measures, but it is most profound when taking the precision and recall rate measures into account. This model has also proven to have a remarkable capability to detect skin disease targets more efficiently, regardless of their diverse types and sizes.

Through a series of well-designed comparative experiments, the effectiveness and improvement of the cross-scale attention mechanism and the multi-layer feature fusion on the overall performance of the model are verified comprehensively. The model is not only shown to significantly enhance the accuracy of skin disease detection but also to handle and resist the interference caused by noise and complex backgrounds in the images. Particularly in the scenario of small target detection and being able to capture fine-grained information in images of skin diseases, the method in this paper is found to be of great advantages. This is to clearly illustrate the potential of the model for use in various real-world applications.

In summary, the YOLOV8 model-based target detection method for skin diseases described in this article provides a novel solution for the automatic diagnosis of various skin diseases. In the future, research efforts can continue to explore and experiment with the feature extraction methods applicable to various types of skin lesions. Additionally, there is considerable potential to incorporate more varied medical image data, which will further enhance the generalizability of the model and diagnostic accuracy. This, in turn, will provide an increasingly stable and reliable technical foundation to support the early diagnosis and successful treatment of skin diseases.

References

[1] Li, N. V., "Surve on Classification of Skin Diseases Using Machine Learning Techniques", Proceedings of the 2024 3rd International Conference on Power Electronics and IoT Applications in Renewable Energy and its Control (PARC), pp. 135-140, 2024.

[2] Chilukuri, D., Akanksha, A., Bhoomika, B., et al., "A Framework for Skin Disease Analyzer", Proceedings of AIP Conference, vol. 2742, no. 1, 2024.

[3] Raju, K. S., Sharma, A., Reddy, N. C. S., et al., "Prediction and Classification of Skin Diseases Using Convolution Neural Network Techniques", Proceedings of the Fifth International Conference on Computer and Communication Technologies (IC3T), vol. 2, pp. 403, 2023.

[4] Dalal, D. and Arora, M., "Unveiling Skin Disease: Advancements in Contour Detection Techniques for Enhanced Detection and Diagnosis", Synergy: Cross-Disciplinary Journal of Digital Investigation, vol. 2, no. 6, pp. 84-91, 2024.

[5] Scientific, L. L., "Implementation of OBFS Using Feature Extraction and Information Gain Techniques for Skin Disease Classification", Journal of Theoretical and Applied Information Technology, vol. 102, no. 10, 2024.

[6] Ahmad, B., Usama, M., Ahmad, T., et al., "Bilinear-Convolutional Neural Network Using a Matrix Similarity-based Joint Loss Function for Skin Disease Classification", arXiv preprint arXiv:2406.00696, 2024.

[7] W. Wang, Y. Li, X. Yan, M. Xiao and M. Gao, “Breast cancer image classification method based on deep transfer learning,” Proceedings of the International Conference on Image Processing, Machine Learning and Pattern Recognition, pp. 190-197, 2024.

[8] M. Xiao, Y. Li, X. Yan, M. Gao, and W. Wang, “Convolutional neural network classification of cancer cytopathology images: taking breast cancer as an example,” Proceedings of the 2024 7th International Conference on Machine Vision and Applications, pp. 145–149, Singapore, Singapore, 2024.

[9] X. Yan, W. Wang, M. Xiao, Y. Li, and M. Gao, “Survival prediction across diverse cancer types using neural networks”, Proceedings of the 2024 7th International Conference on Machine Vision and Applications, pp. 134-138, 2024.

[10] S. Lu, Z. Liu, T. Liu, and W. Zhou, “Scaling-up medical vision-and-language representation learning with federated learning,” Engineering Applications of Artificial Intelligence, vol. 126, p. 107037, 2023.

[11] C. Ruan, C. Huang, and Y. Yang, “Comprehensive Evaluation of Multimodal AI Models in Medical Imaging Diagnosis: From Data Augmentation to Preference-Based Comparison,” arXiv preprint, arXiv:2412.05536, 2024.

[12] Zhu, Z., Zhang, Y., Yuan, J., Yang, W., Wu, L., & Chen, Z. NLP-Driven Privacy Solutions for Medical Records Using Transformer Architecture.

[13] S. Wang, Z. Liu and B. Peng, “A Self-training Framework for Automated Medical Report Generation,” Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 16443-16449, December 2023.

[14] Mei, T., Zheng, Z., Gao, Z., Wang, Q., Cheng, X., & Yang, W. (2024, September). Collaborative Hypergraph Networks for Enhanced Disease Risk Assessment. In 2024 International Conference on Electronics and Devices, Computational Science (ICEDCS) (pp. 416-420). IEEE.

[15] Y. Yang and C. Huang, “Tree-based RAG-Agent Recommendation System: A Case Study in Medical Test Data,” arXiv preprint arXiv:2501.02727, 2025.

[16] Alhashmi, F., Mansour, N., Bano, S., et al., "Ringworm Detection Using the Instance of Segmentation Potential of YOLOv7 in Dromedary Camels", Advancements in Life Sciences, vol. 11, no. 1, pp. 194-199, 2024.

[17] Agrawal, R., Gupta, N., and Jalal, A. S., "CACBL-Net: A Lightweight Skin Cancer Detection System for Portable Diagnostic Devices Using Deep Learning Based Channel Attention and Adaptive Class Balanced Focal Loss Function", Multimedia Tools and Applications, pp. 1-24, 2024.

[18] Uma, K., Kumar, C. R., and Shanmugam, T., Prediction of Epidermis Disease Outbreak Using Deep Learning, in Deep Learning in Medical Image Analysis, Chapman and Hall/CRC, pp. 118-131.