Research on Intelligent Diagnosis of Vitiligo Based on UNet Segmentation and ViT Models

Jiamin Li[1], Sihan Wang[1], Xuwen Zhang[1], Yaxuan Yang[1], Xianyi Chen[1], Mingming Gong[2]
[1]School of Artificial Intelligence and Big Data, Henan University of Technology, Zhengzhou, China
[2]iFlytek Co., Ltd., Hefei, Anhui, China

Abstract: In aviation medical selection, accurate staging of vitiligo faces challenges including low efficiency, high subjectivity, and scarce data. To address this, this paper proposes and implements an AI-assisted vitiligo diagnosis system tailored for aviation recruitment scenarios. The system utilizes dual-modal inputs—clinical daylight images and Wood's lamp images—to construct a two-stage architecture: "UNet segmentation + ViT classification. " First, an improved attention-gated UNet model achieves precise lesion segmentation. Subsequently, a 6-channel dual-stream ViT architecture, combined with hierarchical transfer learning and BalancedFocalLoss, addresses the challenge of training with limited data. The system features a modular design, enabling image upload, visual lesion annotation, staging diagnosis, and treatment recommendations, while enhancing clinical interpretability through a confidence feedback mechanism. Results demonstrate 91. 1% specificity and 82. 1% accuracy on the test set, effectively resolving efficiency and accuracy bottlenecks in aviation medical examinations. This system provides technical support for intelligent vitiligo diagnosis in aerospace medicine, combining clinical utility with military application value.

Keywords: Vitiligo; AI-Assisted Diagnosis; UNet Segmentation; Vision Transformer; Balanced Focal Loss

1. Introduction

Vitiligo is an acquired, localized or generalized depigmenting skin disorder[1][1]. Vitiligo exhibits a global prevalence of 0.1% to 2%, with no statistically significant disparities across racial/ethnic groups or between genders. Epidemiological statistics confirm that over 50% of patients develop the condition prior to age 20, often with onset in adolescence. While not directly life-threatening or organ-impairing, the disease causes characteristic depigmented lesions that substantially affect appearance, leading to marked psychological distress and social barriers, which collectively diminish patients' overall quality of life. The pathogenesis of vitiligo is complex and not fully understood, potentially involving multiple factors such as genetics, autoimmunity, neurohumoral regulation, oxidative stress, and melanocyte self-destruction[2][2]. Current treatment goals primarily focus on controlling disease progression and promoting repigmentation, with common approaches including medication, phototherapy, and surgical transplantation[3][3].

Vitiligo can be classified into progressive and stable phases based on disease activity, clinical presentation, and Wood's lamp examination results. Early diagnosis and accurate classification are crucial for improving treatment efficacy and predicting disease progression. However, clinical assessment of vitiligo lesions still relies primarily on visual inspection and subjective judgment by physicians, which is significantly influenced by lighting conditions, skin tone variations, and physician expertise, resulting in limited objectivity and consistency[4][4].

In recent years, with the rapid advancement of artificial intelligence and medical image analysis technologies, deep learning-based automated skin disease recognition methods have emerged as a research hotspot. While convolutional neural networks excel in image feature extraction [5], they struggle to capture global structural information in skin images and exhibit limited adaptability to lighting variations and skin texture. To address these limitations, this study proposes an intelligent vitiligo diagnosis system integrating the U-Net segmentation network with the Vision Transformer model. The overall system workflow encompasses data acquisition, image preprocessing, lesion segmentation, feature extraction, and intelligent

discrimination. First, dual-image input from visible light and Wood's lamp enhances lesion recognition sensitivity. Subsequently, the U-Net model achieves precise automatic segmentation of vitiligo lesions. Finally, leveraging the global feature modeling capability of the ViT model, the segmentation results undergo classification and diagnostic analysis. The overall workflow is illustrated in Figure 1.

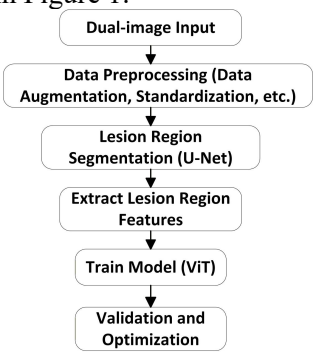

Figure 1. Overall Research Workflow Diagram

This study constructs an intelligent model dedicated to the automatic identification and objective quantitative assessment of vitiligo, equipping clinical practice with more convenient, rapid, and accurate disease recognition and evaluation capabilities. By optimizing algorithmic performance and validating with clinical data, the model effectively enhances the accuracy of early vitiligo diagnosis and improves its clinical applicability and feasibility. Moreover, it furnishes novel research perspectives and technical support for the advancement of intelligent image analysis in dermatological clinical practice and research.

2. Multimodal Image Data Preprocessing

Precise segmentation of vitiligo lesion areas via advanced image processing techniques enables in-depth quantitative analysis of the morphological characteristics (e. g. , size, shape, edge regularity) and spatial distribution patterns of depigmented lesions. This refined lesion delineation not only lays a solid foundation for accurate disease staging and subtype classification but also provides an objective metric for tracking dynamic changes in lesions, thereby effectively facilitating the scientific assessment of treatment efficacy and optimization of personalized therapeutic regimens.

2.1 Data Description

The multimodal image dataset used in this study was provided by the Chinese PLA Air Force Medical Center with granted usage rights. The dataset comprises clinical raw images and Wood's lamp images, categorized and stored according to progressive and stable phases. Each folder contains clinical raw images of affected areas alongside corresponding Wood's lamp images, all in JPG format.

2.2 Data Preprocessing

Analysis of the collected dermatological lesion dataset identified several inherent issues in the raw images, including uneven illumination across imaging scenes, noticeable contrast variations between lesion and normal skin regions, and unavoidable image noise induced by imaging equipment or environmental interference. These confounding factors are prone to distort lesion features and reduce data quality, thereby potentially undermining the stability and performance of subsequent model training. Consequently, this study implemented a targeted series of data preprocessing measures to mitigate such drawbacks and enhance overall image quality.

By analyzing the characteristics of Wood's lamp images and clinical images, specialized augmentation strategies were designed. The comprehensive data augmentation workflow[6]was implemented using the Albumentations library, incorporating geometric transformations such as horizontal and vertical flipping to increase data diversity. To address the uneven lighting issue in raw images, targeted techniques including adaptive brightness/contrast adjustment and CLAHE (Contrast-Limited Adaptive Histogram Equalization) were applied to standardize illumination. Gaussian blur and Gaussian noise augmentation were introduced to enhance the model’s robustness against inherent and environmental noise. For minority samples, stronger data augmentation parameters (e. g. rotation, flipping) were adopted to mitigate class imbalance, thereby optimizing the data distribution and improving the model’s generalizability.

This study employs a dual-modal data approach for joint analysis. Image pairing accuracy is ensured through filename cleansing and precise matching, with unmatched pairs excluded. All images are uniformly resized to the standard 224×224 pixel dimensions and normalized using the mean and standard deviation from the

ImageNet dataset, effectively accelerating model convergence.

Prior to implementing automatic segmentation, clinical original images and affected regions in Wood's lamp images within the dataset were annotated using LabelMe to generate a labeled dataset for model training. Figure 2 displays a portion of these annotations.

Figure 2. Label Display

3. Lesion Segmentation and Classification Model Construction

3.1 Improved U-Net Model with Fusion Attention Mechanism

In the intelligent vitiligo diagnosis system, precise segmentation of lesion areas serves as a core prerequisite for reliable quantitative assessment and dynamic trend analysis of the disease. The U-Net model, renowned for its symmetric encoder-decoder architecture and ingenious skip-connection design, efficiently integrates fine-grained local details with high-level global semantic features, thus establishing itself as a classic and widely adopted paradigm in the field of medical image segmentation[7]. Nevertheless, the model still exhibits inadequate segmentation accuracy in challenging clinical scenarios, such as when hair occlusion obscures lesion regions and in cases where blurred boundaries of vitiligo lesions appear in dermatoscopic images. These limitations may hinder its broader adoption and practical clinical application.

To address these issues, this paper proposes an enhanced U-Net model incorporating attention mechanisms, specifically optimized for the unique characteristics of vitiligo images. In the data processing stage, a specialized dataset is constructed using clinically acquired progressive and stable-phase vitiligo images. Sample augmentation is systematically achieved through geometric transformations such as rotation and photometric adjustments like contrast enhancement, while a local adaptive algorithm [8] is applied to enhance texture differences between vitiligo patches and normal skin. The preprocessing pipeline also adapts to two different mask formats to enable unified loading and processing of multi-source data. Architecturally, the improved model embeds channel attention modules within the skip connections of the traditional U-Net framework to dynamically allocate feature weights and enhance focus on subtle vitiligo boundary information. The encoder component employs deep separable convolutions to substantially reduce parameter count and computational complexity, while the decoder progressively restores spatial resolution through carefully designed upsampling operations. To visually demonstrate the model's enhanced capability in capturing vitiligo features, Figure 3 presents detailed feature maps processed by the improved U-Net model, clearly illustrating the selective focusing capabilities of different convolutional channels on vitiligo regions and their boundaries.

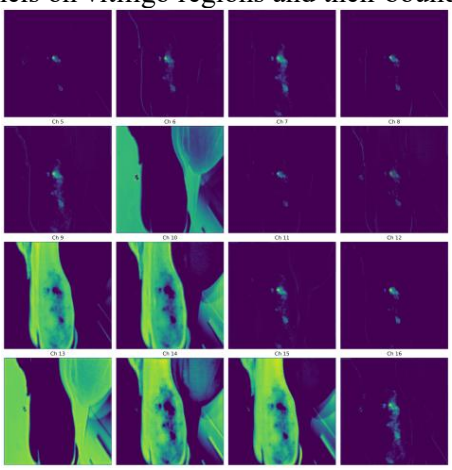

Figure 3. Feature Mapping Results of the Improved U-Net model

3.2 Multi-dimensional Architecture Design of the Enhanced SimplifiedViT Classification Module

Centered on an enhanced SimplifiedViT model, the classification module employs multidimensional architecture optimization to efficiently utilize multimodal imaging information, adapting to clinical demands for vitiligo staging diagnosis. To fully integrate complementary features from segmented clinical images and Wood's lamp images, the model modifies the standard ViT input layer to a 6-channel design: 3 channels of clinical images and 3 channels of Wood's lamp images are directly concatenated into a unified tensor. This

enables the model to simultaneously learn key information from both modalities within a single coherent architecture, effectively avoiding the feature fragmentation issues inherent in traditional multimodal fusion [9].

Addressing the limitation of standard ViT's rigid grid-based patch segmentation failing to adapt to vitiligo's irregular lesions, the model innovatively introduces a deformable patch embedding module. Based on deformable convolution principles, this module learns two-dimensional offset parameters (Δx, Δy) for each patch center. This function enables dynamic grid adjustment to precisely cover lesion boundaries and areas with heterogeneous pigment distribution, thereby significantly enhancing the accuracy of feature representation for irregular vitiligo lesions. To strike a balance between capturing fine local texture details and integrating comprehensive global contextual information, the model incorporates a multi-scale attention mechanism into the ViT encoder. Specifically, fine-grained texture details and global spatial distribution features are extracted in parallel through self-attention windows of different sizes (e. g. , 8×8, 32×32), which are then adaptively fused via a gated fusion network to optimize feature integration. This achieves an organic combination of the local feature extraction advantages of convolutional neural networks and the global modeling capabilities of Transformers. The effectiveness of this architectural design is visually demonstrated by the feature distribution map in Figure 4 below. The vertical axis represents the probability density of feature values, reflecting the relative frequency of feature occurrence across different numerical intervals. Higher density values indicate that features within that interval are more concentrated and active within the model. The feature distribution across the entire workflow—from multimodal fusion , classifier input , intermediate layer , to final output —reveals that each layer effectively performs nonlinear transformations and information refinement on features. The feature distribution in the fusion layer is relatively dispersed, reflecting the integration of multi-source information. Feature distributions across classifier stages gradually converge and exhibit a bimodal pattern, indicating the model has successfully learned discriminative features that provide robust support for final classification decisions.

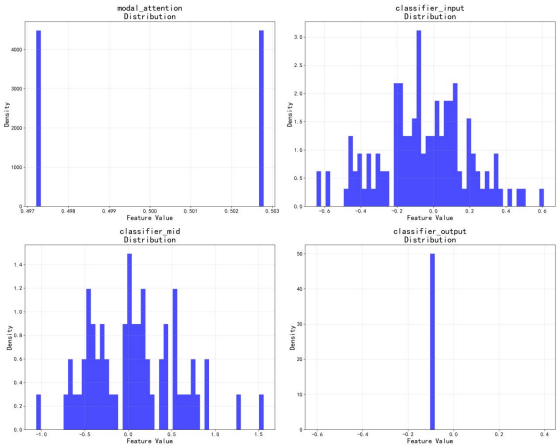

Figure 4. Core Feature Visualization

The modal attention distribution of the enhanced SimplifiedViT model presents a distinct bimodal pattern, which vividly reflects the differentiated integration strategy of multimodal imaging information. Specifically, this pattern indicates that the model assigns adaptive attention weights to different modal data based on their information value, ensuring effective synergy between complementary features.

Notably, the input feature distribution of the classifier is remarkably uniform and stable, free from extreme fluctuations or bias, which provides high-quality and reliable data support for subsequent vitiligo staging classification. Furthermore, the features extracted from the classifier's intermediate layer undergo thorough refinement and rational reorganization through the model's internal mechanism, ultimately forming a multi-peak and highly discriminative distribution—this characteristic enables clear distinction between different disease stages and lesion types. In the final output stage, the classifier's output features exhibit a well-converged distribution, with distinct clustering of features corresponding to different diagnostic categories, thereby strongly supporting the precise decision-making process for vitiligo staging diagnosis. Collectively, these prominent distribution characteristics, which span the entire technical chain from multimodal information fusion to classification output, fully validate the enhanced SimplifiedViT model's superior capability to efficiently utilize multimodal imaging resources. This validation underscores the effectiveness of the model's multidimensional architectural optimizations in boosting diagnostic performance.

Considering deployment requirements in resource-constrained scenarios like aviation medical examinations, the model undergoes

further lightweight optimization. It replaces some fully connected layers with separable convolutions and incorporates sparse attention mechanisms, prioritizing computations within pre-identified lesion candidate regions. This approach maintains diagnostic accuracy while meeting the efficiency demands of real-time clinical diagnosis. The design of this enhanced SimplifiedViT model provides efficient and precise technical support for vitiligo staging and classification. Its multimodal adaptability and lightweight characteristics enable flexible adaptation to diverse clinical application scenarios.

4. Experimental Results

4.1 U-Net Model Performance Evaluation Metrics

The model's training strategy adopts a combined loss function integrating Dice loss and cross-entropy loss, which is specifically designed to optimize segmentation performance for vitiligo lesions. This hybrid loss design effectively balances two critical objectives: the pixel-level classification accuracy ensured by cross-entropy loss and the accurate overlap of predicted lesion regions with ground truth facilitated by Dice loss, thereby addressing the class imbalance issue common in medical image segmentation [10].

As shown in Table 1, experimental results on the independent validation set demonstrate the superior performance of the improved model: it achieves a Dice coefficient of 0. 9429, a key metric for evaluating segmentation consistency, while the Intersection over Union metric remains consistently above 0. 89. These results not only reflect high precision in lesion delineation but also represent a significant and statistically meaningful improvement over the traditional U-Net model. Such enhancements underscore the rationality of the proposed training strategy and the model's potential for clinical application in vitiligo diagnosis.

Table 1. Model Comparison: Dice Coefficient and IoU Metrics

Model Category	Dice Coefficient	IoU
Traditional U-Net	0. 7523	0. 6413
Improved U-Net	0. 9429	0. 8912

The lesion masks generated by this segmentation module can serve as input features for the ViT model, providing quantitative evidence for subsequent vitiligo progression prediction and establishing an integrated "segmentation-analysis" intelligent diagnosis workflow.

4.2 Visual Comparison of Diagnostic Results

To intuitively showcase the practical performance of the intelligent diagnosis system and pinpoint the key factors that exert an impact on diagnostic accuracy, this study carried out in-depth and detailed analysis on representative cases of progressive vitiligo. These cases were carefully selected for their typical clinical characteristics, ensuring the analysis's relevance and reliability. Figure 5 displays the results generated by this U-Net model for these selected cases, providing a clear baseline for evaluating the advantages of the improved system.

Figure 5. Visual Comparison of Diagnosis Results from Dual-Modality Inputs for Vitiligo

Figure 5: Left panel shows the Wood's lamp image of this case, revealing multiple irregular blue-white fluorescent areas on the skin surface corresponding to clinically observed depigmented lesions. However, the image also contains interfering information such as localized strong light reflections and equipment shadows. The right panel displays the lesion segmentation results from the U-net model, with a heatmap visually illustrating the system's detection of affected areas. Red areas indicate suspected lesion regions identified by the system, with color intensity reflecting confidence levels. The heatmap demonstrates successful detection of most lesions, though image quality issues result in lower confidence in certain border regions.

This visual comparison not only demonstrates the system's performance characteristics but also provides a basis for optimizing image acquisition conditions and improving algorithms. Simultaneously, this "raw image + segmentation heatmap" presentation transforms the AI diagnostic process from a "black box" into an explainable one: clinicians can quickly grasp the system's recognition priorities through the heatmap while manually reviewing low-confidence areas based on their expertise. This establishes an efficient "AI preliminary screening + human verification" diagnostic

model, further enhancing the system's practicality and reliability in clinical settings.

5. Conclusion

This study developed an AI-assisted diagnostic system for vitiligo tailored to aviation recruitment scenarios. It integrates the precise lesion segmentation capability of an improved attention-gated UNet model with the multimodal feature classification capability of a 6-channel dual-stream ViT architecture. Through dual-modal input processing of clinical daylight and Wood's lamp images, lesion segmentation extraction, and the integration of hierarchical transfer learning strategies with the BalancedFocalLoss function, the system successfully achieves precise staging classification of vitiligo while effectively addressing the challenge of training with limited sample data.

This system not only reduces the workload for physicians but also minimizes issues arising from subjective experience differences through objective, quantifiable diagnostic results. It provides an effective and standardized technical solution for skin disease screening in aviation medicine, serving as a reference for medical AI applications in primary care and specialized settings.

Future research may focus on two key areas: First, enhancing the model's adaptability to more complex cases; second, integrating real-time monitoring data from wearable devices to provide more comprehensive decision support for personalized clinical treatment plans.

References

[1] Kaushik H, Kumar V, Parsad D. Dysregulated keratinocyte proliferation and maturation might lead to loss of melanocytes causing depigmentation in vitiligo. Journal of the American Academy of Dermatology, 2025, 93(3S): AB179-AB179. DOI: 10. 1016/J. JAAD. 2025. 05. 711.

[2] Wang J, Zhang C, Wu H, et al. Recent Advances in the Pathogenesis of Vitiligo and the Application of Novel Drug Delivery Systems in Its Treatment. International Journal of Pharmaceutics:X, 2025, 10100397-100397. DOI:10. 1016/J. IJPX. 2025. 10039.

[3] Sallehuddin N, Fadilah M I N, Fauzi B M, et al. A Scoping Review of Pathogenesis, Current Treatments, and Novel Approaches for Vitiligo. Journal of Cosmetic Dermatology, 2025, 24(10): e70444. DOI: 10. 1111/JOCD. 70444.

[4] Peralta-Pedrero ML, Morales-Sánchez MA, Jurado-Santa Cruz F, et al. Systematic Review of Clinimetric Instruments to Determine the Severity of Non-segmental Vitiligo. Australas J Dermatol. 2019;60(3):e178-e185.

[5] Shabir A, Ahmed T K, Mahmood A, et al. Deep image features sensing with multilevel fusion for complex convolutional neural networks & cross-domain benchmarks. PloS one, 2025, 20(3):e0317863. DOI:10. 1371/JOURNAL. PONE. 0317863.

[6] Xu Chao, Wang Yunjian, Liu Yang, et al. Automatic Diagnosis of Knee Osteoarthritis X-ray Images Based on an Improved SWIN Transformer. Electronic Measurement Technology, 2024, 47(19): 155-163. DOI:10. 19651/j. cnki. emt. 2416040.

[7] Fan Zhuangzhuang. Research on Deep Learning-Based Quantitative Evaluation Methods for Vitiligo Treatment Outcomes. Anhui Medical University, 2025. 17-19.

[8] Wang H, Liu X, Liu L, et al. UHF RFID Sensing for Dynamic Tag Detection and Behavior Recognition: A Multi-Feature Analysis and Dual-Path Residual Network Approach. Sensors, 2025, 25(17): 5540. DOI:10. 3390/S25175540.

[9] Wei Xiaoya. Research on Deep Learning-Based Multimodal Medical Image Fusion Algorithms. Anhui University of Science and Technology, 2025. DOI:10. 26918/d. cnki. ghngc. 2025. 001274.

[10] Thinley J, O'Keefe S, Ndehedehe C. Mapping Tree Cover Patterns in an Urban Arboretum from Multispectral Drone Imagery Using Pixel-Based Classification and Object-Based Image Analysis. Remote Sensing in Earth Systems Sciences, 2025, 9(1): 3-3. DOI: 10. 1007/S41976-025-00257-W.

