Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Automated Grading of Radiation Dermatitis in Breast Cancer Radiotherapy with Dual-Head Supervision and Rule-Based Fusion: A Multicenter Study

Hongfei Sun
Department of Radiation Oncology, Xijing Hospital, Fourth Military Medical University
Zihan Shi
Department of Radiation Oncology, Xijing Hospital, Fourth Military Medical University
Cong Zhang
Department of Radiation Oncology, Xijing Hospital, Fourth Military Medical University
Qilong Huang
Department of Radiation Oncology, Xijing Hospital, Fourth Military Medical University
Jing Guo
Department of Radiation Oncology, Xijing Hospital, Fourth Military Medical University
Yixin Cui
School of Automation, Northwestern Polytechnical University
Yu Tian
Department of Radiation Oncology, Xijing Hospital, Fourth Military Medical University
Bo Lv
Department of Radiation Oncology, Xijing Hospital, Fourth Military Medical University
Guangjin Chai
Department of Radiation Oncology, Xijing Hospital, Fourth Military Medical University
Yan Huang
Xi'an International Medical Center Hospital
Wenjing Zhang
Xi'an International Medical Center Hospital
Jie Li
Department of Radiation Oncology, Xijing Hospital, Fourth Military Medical University
Yufen Liu
Department of Radiation Oncology, Xijing Hospital, Fourth Military Medical University
Fan Meng
Department of Radiation Oncology, Xijing Hospital, Fourth Military Medical University
Yan Cui
Department of Radiation Oncology, Xijing Hospital, Fourth Military Medical University

Meixi Liu
Department of Radiation Oncology, Xijing Hospital, Fourth Military Medical University
Xiaomeng Wang
Department of Radiation Oncology, Xijing Hospital, Fourth Military Medical University
Jie Gong
18700182707@163.com
Department of Radiation Oncology, Xijing Hospital, Fourth Military Medical University
Lina Zhao
Department of Radiation Oncology, Xijing Hospital, Fourth Military Medical University

---

Article

Keywords: Breast cancer, Radiotherapy, Radiation dermatitis, Deep learning, Automated grading

Posted Date: May 21st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9478524/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Automated Grading of Radiation Dermatitis in Breast Cancer Radiotherapy with Dual-Head Supervision and Rule-Based Fusion: A Multicenter Study

Hongfei Sun[1,4]; Zihan Shi[1,4]; Cong Zhang[1,4]; Qilong Huang[1,4]; Jing Guo[1,4]; Yixin Cui[2,4]; Yu Tian[1]; Bo Lv[1]; Guangjin Chai[1]; Yan Huang[3]; Wenjing Zhang[3]; Jie Li[1]; Yufen Liu[1]; Fan Meng[1]; Yan Cui[1]; Meixi Liu[1]; Xiaomeng Wang[1]; Jie Gong[1,*]; Lina Zhao[1,*]

1 Department of Radiation Oncology, Xijing Hospital, Fourth Military Medical University, Xi'an, Shaanxi, China

2 School of Automation, Northwestern Polytechnical University, Xi'an, Shaanxi, China

3 Xi'an International Medical Center Hospital, Xi'an, Shaanxi, China

4 These authors contributed equally: Hongfei Sun, Zihan Shi, Cong Zhang, Qilong Huang, Jing Guo and Yixin Cui.

Corresponding Author:

Lina Zhao, Ph.D
Department of Radiation Oncology
Xijing Hospital, Fourth Military Medical University
Xi'an, China
Email: zhaolina@fmmu.edu.cn

Co-corresponding Author:

Jie Gong, Ph.D
Department of Radiation Oncology
Xijing Hospital, Fourth Military Medical University
Xi'an, China
Email: 18700182707@163.com

Running Title: Automated Grading of Radiation Dermatitis

Abstract

Purpose: To develop and validate an automated grading framework that integrates dual-head supervision and rule-based fusion for severity assessment of radiation dermatitis during breast cancer radiotherapy, and to evaluate its performance across centers and cancer types.

Materials and Methods: This multicenter study included 10,685 skin images acquired during radiotherapy from 1,021 patients. The primary cohort comprised 833 patients with breast cancer, the independent external validation cohort comprised 143 patients with breast cancer, and the cross-cancer validation cohort comprised 45 patients with nasopharyngeal carcinoma. Using prospectively acquired skin photographs, we developed a dual-head deep learning framework consisting of a cross-entropy classification head, an ordinal regression head, and a rule-based fusion strategy. Performance was evaluated with internal validation, independent external validation, ablation analysis, AI-versus-clinician comparison, and cross-cancer validation.

Results: In the internal validation cohort, the proposed model achieved an accuracy, macro-F1, and quadratic weighted kappa (QWK) of 0.947, 0.947, and 0.958, respectively. The corresponding values in the independent external validation cohort were 0.909, 0.908, and 0.918. Ablation analysis showed that the fusion model outperformed the cross-entropy–only head, the ordinal-only head, and the dual-head model without fusion. In the AI-versus-clinician comparison, the model achieved an overall accuracy of 93.3%, exceeding that of three clinicians. In cross-cancer validation, the model maintained an accuracy of 0.891 and a QWK of 0.907.

Conclusion: Dual-head supervision with rule-based fusion improved automated grading of radiation dermatitis in breast cancer radiotherapy and showed consistent performance across centers and cancer types.

Keywords: Breast cancer; Radiotherapy; Radiation dermatitis; Deep learning; Automated grading

Introduction

With broader population-based screening, more refined molecular subtyping, and continuing therapeutic advances, outcomes for patients with breast cancer have improved substantially [1-3]. In China, data from the National Cancer Center show a marked increase in 5-year breast cancer survival over time [4]. In the United States, breast cancer mortality declined by 44% between 1989 and 2022 [5]. As a key component of local treatment for breast cancer, particularly after breast-conserving surgery, adjuvant radiotherapy reduces recurrence and improves long-term outcomes [6-8]. The 2024 ESMO Clinical Practice Guideline states that radiotherapy remains part of standard local treatment for most patients after breast-conserving surgery, whereas postmastectomy radiotherapy is primarily indicated for patients at higher risk of locoregional recurrence [9]. In the landmark EBCTCG individual patient data meta-analysis, radiotherapy after breast-conserving surgery reduced the 10-year risk of first recurrence from 35.0% to 19.3% and the 15-year risk of breast cancer death from 25.2% to 21.4% [10]. However, these gains in local control and long-term survival are accompanied by treatment-related toxicities, among which radiation dermatitis is one of the most common adverse effects [11-13]. In breast cancer radiotherapy, the chest wall and regional nodal fields often place a broad skin surface close to high superficial dose regions, increasing the likelihood of grade 2 or higher radiation dermatitis and adversely affecting treatment experience and quality of life. Timely, accurate, and reproducible grading of

radiation dermatitis is therefore clinically important for toxicity management, treatment continuity, and patient-centered care [14].

At present, severity assessment for radiation dermatitis in breast cancer radiotherapy still relies primarily on clinical grading systems such as the Radiation Therapy Oncology Group (RTOG) and the Common Terminology Criteria for Adverse Events (CTCAE) [15,16]. These tools are practical and widely used in both clinical care and research, but grading remains based largely on visual inspection of erythema, pigmentation change, and dry or moist desquamation, and therefore depends substantially on reader experience [17]. A systematic review including 423 studies and 227 randomized controlled trials found that 58 different tools have been used to assess radiation dermatitis, underscoring the lack of consensus in this field [18]. Further studies have shown limited correlation between existing clinician-reported scales and selected physiologic skin measurements, as well as suboptimal agreement between patient-reported outcomes and clinician assessments [19,20]. Together, these limitations highlight the need for an objective and consistent adjunctive assessment approach for radiation dermatitis in breast cancer radiotherapy.

In recent years, artificial intelligence has opened new avenues for more objective assessment of radiation dermatitis [21-24]. Lee et al. analyzed 120 patients with breast cancer treated with volumetric modulated arc therapy and integrated clinical, radiomic, and dosiomic features to predict grade 2 or higher radiation dermatitis, reporting area under the receiver operating characteristic curve (AUC) values of 0.82 and 0.83 for logistic regression and random forest models, respectively [25]. Park et al. developed a deep learning model with dilated convolutions and skip connections for automated segmentation of radiation dermatitis regions using 73 skin photographs from patients receiving radiotherapy for head and neck tumors and breast cancer, and reported mean Dice similarity coefficients of 0.62 for mild dermatitis (CTCAE grade 1) and 0.69 for severe dermatitis (CTCAE grade 2) [26]. These studies suggest that AI can be used not only for risk prediction of radiation dermatitis but also for objective quantification of lesion extent. However, prior work has focused mainly on toxicity prediction or lesion segmentation rather than direct severity grading from skin photographs. By contrast, Ni et al. developed a DenseNet121-based automated grading model using 1,205 head and neck skin photographs from patients with nasopharyngeal carcinoma, with an overall test-set accuracy of 83.0% and class-specific accuracies of 90.5%, 67.2%, 93.8%, and 100% for grades 0, 1, 2, and greater than 2, respectively [27]. Nevertheless, head and neck radiotherapy differs substantially from breast cancer radiotherapy in anatomy, dose distribution, and cutaneous phenotype, and that study did not include cross-cancer validation. The generalizability of those findings to breast cancer remains uncertain.

A recent review found that studies on automated grading of radiation dermatitis remain limited and are commonly constrained by small datasets and insufficient cross-center validation [28]. To address these gaps, we developed an image-based automated grading framework for radiation dermatitis in breast cancer radiotherapy using a large prospectively collected clinical photo dataset. Because radiation dermatitis grading reflects an ordered progression of skin injury severity, we incorporated both a cross-entropy (CE) head and an ordinal regression head on top of a shared backbone. The former directly predicts grades 0, 1, and 2, whereas the latter constrains grade ordering through stepwise predictions of whether a case has reached grade 1 or higher and grade 2 or higher, thereby reducing implausible between-grade errors. We further introduced a rule-based fusion strategy to integrate outputs from the two heads for boundary cases and improve stability, robustness, and clinical consistency. To assess effectiveness and transportability, we performed internal validation, independent external validation, ablation analysis, and cross-cancer validation in radiation dermatitis associated with nasopharyngeal carcinoma radiotherapy. Our aim was to develop a more objective, reproducible, and real-time method for grading radiation dermatitis in breast cancer radiotherapy and to provide technical support for toxicity monitoring and refined clinical management.

Materials and Methods

An overview of the multicenter study design is shown in Figure 1. Briefly, we prospectively enrolled a breast cancer cohort and collected skin photographs during radiotherapy, which were paired with clinical grades to construct datasets for model development and validation. After image processing and label definition, we developed an automated grading model integrating dual-head supervision and rule-based fusion. Model performance was evaluated through internal validation, independent external validation, and cross-cancer generalization analysis. We also performed ablation studies to clarify the contribution of individual components. Details of study design, image acquisition, model development, validation, and interpretability analysis are provided below.

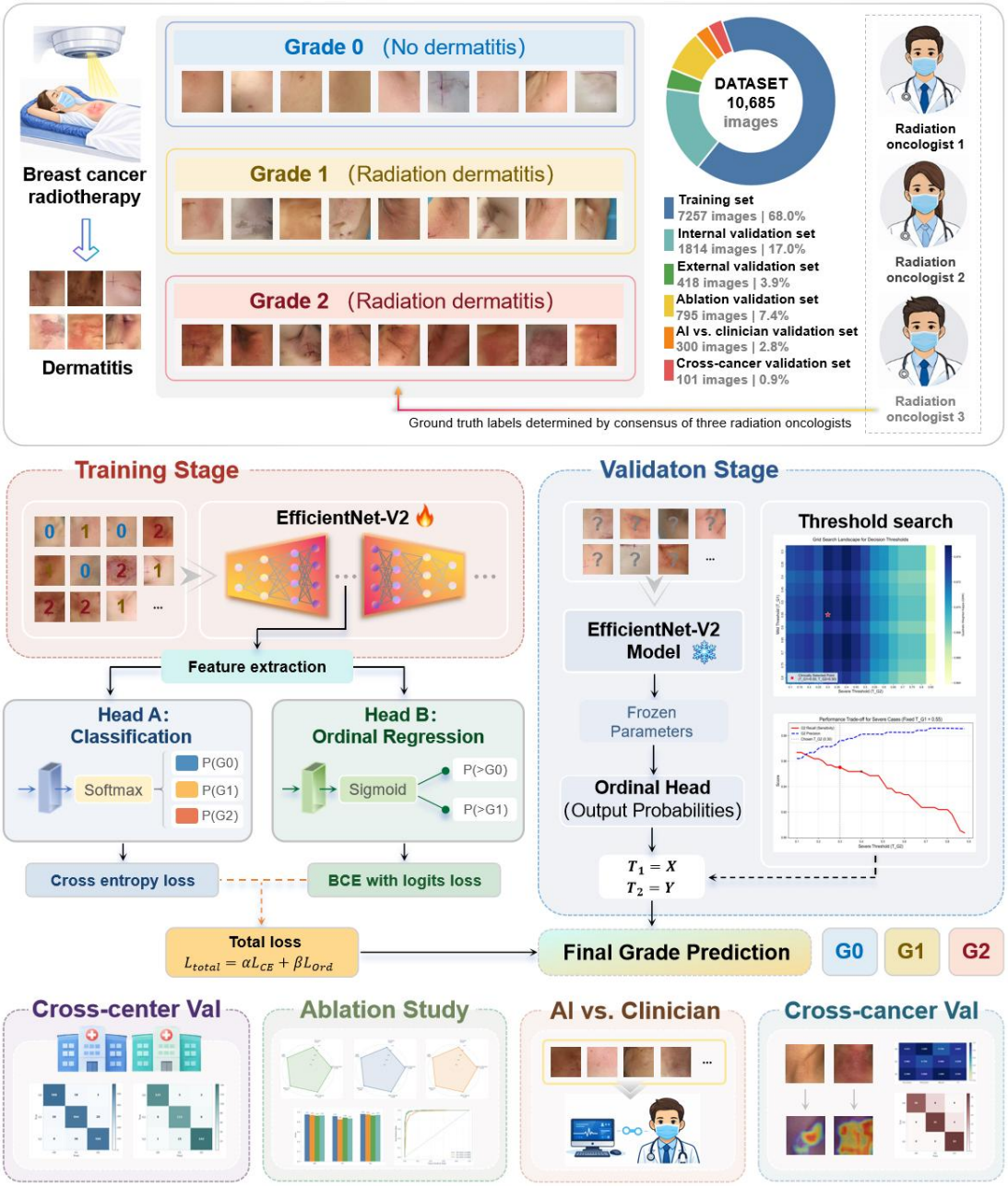


Figure 1. *Overview of the proposed automated grading framework for radiation dermatitis in breast cancer radiotherapy. The workflow includes dataset construction, reference-standard labeling, model development,*

and multilevel validation. Skin photographs were assigned to grade 0, grade 1, or grade 2 by consensus of three radiation oncologists. The proposed dual-head model integrated classification and ordinal regression and was evaluated with cross-center validation, ablation analysis, AI-versus-clinician comparison, and cross-cancer validation.

Study Design and Patient Cohorts

This study aimed to develop and validate a deep learning model for automated grading of radiation dermatitis in breast cancer radiotherapy as an objective tool for assessing skin toxicity during treatment. A total of 1,021 patients from multiple centers were included. The primary cohort consisted of 833 patients with breast cancer prospectively enrolled at Xijing Hospital and was used for model training, internal validation, ablation analysis, and AI-versus-clinician comparison. Data from 143 patients with breast cancer from Xi'an International Medical Center Hospital served as an independent external validation cohort. An additional 45 patients with radiation dermatitis related to nasopharyngeal carcinoma radiotherapy from Xijing Hospital were included as a cross-cancer validation cohort. This study was approved by the Ethics Committee of Xijing Hospital (approval No. KY20242059-F-1). Written informed consent was obtained from all patients, and the study was conducted in accordance with the Declaration of Helsinki and relevant medical ethics guidelines.

In the primary cohort, the mean age of patients with breast cancer was 48.91 ± 10.03 years. In the independent external validation cohort, the mean age was 49.22 ± 10.42 years. Both breast cancer cohorts included only women. In the cross-cancer validation cohort, patients with nasopharyngeal carcinoma had a mean age of 49.58 ± 10.45 years, including 37 men (82.22%) and 8 women (17.78%). Radiotherapy regimens were broadly similar between the two breast cancer cohorts, with 50 Gy in 25 fractions being the most common regimen, accounting for 48.50% (404/833) and 45.45% (65/143), respectively. The second most common regimen was 40.05 Gy in 15 fractions plus a 10.68 Gy boost in 4 fractions, accounting for 33.49% (279/833) and 38.46% (55/143), followed by 50 Gy in 25 fractions plus a 10 Gy boost in 5 fractions, accounting for 13.93% (116/833) and 11.89% (17/143). Other less common regimens accounted for 4.08% (34/833) and 4.20% (6/143), respectively. Patients in the nasopharyngeal carcinoma cohort mainly received definitive concurrent chemoradiotherapy. Prescription doses were 70-72 Gy in 33 fractions to the primary tumor and 68-70 Gy in 33 fractions to metastatic lymph nodes. Detailed cohort characteristics are shown in Table 1.

Table 1. *Baseline characteristics of the primary cohort, independent external breast cancer validation cohort, and cross-cancer cohort.*

Variables	PC (n = 833)	EBC (n = 143)	NPC-C (n = 45)
Age, Mean ± SD	48.91 ± 10.03	49.22 ± 10.42	49.58 ± 10.45
Gender			
Male	0 (0)	0 (0)	37 (82.22)
Female	833 (100)	143 (100)	8 (17.78)
Radiation dose, n (%)			
50Gy/25F	404 (48.50)	65 (45.45)	-
40.05Gy/15F+10.68Gy/4F	279 (33.49)	55 (38.46)	-
50Gy/25F+10Gy/5F	116 (13.93)	17 (11.89)	-
Others	34 (4.08)	6 (4.20)	-

Abbreviations: *PC, primary cohort; EBC, independent external breast cancer validation cohort; NPC-C, nasopharyngeal carcinoma cross-cancer cohort.*

Reference Standard

Radiation dermatitis severity was graded according to the RTOG criteria. Skin photographs acquired during radiotherapy were independently reviewed by three radiation oncologists, and consensus grading was used as the reference standard. Cases with discrepant assessments were discussed until a final grade was assigned. Because the number of high-grade cases was limited and the present study focused on the more common image-level grading scenario in clinical practice, radiation dermatitis was ultimately grouped into three grades: grade 0, grade 1, and grade 2 or higher. Grade 0 indicated no obvious skin reaction, grade 1 indicated mild skin reaction, and grade 2 or higher indicated more substantial skin injury. The automated grading task was therefore formulated as a three-class classification problem.

Image Processing and Data Augmentation

All images were acquired prospectively by uniformly trained radiation therapists under a standardized protocol with a fixed background, standardized shooting distance, and as consistent an angle as possible. To minimize external influences, patients were instructed not to apply local products, including cosmetics, moisturizers, or cleansers, for at least 8 hours before photography. Skin photographs were obtained at prespecified treatment time points corresponding to the 5th, 10th, 15th, 20th, and 25th fractions. All images were manually cropped by radiation oncologists to a 60 × 60 mm region containing the most prominent dermatitis findings, thereby reducing background distraction and emphasizing local skin phenotypes. Grade 0 images were cropped using the same procedure.

Before model input, images were resized to 384 × 384 pixels by bilinear interpolation and normalized using the same mean and standard deviation as the ImageNet pretrained model. To mitigate class imbalance, grade 0, 1, and 2 images in the training set were augmented to different extents, resulting in 2,418, 2,421, and 2,418 samples, respectively. Data augmentation included horizontal and vertical flipping, random rotation within ±15°, and perturbations of brightness and contrast to improve robustness to variation in camera angle and illumination. No data augmentation was applied outside the training cohort. The final numbers of images in each cohort and grade are summarized in Figure 2.

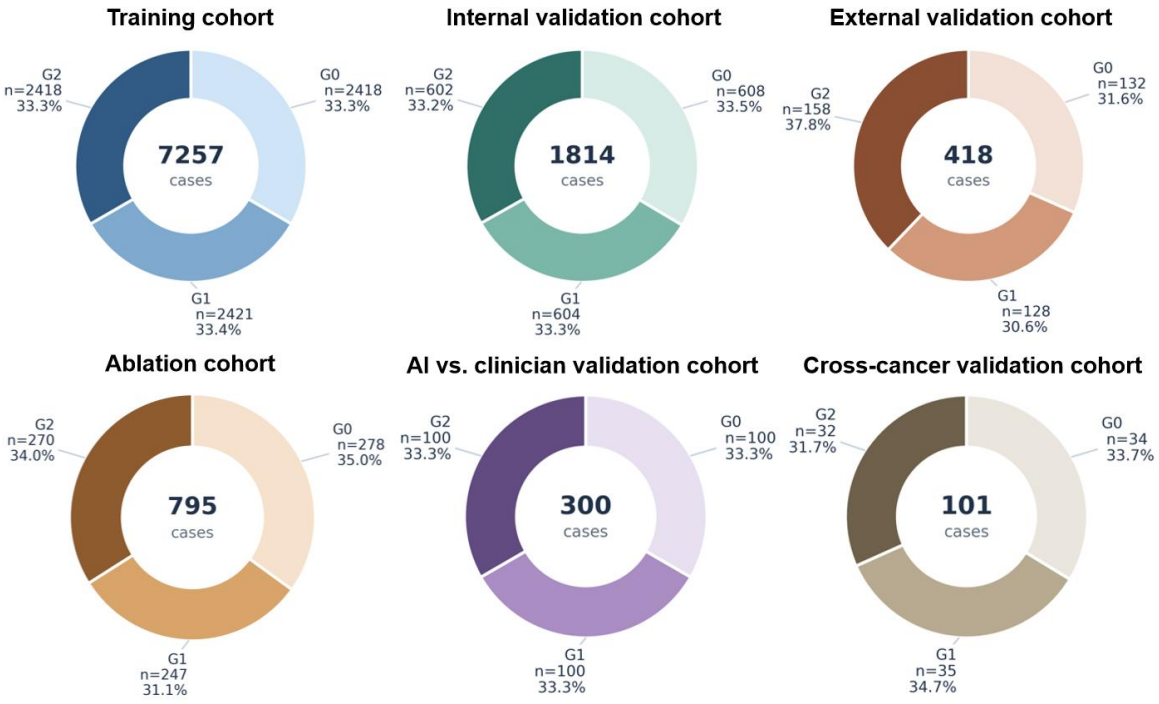


Figure 2. *Distribution of radiation dermatitis images across the study cohorts.*

Overall Network Architecture

We developed a dual-head deep learning framework for automated grading of radiation dermatitis in breast cancer radiotherapy. The model consists of a shared feature-extraction backbone and two parallel prediction heads. After cropped and normalized local skin photographs were fed into the network, the backbone extracted multilevel visual representations, which were then passed to a CE head and an ordinal regression head to jointly perform standard multiclass discrimination and severity-aware ordered prediction [29]. This design was motivated by the fact that radiation dermatitis grades are not isolated categories but ordered representations of increasing skin injury severity. Although a standard three-class classifier can directly output labels 0, 1, and 2, it does not exploit the intrinsic ordering among grades. We therefore retained conventional classification while introducing an ordinal branch to better reflect the logic of clinical grading.

EfficientNet-V2-L was used as the backbone [30]. As shown in Figure 3, the network includes a 3 × 3 convolutional stem, three Fused-MBConv stages, four MBConv stages, a 1 × 1 convolutional head, global average pooling, and a linear projection layer. The early Fused-MBConv modules improve the extraction of superficial texture and color features, whereas the later MBConv modules strengthen higher-level semantic modeling, which is well suited to capturing subtle erythema, pigmentation change, desquamation, and textural heterogeneity in radiation dermatitis images. Compared with traditional backbones such as ResNet or DenseNet, EfficientNet-V2-L provides a favorable balance between parameter efficiency and fine-grained phenotypic representation and was therefore used as the shared feature extractor in this study.

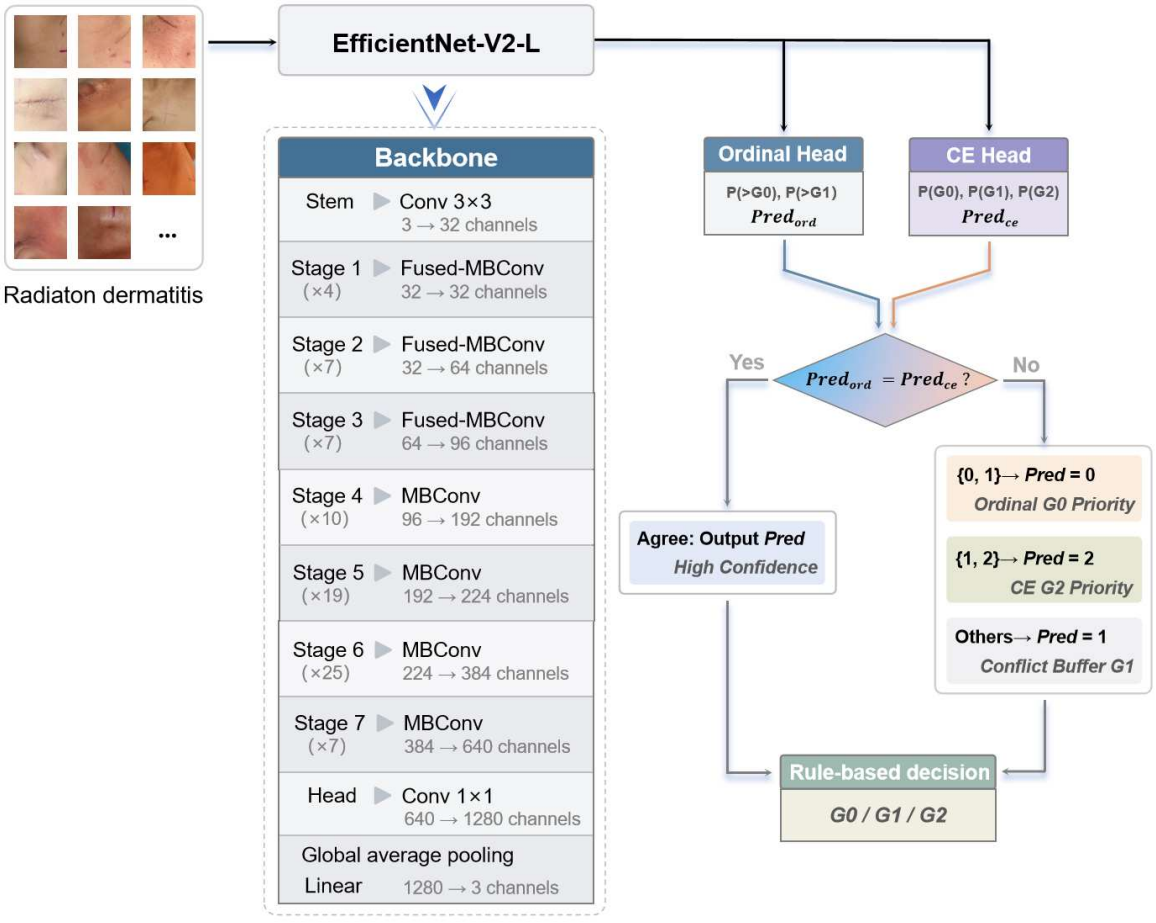


Figure 3. *Architecture of the proposed dual-head automated grading framework with rule-based fusion. CE, cross-entropy; $Pred_{ce}$, prediction from the cross-entropy classification head; $Pred_{ord}$, prediction from the ordinal regression head; G0, grade 0; G1, grade 1; G2, grade 2.*

Dual-Head Prediction and Rule-Based Fusion Strategy

Following shared feature extraction, the model generates predictions from two heads. The CE head outputs probabilities for the three classes and performs multiclass discrimination among grades 0, 1, and 2. Its main advantage is the direct learning of class boundaries. The ordinal head reformulates the three-class task as two sequential binary decisions that determine whether a case reaches a higher severity threshold. In other words, the ordinal head does not directly answer "which class is it?" but rather "has the case crossed the current grade threshold?" This design allows the model to learn both interclass differences and the ordered progression from mild to severe disease, thereby reducing clinically implausible nonadjacent errors. Specifically, the CE head outputs three class probabilities, and the class with the highest probability is taken as the CE prediction. The ordinal head outputs two probabilities corresponding to grade 1 or higher and grade 2 or higher, respectively. Rather than using fixed empirical cutoffs, we performed grid search on the internal validation set to determine the optimal two thresholds under an ordering constraint to preserve logical consistency in grading. The final threshold combination was selected according to the performance of the fused model on the internal validation set. The two ordinal probabilities were then converted into a discrete grade prediction: cases exceeding the higher threshold were assigned to grade 2 or higher; cases below the higher threshold but above the lower threshold were assigned to grade 1; and cases below both thresholds were assigned to grade 0. This allowed the ordinal head to generate predictions in the same label space as the CE head.

After predictions from the CE and ordinal heads were obtained, we designed a rule-based fusion strategy to generate the final grade (Figure 3). Although the dual-head framework learns both class boundaries and ordinal structure, the two heads may still disagree on boundary cases. Selecting one head alone as the final output would discard potentially useful information from the other and would not fully exploit the advantage of joint supervision. We therefore integrated the two predictions at inference to improve stability and clinical consistency. If the two heads produced the same grade, that grade was used as the final output. If they disagreed, predetermined rules were applied according to the type of discordance. When the ordinal head predicted grade 0 and the CE head predicted grade 1, the final output was set to grade 0. This scenario usually corresponds to very subtle or equivocal skin change, and the ordinal head was considered more conservative at the lower threshold, where the key question is whether dermatitis is present at all. When the ordinal head predicted grade 1 and the CE head predicted grade 2, the final output was set to grade 2. Because higher-grade skin toxicity has clearer clinical implications, the CE head was considered more sensitive to overt severe phenotypes in this setting, and the final decision was therefore shifted toward the higher grade. All other discordant combinations were mapped to grade 1. This strategy allowed the final prediction to combine the direct category discrimination of the CE head with the ordered severity information encoded by the ordinal head within a single decision framework.

Model Training Details

Model training followed the framework described above. All input images were resized to 384 × 384 pixels after processing. Losses for the CE head and ordinal head were equally weighted. The optimizer was AdamW with an initial learning rate of 1×10^{-4} and weight decay of 1×10^{-4}. Learning rate scheduling used CosineAnnealingWarmRestarts, with T_0 set to 10 times the number of iterations per epoch and T_{mult} set to 2. The batch size was 8, and gradient accumulation over 4 steps yielded an effective batch size of 32. Models were trained for 100 epochs with a random seed of 42 for reproducibility.

At the end of each epoch, model performance was evaluated on the internal validation set, and metrics for the CE head, ordinal head, and fused output were recorded. The best model was selected on the basis of macro-F1 for the fused output on the internal validation set. Early stopping used a patience of 50 epochs to avoid premature termination. Total loss and branch-specific losses were saved throughout training for monitoring

and troubleshooting. All experiments were conducted in Python 3.13.5 on a computer equipped with an Intel Core i9-12900K CPU, 64 GB of RAM, and an NVIDIA GeForce RTX 4090 GPU with 24 GB of memory.

Baseline Classification Models

To benchmark the proposed dual-head framework against conventional single-head classifiers, we trained multiple convolutional neural networks as comparators: DenseNet121, DenseNet169, DenseNet201, DenseNet264, ResNet18, ResNet34, ResNet50, and ResNet101, and a single-head classifier built on the EfficientNet-V2-L backbone. For the DenseNet and ResNet series, input images were resized to the same resolution as the proposed model. The initial learning rate was 1×10^{-5}. The StepLR scheduling was applied with 0.8 decay every 10 epochs, and the maximum number of epochs was 100. The batch size was 4. All comparator models were trained with cross-entropy loss. To ensure a fair backbone-level comparison, we also trained a single-head EfficientNet-V2-L classifier using the same backbone, input size, and training settings as the proposed model, with the output head changed to a conventional single classification head.

Evaluation and Validation

Model performance was evaluated from four aspects: overall discrimination, grading agreement, cross-dataset generalizability, and interpretability. Standard metrics included overall accuracy, macro-recall, macro-F1, and the classification report [31-33]. Because radiation dermatitis has an intrinsic ordinal structure, we additionally used quadratic weighted kappa (QWK) to assess agreement between model predictions and the reference standard [34].

Validation experiments included internal validation, independent external validation, cross-cancer validation, ablation analysis, AI-versus-clinician comparison, and interpretability analysis. Internal validation was performed on the validation subset of the primary cohort and was used for model selection, threshold search, and performance assessment. Independent external validation used the external breast cancer cohort to evaluate cross-center generalizability within the same cancer type. Cross-cancer validation used radiation dermatitis images from nasopharyngeal carcinoma to evaluate transferability across anatomy, skin phenotype, and radiotherapy context. In addition, an AI-versus-clinician cohort was used to compare the grading performance between the AI system and clinicians. Ablation experiments were designed to quantify the contributions of the CE head, ordinal head, and rule-based fusion to overall performance. For interpretability analysis, Grad-CAM was used to visualize heatmaps from the CE head, ordinal head, and fused output to identify image regions that contributed most to grading across different classes [35].

Results

Performance in the Internal and Independent External Validation Cohorts

In the internal validation cohort, the proposed model outperformed all comparator models across overall metrics (Figure 4A). The proposed model achieved an accuracy, macro-F1, and QWK of 0.947, 0.947, and 0.958, respectively, representing the best overall performance among all models. Class-specific analysis showed F1 scores of 0.967, 0.923, and 0.951 for G0, G1, and G2, respectively (Figure 4B), all higher than those of the single-head comparators. Notably, G1 was the most challenging class across all models, yet the proposed model still achieved the highest F1 score in this class, suggesting that dual-head supervision and rule-based fusion are particularly helpful for boundary cases between adjacent grades. One-versus-rest receiver operating characteristic (ROC) analysis also showed excellent discriminatory performance, with AUCs of 0.996, 0.986, and 0.995 for G0, G1, and G2, respectively (Figure 4C). The confusion matrix further showed that misclassifications were concentrated between adjacent grades (Figure 4D), indicating that the model maintained high accuracy while largely avoiding clinically implausible skipped-grade errors.

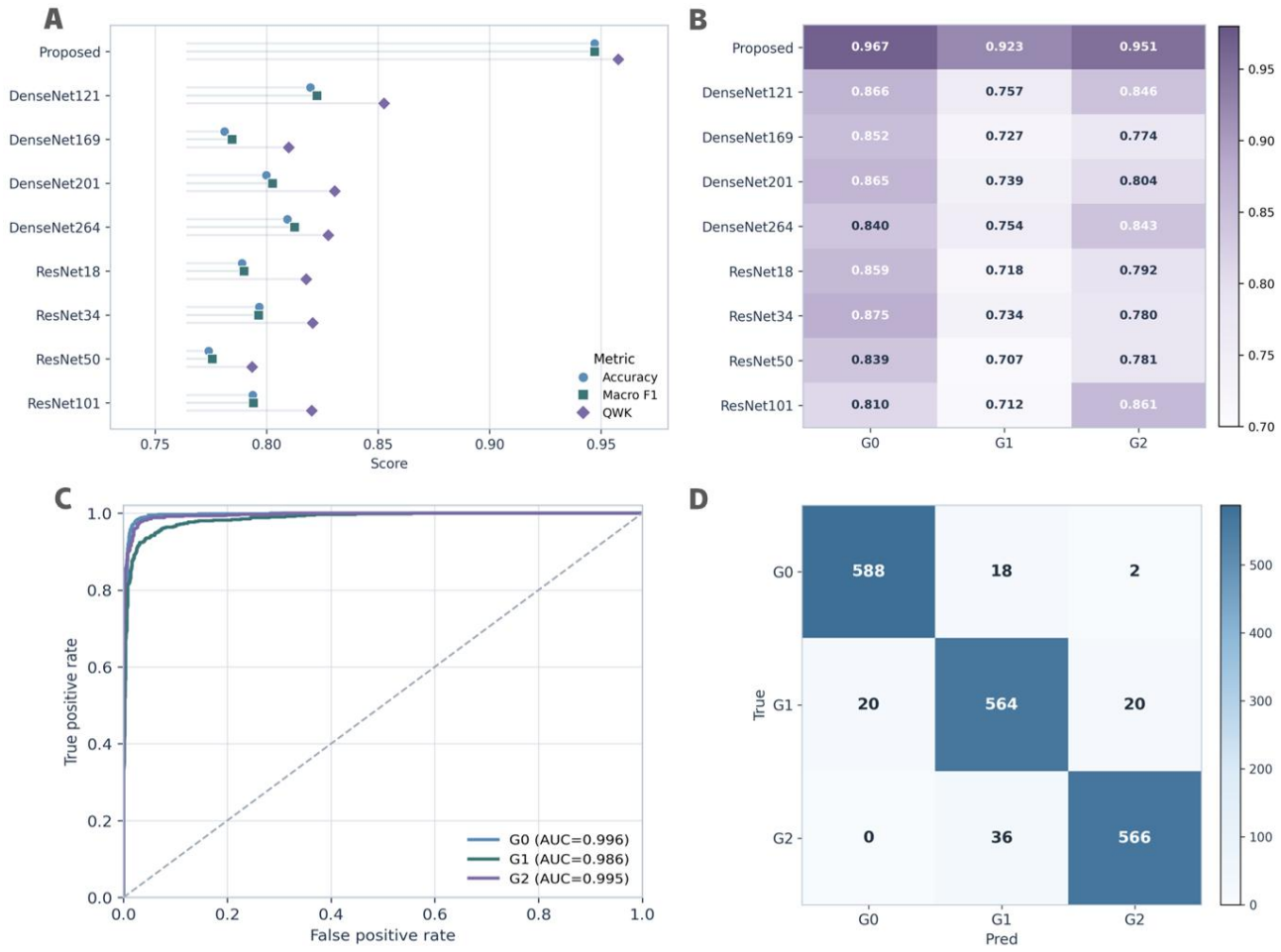


Figure 4. *Performance of the proposed model and comparator models in the internal validation cohort. (A) Overall performance across models. (B) Heatmap of class-specific F1 scores. (C) One-versus-rest ROC curves for the proposed model. (D) Confusion matrix for the proposed model.*

In the independent external validation cohort, performance was slightly lower than that in the internal validation cohort but remained stable overall, and the proposed model continued to outperform all comparators (Figure 5A). Accuracy, macro-F1, and QWK were 0.909, 0.908, and 0.918, respectively. Class-specific F1 scores were 0.936, 0.866, and 0.922 for G0, G1, and G2, respectively (Figure 5B), all higher than those of the comparator models, indicating good grading performance in a cross-center setting. ROC analysis showed AUCs of 0.992, 0.971, and 0.990 for G0, G1, and G2, respectively (Figure 5C), indicating preserved discriminatory ability on external data. The external validation confusion matrix also showed that errors were concentrated primarily between adjacent grades, with only 1 G2 case misclassified as G0 and 2 G0 cases misclassified as G2 (Figure 5D), further supporting the robustness of the model across centers.

To provide a more complete view of model behavior across cohorts, confusion matrices and ROC curves for all comparator models in the internal validation cohort are shown in Supplementary Figure S1 and S2, respectively, and the corresponding results for the independent external validation cohort are shown in Supplementary Figure S3 and S4. Overall, the proposed model showed consistent performance advantages in both internal and external validation, with error patterns largely confined to adjacent grades, supporting its reliability and generalizability for automated grading of radiation dermatitis in breast cancer radiotherapy.

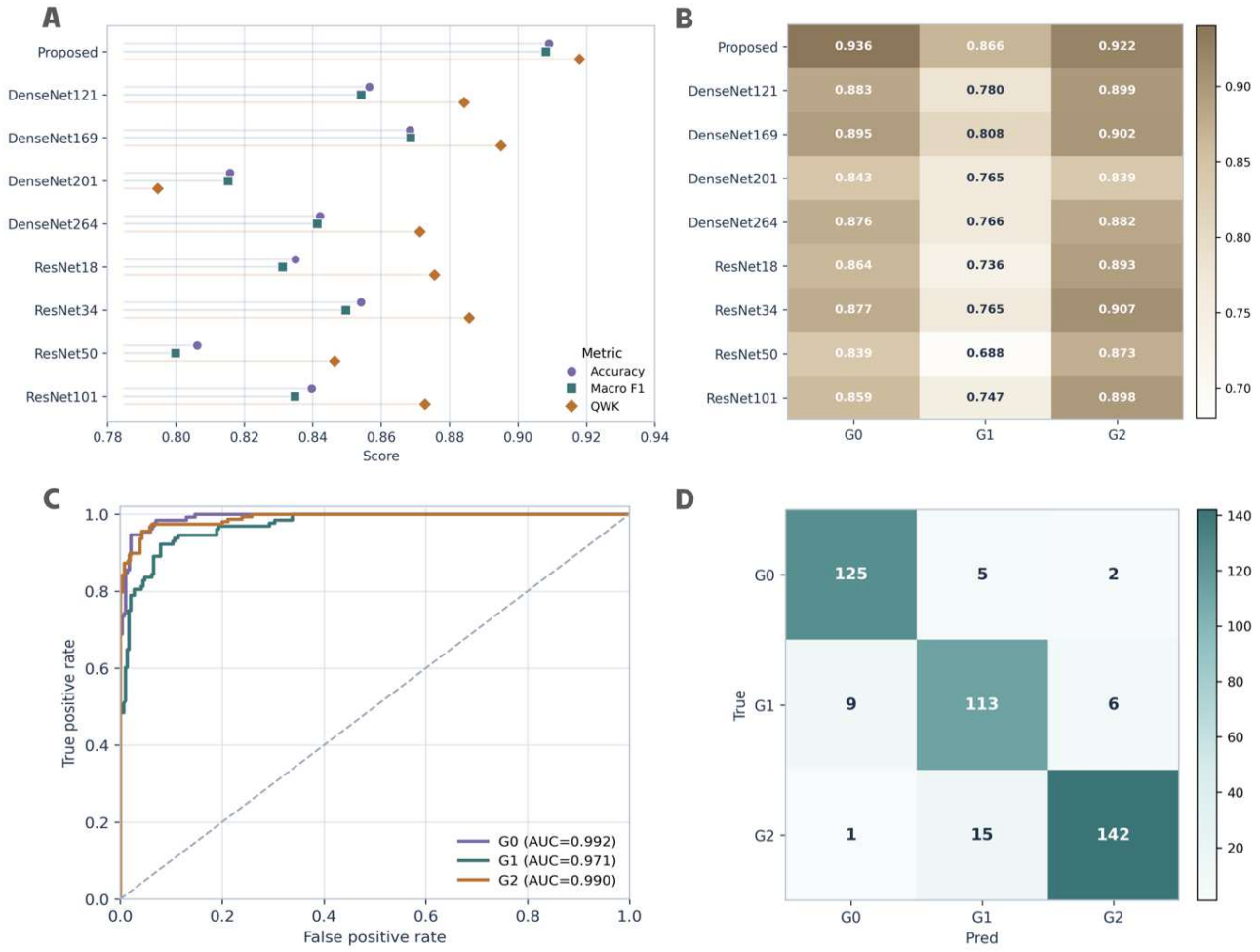


Figure 5. *Performance of the proposed model and comparator models in the independent external validation cohort. (A) Overall performance across models. (B) Heatmap of class-specific F1 scores. (C) One-versus-rest ROC curves for the proposed model. (D) Confusion matrix for the proposed model.*

Ablation Analysis of Dual-Head Supervision and Rule-Based Fusion

To further evaluate the contribution of dual-head supervision and rule-based fusion, we performed a systematic ablation analysis (Figure 6). The fusion model outperformed all ablation variants, including the CE head alone, the ordinal head alone, and the dual-head model without rule-based fusion. Specifically, the proposed model achieved an accuracy, macro-F1, and QWK of 0.917, 0.916, and 0.936, compared with 0.914, 0.913, and 0.935 for the dual-head model, 0.879, 0.878, and 0.903 for the ordinal head alone, and 0.886, 0.884, and 0.906 for the CE head alone (Figure 6A). These findings suggest that the CE and ordinal heads provide complementary information on top of the shared backbone and that rule-based fusion yields additional benefit. Class-specific analysis supported the same conclusion: the proposed model achieved F1 scores of 0.933, 0.871, and 0.944 for G0, G1, and G2, respectively, outperforming both the CE head (0.908, 0.826, 0.918) and ordinal head (0.914, 0.815, 0.904), and showing further improvement over the dual-head model without rule-based fusion (0.933, 0.864, 0.941) (Figure 6C). G1 remained the most difficult class, but its F1 score improved from 0.864 to 0.871 after fusion, which was the best among all configurations. This finding suggests that integrating information from both heads is important for classifying intermediate-grade cases.

In terms of misclassification burden, the proposed model produced 21, 27, and 18 errors in G0, G1, and G2, respectively. Compared with the dual-head model without rule-based fusion, the number of G0 errors

remained unchanged (21 in both models), G1 errors decreased from 34 to 27, and G2 errors increased from 13 to 18. Nevertheless, the proposed model maintained better overall control of errors in three grades than either the CE or ordinal head alone (Figure 6B). These results suggest that the main gain from rule-based fusion lies in reducing uncertainty for boundary cases in the intermediate class while preserving overall class balance. In the case-level analysis of discordant predictions, rule-based integration of the CE and ordinal heads made the final output more consistent (Figure 6D). The confusion matrices showed the strongest diagonal concentration for the proposed model, particularly in G1, where the number of correctly classified cases reached 220, compared with 213 for the dual-head model, 205 for the ordinal head, and 206 for the CE head. Recognition of G2 also remained high at 252 correctly classified cases (Figure 6E). Together, these results indicate that dual-head supervision improves the use of ordinal grading information and that rule-based fusion further refines the final decision for boundary cases.

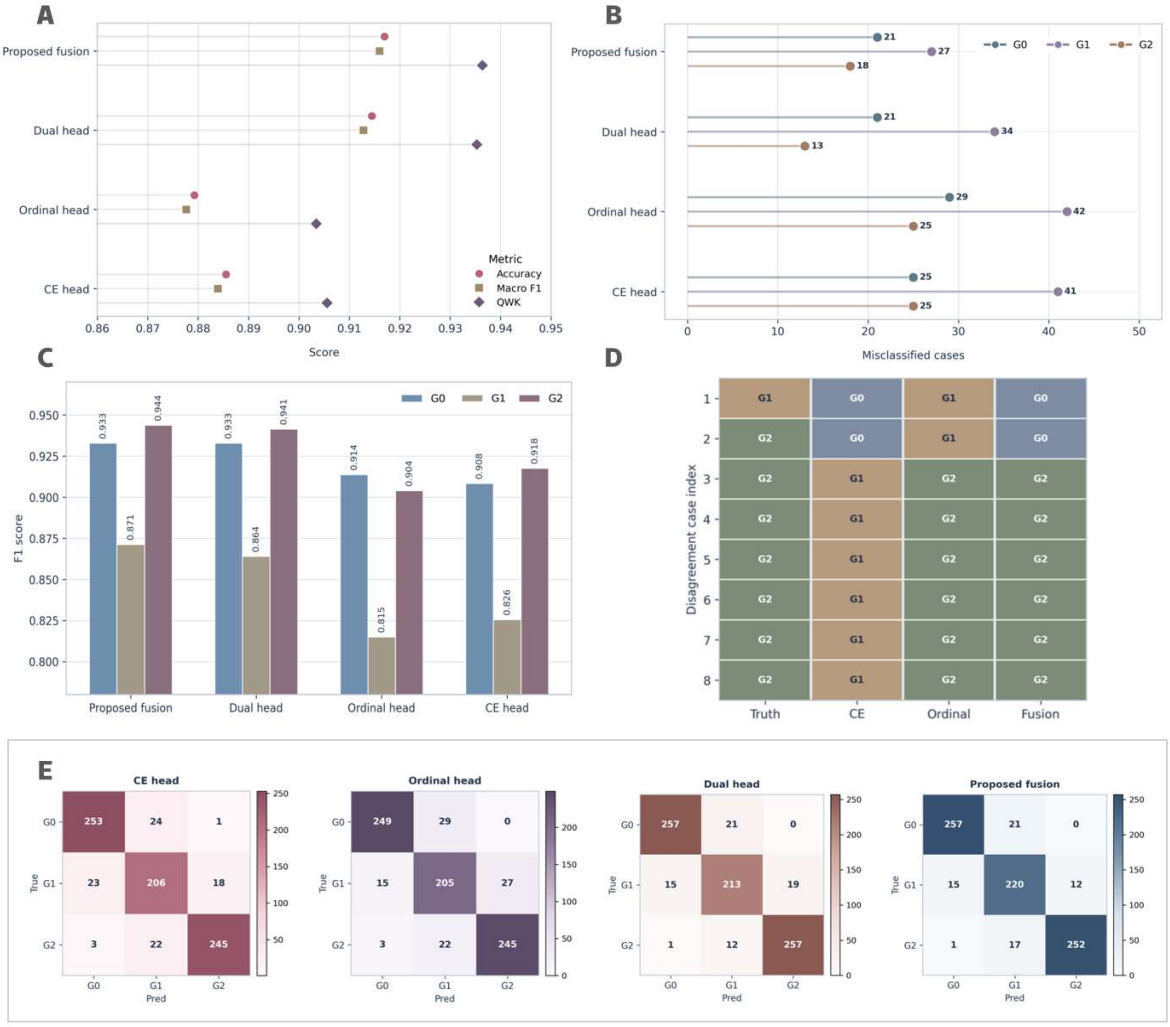


Figure 6. *Performance comparison of the proposed model and ablation variants. (A) Overall performance. (B) Error burden by true grade. (C) Class-specific F1 scores. (D) Case-level output patterns in the disagreement subset. (E) Confusion matrices of the ablation variants.*

To better understand the focus patterns of different models, we compared visual responses of the proposed model and ablation variants across representative cases from different grades (Figure 7). In representative G0 cases, the proposed model showed relatively restricted activation, mainly near local edges or marker regions, whereas the CE head showed broader and more diffuse responses, suggesting less nonspecific activation in the

proposed model. In G1 and G2 cases, both the proposed model and the dual-head model without rule-based fusion were able to focus on lesion-related regions such as erythema, desquamation, and pigmentation change. However, the proposed model showed more complete and accurate coverage, with better correspondence to the visible lesion extent. By contrast, the ordinal head tended to show fragmented responses, whereas the CE head more often exhibited background interference or diffuse high-activation regions. Additional examples for G0, G1, and G2 are provided in Supplementary Figure S5-S7. These showed patterns consistent with the main results and further support the notion that the proposed fusion model provides more stable and clinically relevant attention across grades.

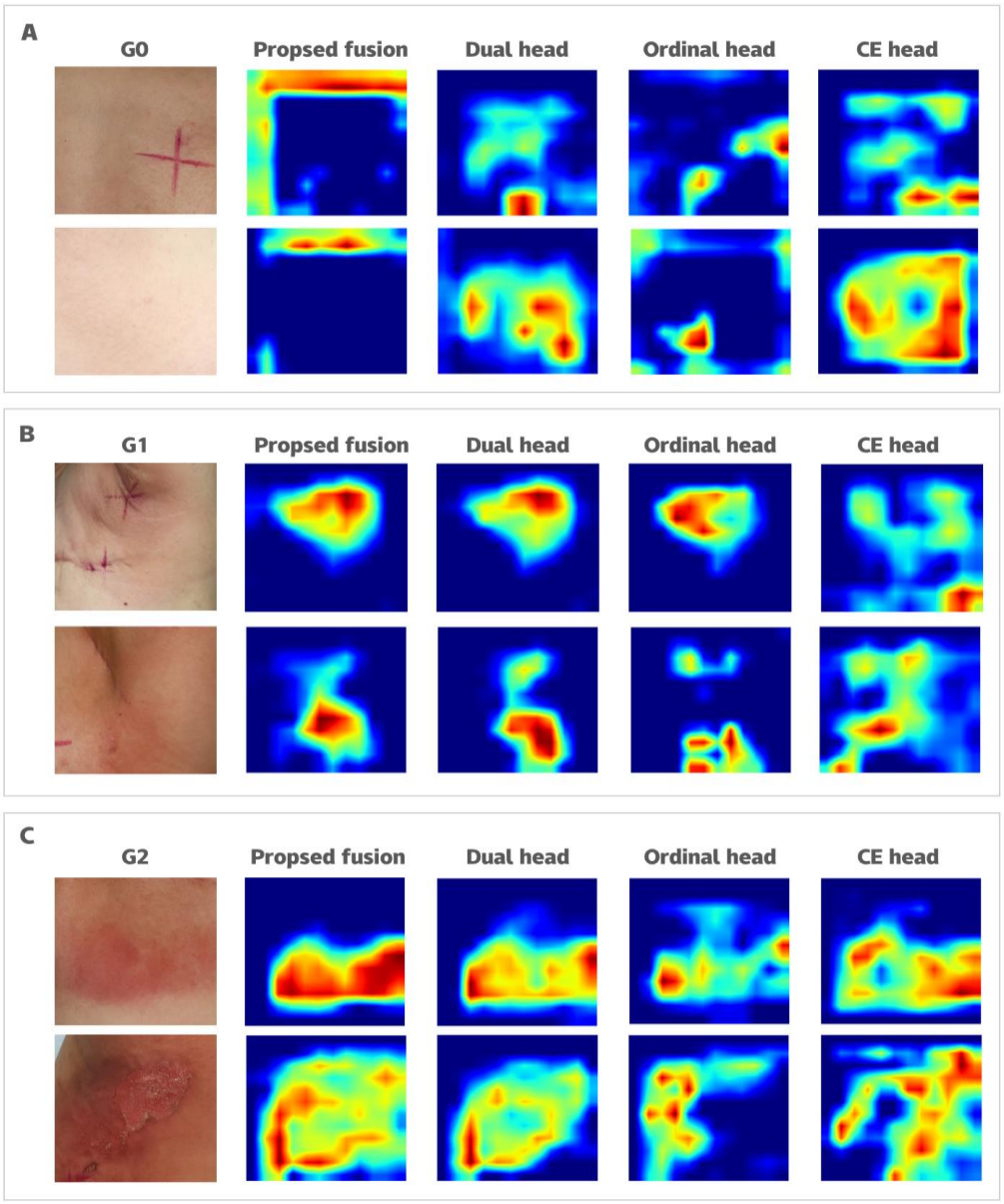


Figure 7. *Representative Grad-CAM visualizations of the proposed model and ablation variants for grade 0 (A), grade 1 (B), and grade 2 (C) radiation dermatitis.*

Comparison Between the AI Model and Clinicians

In the AI-versus-clinician cohort, the proposed model outperformed all three clinicians (Figure 8). The AI system achieved an overall accuracy of 93.3%, higher than that of the senior clinician (Clinician 1, 88.7%)

and the two junior clinicians (Clinician 2, 83.3%; Clinician 3, 80.3%) (Figure 8A). Errors made by the AI system were concentrated between adjacent grades, with a one-step error rate of 6.7%, lower than the corresponding rates for Clinician 1 (11.3%), Clinician 2 (16.7%), and Clinician 3 (19.3%). The AI system showed no two-step errors, whereas Clinician 3 showed a 0.3% two-step error rate. These findings suggest that the proposed model not only achieved higher overall accuracy but also avoided clearly implausible misclassifications that violate clinical grading logic.

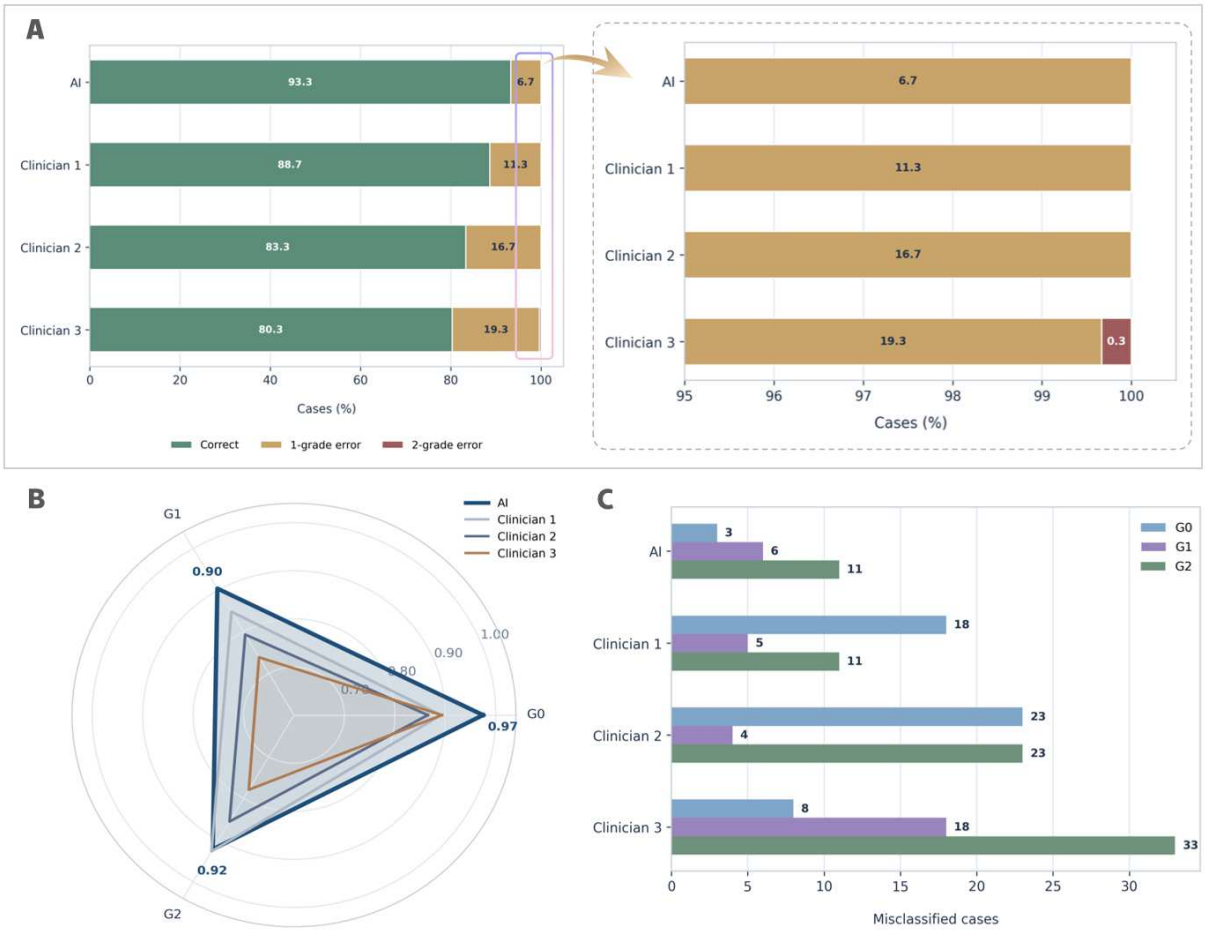


Figure 8. *Comparison of grading performance between the proposed model and three clinicians. (A) Overall correct classification and error-step distribution. (B) Radar plot of class-specific F1 scores. (C) Error burden by true grade.*

Subgroup analysis by dermatitis grade further supported these findings. The radar plot showed F1 scores of 0.97, 0.90, and 0.92 for AI in G0, G1, and G2, respectively, which were higher than those of all three clinicians (Figure 8B). Among the clinicians, Clinician 1 performed most similarly to AI, whereas the two junior clinicians showed greater variation across grades, particularly for G1 and G2. Overall, AI performance was more balanced across grades, indicating more consistent automated grading. In case-level error counts, AI produced only 3 errors in G0, fewer than Clinician 1, Clinician 2, and Clinician 3 (18, 23, and 8 errors, respectively) (Figure 8C). For G2, AI misclassified 11 cases, equal to Clinician 1 and markedly fewer than Clinicians 2 and 3 (23 and 33 cases, respectively). For G1, AI misclassified 6 cases, close to Clinician 1 (5 cases) and Clinician 2 (4 cases), and substantially fewer than Clinician 3 (18 cases). These findings suggest that the main advantage of AI lies in more stable identification of no dermatitis cases, while maintaining performance for mild and severe dermatitis that is comparable to that of an experienced clinician.

Cross-Cancer Validation in Nasopharyngeal Carcinoma

To examine model transportability, we applied the model trained on radiation dermatitis in breast cancer radiotherapy to a cohort of radiation dermatitis associated with nasopharyngeal carcinoma radiotherapy. In this cross-cancer validation cohort, the proposed fusion model maintained clearly better overall performance than either the CE head alone or the ordinal head alone (Figure 9A). Specifically, the fusion model achieved a QWK, macro-precision, macro-recall, macro-F1, and accuracy of 0.907, 0.895, 0.893, 0.882, and 0.891, respectively, substantially higher than the corresponding values for the ordinal head (0.322, 0.456, 0.471, 0.374, and 0.455) and the CE head (0.001, 0.103, 0.337, 0.164, and 0.317). These findings indicate that the proposed framework retained substantial overall classification capability despite changes in cancer type, anatomy, and skin phenotype.

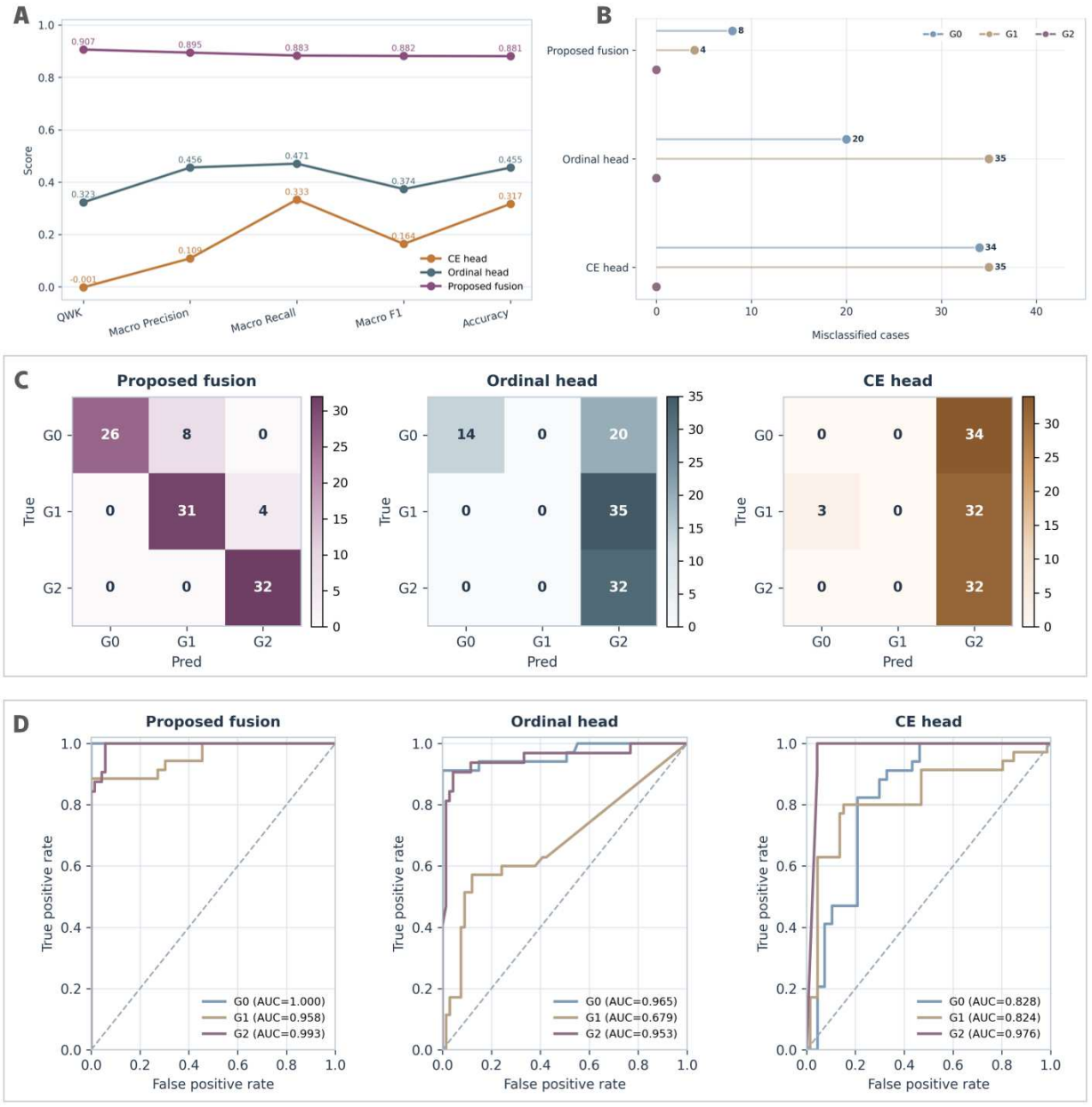


Figure 9. *Quantitative performance of the proposed model in the nasopharyngeal carcinoma cross-cancer validation cohort. (A) Cross-cancer validation metrics. (B) Error burden by true grade. (C) Confusion matrices. (D) One-versus-rest ROC curves.*

Misclassifications in the cross-cancer validation cohort were again concentrated between adjacent grades (Figure 9B, C). The proposed model produced 8, 4, and 0 errors in G0, G1, and G2, respectively. By comparison, the ordinal head misclassified 20 G0 cases and 35 G1 cases, whereas the CE head misclassified 34 G0 cases and 35 G1 cases, indicating much higher error rates for low- and intermediate-grade cases. The confusion matrix showed that the proposed model correctly classified all 32 G2 cases, as well as 26 of 34 G0 cases and 31 of 35 G1 cases. Remaining errors were mainly G0 cases classified as G1 or G1 cases classified as G2, with no obvious nonadjacent errors inconsistent with clinical grading logic. In contrast, the ordinal head tended to assign large numbers of G0 and G1 cases to G2, whereas the CE head was largely unable to identify G0 and G1 and showed a pronounced bias toward higher grades. ROC analysis further supported these findings (Figure 9D). AUCs for the proposed model were 1.000, 0.958, and 0.993 for G0, G1, and G2, respectively, all higher than those of the ordinal head (0.965, 0.679, and 0.953) and the CE head (0.828, 0.824, and 0.976). The advantage of the proposed model was most apparent in G1, supporting the value of rule-based fusion for intermediate-grade cases in the cross-cancer setting. Overall, these results suggest that although the model was trained on breast cancer radiotherapy, it retained promising generalizability in radiation dermatitis associated with nasopharyngeal carcinoma.

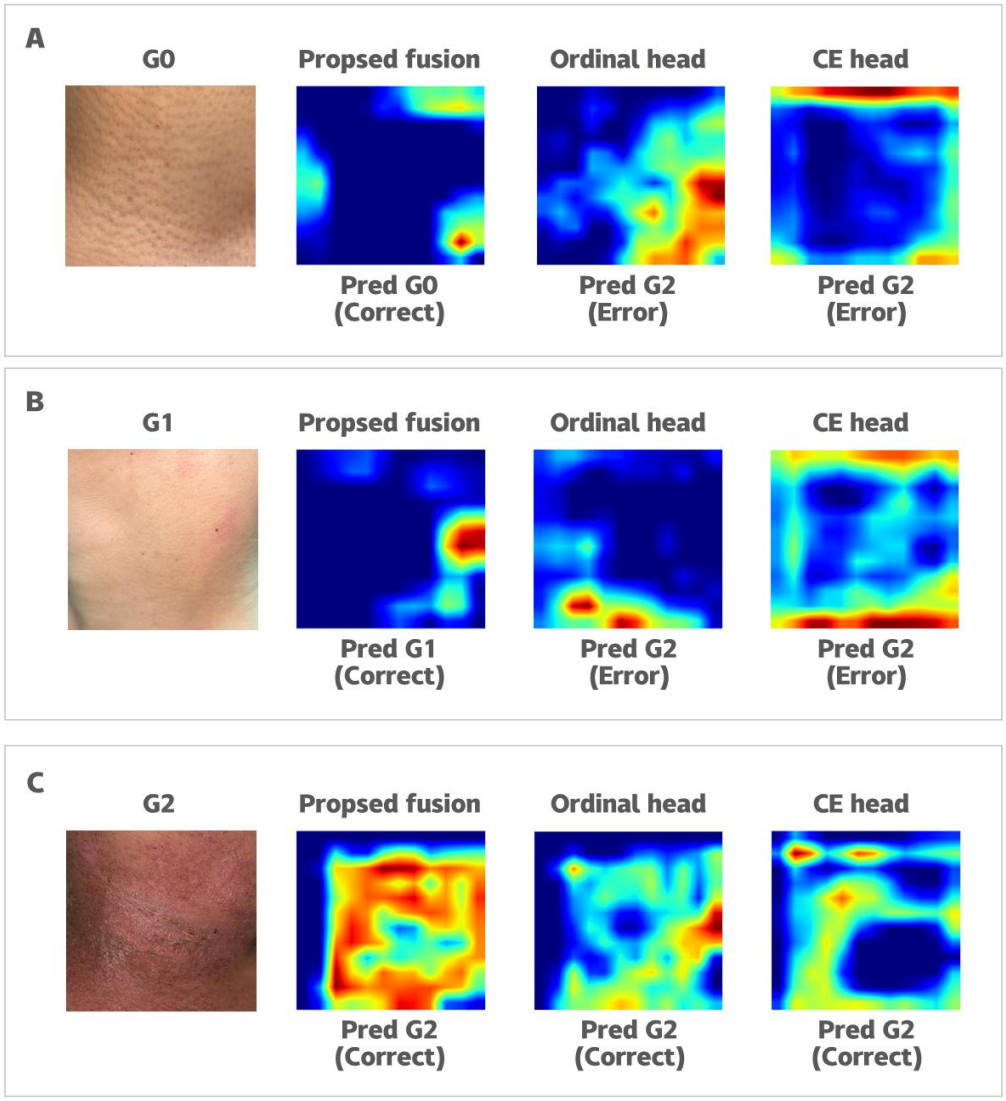


Figure 10. *Representative activation maps for grade 0 (A), grade 1 (B), and grade 2 (C) radiation dermatitis in the nasopharyngeal carcinoma cross-cancer validation cohort.*

Visualization results further illustrated differences in attention patterns across models in the cross-cancer setting (Figure 10). In representative G0 and G1 cases, the proposed model produced correct predictions, whereas both the ordinal and CE heads misclassified them as G2. The heatmaps showed that the proposed model yielded more restricted responses, concentrated mainly around local edges or regions associated with subtle skin change, whereas the single heads showed broader and more diffuse activation, suggesting greater susceptibility to background texture in cross-cancer transfer. In representative G2 cases, all three models produced the correct class label, but the proposed model showed closer correspondence between activated regions and the visible lesion extent. Taken together, the visualization results were consistent with the quantitative findings and further support more stable and anatomically appropriate attention patterns for the proposed model in the cross-cancer setting.

Discussion

In this multicenter study, we developed and validated a deep learning framework for automated grading of radiation dermatitis in breast cancer radiotherapy. Using skin photographs acquired during treatment, the model classified radiation dermatitis into grade 0, grade 1, and grade 2 or higher. Its performance was evaluated through internal validation, independent external validation, AI-versus-clinician comparison, ablation analysis, and cross-cancer validation. Overall, the proposed model outperformed multiple single-head convolutional neural network baselines in both the internal and external breast cancer cohorts, exceeded the performance of three clinicians in the AI-versus-clinician comparison, and maintained promising transportability in the nasopharyngeal carcinoma cohort. These findings support the technical feasibility of image-based automated grading of radiation dermatitis and its potential to improve the objectivity and consistency of toxicity assessment.

A consistent observation across internal validation, independent external validation, and AI-versus-clinician comparison was that G1 was the most challenging class. This is clinically plausible. Mild radiation dermatitis represents a transition state between normal skin and unequivocal lesion. Its phenotype is often subtle and may be influenced by baseline skin tone, local anatomy, and lighting conditions during photography, all of which also contribute to interobserver variability in manual assessment. Accordingly, more than overall accuracy, the ability to identify G1 reflects how well a model handles boundary cases. In the present study, the proposed model still achieved the highest F1 score in G1 among all models, and misclassifications were concentrated primarily between adjacent grades. This suggests that the gain in performance did not simply arise from better recognition of typical G0 or G2 cases, but also from more stable handling of borderline cases. The practical value of dual-head supervision and rule-based fusion lies precisely here: retaining standard multiclass discrimination while explicitly incorporating the ordered nature of clinical grading, thereby reducing clearly implausible skipped-grade errors.

From a methodological perspective, the performance improvement was driven less by the backbone itself than by the complementarity created by dual-head supervision and rule-based fusion. The CE head directly learns class boundaries among grades, whereas the ordinal head explicitly encodes the ordered progression of radiation dermatitis through stepwise decisions about whether a case reaches a higher severity threshold. For a clinical task such as dermatitis grading, this design better reflects how grading is actually performed: the process is not a purely mechanical separation of three independent categories, but rather a discretized expression of a continuous judgment of skin injury severity. The ablation analysis supports this interpretation. Neither the CE head nor the ordinal head alone achieved the best performance. Introducing both forms of supervision on top of a shared backbone improved results, and adding rule-based fusion yielded the best accuracy, macro-F1, and QWK. Importantly, the main gain was observed in boundary cases, particularly for G1. This suggests that the value of the dual-head design lies not merely in adding an extra branch, but in

retaining ordinal constraints alongside class discrimination, while the fusion strategy translates the two sources of information into a final decision that is more stable and clinically interpretable.

Two additional findings have clear clinical relevance. First, in the AI-versus-clinician comparison, the proposed model outperformed three radiation oncologists overall. Its performance was closest to that of the senior clinician, but it provided better overall accuracy and better control of misclassification. This suggests a role for the model as a consistency aid and decision-support tool. In scenarios in which skin toxicity must be repeatedly documented during radiotherapy, the potential value of such a system lies in reducing subjective variation across readers and making toxicity assessment more reproducible. Second, we performed cross-cancer validation by applying a model trained on breast cancer radiotherapy to radiation dermatitis associated with nasopharyngeal carcinoma. Despite substantial differences in anatomy, irradiation background, and skin phenotype between the two settings, the proposed model remained clearly superior to the CE and ordinal heads alone. This suggests that the framework captured visual features of radiation dermatitis that are not entirely specific to breast cancer. Cross-cancer validation remains uncommon in prior studies of automated radiation dermatitis grading and extends the present study beyond a single-center, single-cancer setting.

Several issues should be addressed before the framework can be applied more broadly in clinical practice. First, the current model was trained mainly on locally cropped skin photographs acquired under relatively standardized conditions. This design helps reduce background interference and improve internal consistency, but it also means that the model remains dependent on a relatively controlled acquisition workflow. Future work should evaluate robustness under more natural clinical conditions, including different devices, shooting distances, lighting environments, and more complex backgrounds. Second, although EfficientNet-V2-L achieved favorable grading performance, the model remains relatively large and may be demanding for deployment in resource-constrained environments. Exploration of lighter architectures, knowledge distillation, or parameter compression may be helpful for practical deployment on mobile devices. Third, the present framework is image-based and does not yet incorporate dosimetry variables, treatment fractionation, symptoms, or patient-reported outcomes. Because radiation dermatitis is a toxicity endpoint closely related to temporal evolution and local dose exposure, integrating images with dose, time, and symptom information may further improve recognition of borderline cases and dynamic progression. Finally, although the study included an independent external breast cancer cohort and an additional cross-cancer validation cohort, the overall data source remains limited, particularly for cross-cancer validation. Broader validation in larger, multicenter datasets and in more heterogeneous clinical settings will be needed to support wider application across tumor types, treatment sites, and clinical environments.

Conclusion

In summary, we developed and validated a deep learning framework for automated grading of radiation dermatitis in breast cancer radiotherapy. A model integrating cross-entropy, ordinal regression, and rule-based fusion showed favorable grading performance in internal and external validation, AI-versus-clinician comparison, and cross-cancer testing. These findings suggest that augmenting conventional classification with ordinal constraints and decision-level fusion can improve the stability and clinical consistency of automated radiation dermatitis grading, and provide a useful technical basis for more objective assessment of skin toxicity during radiotherapy.

Declarations

Ethics approval and consent to participate

This study was approved by the Ethics Committee of Xijing Hospital (approval No. KY20242059-F-1). Written informed consent was obtained from all participants before enrollment. All procedures involving human participants were conducted in accordance with the Declaration of Helsinki.

Consent for publication

Not applicable.

Availability of data and materials

The datasets used during the current study are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the National Natural Science Foundation of China (No. 82404071 and No. 12505420), and supported by the Shaanxi Provincial Health Research Innovation Capacity Enhancement Program Funding (2025PT-02).

Authors' contributions

Conceptualization: HF Sun. and J Gong. **Methodology:** HF Sun, QL Huang and YX Cui. **Validation:** YX Cui, C Zhang, F Meng and J Guo. **Formal analysis:** HF Sun, B Lv and GJ Chai. **Investigation:** HF Sun, ZH Shi, J Li and Y Cui. **Resources:** LN Zhao and J Guo. **Data curation**: C Zhang, J Guo, Y Huang, WJ Zhang, MX Liu and XM Wang. **Writing-original draft preparation:** HF Sun and ZH Shi. **Writing-review and editing:** J Gong. **Visualization:** Y Tian and YF Liu. **Supervision:** J Gong and LN Zhang. Funding acquisition: HF Sun, J Gong and LN Zhao.

Acknowledgements

Not applicable.

References

1. Bray F, Laversanne M, Sung H, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2024;74(3):229-263.
2. Plevritis SK, Munoz D, Kurian AW, et al. Association of Screening and Treatment With Breast Cancer Mortality by Molecular Subtype in US Women, 2000-2012. JAMA. 2018;319(2):154-164.
3. Lawton TJ. Update on the Use of Molecular Subtyping in Breast Cancer. Adv Anat Pathol. 2023;30(6):368-373.
4. Zeng H, Zheng R, Sun K, et al. Cancer survival statistics in China 2019-2021: a multicenter, population-based study. J Natl Cancer Cent. 2024;4(3):203-213.
5. Giaquinto AN, Sung H, Newman LA, et al. Breast cancer statistics 2024. CA Cancer J Clin. 2024;74(6):477-495.
6. Chua BH. Omission of radiation therapy post breast conserving surgery. Breast. 2024;73:103670.
7. EBCTCG (Early Breast Cancer Trialists' Collaborative Group), McGale P, Taylor C, et al. Effect of radiotherapy after mastectomy and axillary surgery on 10-year recurrence and 20-year breast cancer mortality: meta-analysis of individual patient data for 8135 women in 22 randomised trials. Lancet. 2014;383(9935):2127-2135.
8. Speers C, Pierce LJ. Postoperative radiotherapy after breast-conserving surgery for early-stage breast cancer: a review. JAMA Oncol. 2016;2(8):1075-1082.

9. Loibl S, André F, Bachelot T, et al. Early breast cancer: ESMO Clinical Practice Guideline for diagnosis, treatment and follow-up. Ann Oncol. 2024;35(2):159-182.
10. Early Breast Cancer Trialists' Collaborative Group (EBCTCG), Darby S, McGale P, et al. Effect of radiotherapy after breast-conserving surgery on 10-year recurrence and 15-year breast cancer death: meta-analysis of individual patient data for 10,801 women in 17 randomised trials. Lancet. 2011;378(9804):1707-1716.
11. Xie Y, Wang Q, Hu T, et al. Risk Factors Related to Acute Radiation Dermatitis in Breast Cancer Patients After Radiotherapy: A Systematic Review and Meta-Analysis. Front Oncol. 2021;11:738851.
12. Kole AJ, Kole L, Moran MS. Acute radiation dermatitis in breast cancer patients: challenges and solutions. Breast Cancer (Dove Med Press). 2017;9:313-323.
13. Ramseier JY, Ferreira MN, Leventhal JS. Dermatologic toxicities associated with radiation therapy in women with breast cancer. Int J Womens Dermatol. 2020;6(5):349-356.
14. Zhu G, Shi Q, Cai T, et al. CACA guidelines for holistic integrative management of anticancer treatment-induced cutaneous adverse events. Holist Integr Oncol. 2024;3:33.
15. Rahman SM, Patel P, Finkelstein S, et al. Delphi consensus-based recommendations of tools and outcomes for the assessment of radiation dermatitis severity. Radiother Oncol. 2025;206:110846.
16. Behroozian T, Milton LT, Shear NH, et al. Radiation dermatitis assessment tools used in breast cancer: a systematic review of measurement properties. Support Care Cancer. 2021;29(5):2265-2278.
17. Bontempo PSM, Menêses AG, Ciol MA, Ferreira EB, dos Reis PED. Instruments and scales for the evaluation of acute radiation dermatitis: a systematic review. Crit Rev Oncol Hematol. 2023;191:104116.
18. Patel P, Rahman SM, Behroozian T, et al. Systematic review of the tools and outcomes for the assessment of acute radiation dermatitis severity. Clin Transl Radiat Oncol. 2025;53:100977.
19. Huang CJ, Hou MF, Luo KH, et al. RTOG, CTCAE and WHO criteria for acute radiation dermatitis correlate with cutaneous blood flow measurements. Breast. 2015;24(3):230-236.
20. Behroozian T, Milton L, Zhang L, et al. How do patient-reported outcomes compare with clinician assessments? A prospective study of radiation dermatitis in breast cancer. Radiother Oncol. 2021;159:98-105.
21. Cilla S, Romano C, Macchia G, et al. Machine-learning prediction model for acute skin toxicity after breast radiation therapy using spectrophotometry. Front Oncol. 2023;12:1044358.
22. Feng H, Wang H, Xu L, et al. Prediction of radiation-induced acute skin toxicity in breast cancer patients using data encapsulation screening and dose-gradient-based multi-region radiomics technique: a multicenter study. Front Oncol. 2022;12:1017435.
23. Aldraimli M, Osman S, Grishchuck D, et al. Development and optimization of a machine-learning prediction model for acute desquamation after breast radiation therapy in the multicenter REQUITE cohort. Adv Radiat Oncol. 2022;7(3):100890.
24. Saadatmand P, Mahdavi SR, Nikoofar A, et al. A dosiomics model for prediction of radiation-induced acute skin toxicity in breast cancer patients: machine learning-based study for a closed bore linac. Eur J Med Res. 2024;29(1):282.
25. Lee TF, Chang CH, Chi CH, et al. Utilizing radiomics and dosiomics with AI for precision prediction of radiation dermatitis in breast cancer patients. BMC Cancer. 2024;24(1):965.
26. Park YI, Choi SH, Hong CS, et al. A New Approach to Quantify and Grade Radiation Dermatitis Using Deep-Learning Segmentation in Skin Photographs. Clin Oncol (R Coll Radiol). 2023;35(1):e10-e19.
27. Ni R, Zhou T, Ren G, et al. Deep learning-based automatic assessment of radiation dermatitis in patients with nasopharyngeal carcinoma. Int J Radiat Oncol Biol Phys. 2022;113(3):685-694.
28. Shi Z, Zhang C, Liu Y, et al. Artificial intelligence-based automated grading and severity prediction of radiation dermatitis: a review. Radiat Oncol. 2025;21(1):17.

29. Cao W, Mirjalili V, Raschka S. Rank consistent ordinal regression for neural networks with application to age estimation. Pattern Recognit Lett. 2020;140:325-331.
30. Tan M, Le QV. EfficientNetV2: Smaller Models and Faster Training. In: Proceedings of the 38th International Conference on Machine Learning (ICML). 2021;139:10096-10106.
31. Miller C, Portlock T, Nyaga DM, O'Sullivan JM. A review of model evaluation metrics for machine learning in genetics and genomics. Front Bioinform. 2024;4:1457619.
32. Rainio O, Teuho J, Klén R. Evaluation metrics and statistical tests for machine learning. Sci Rep. 2024;14(1):6086.
33. Kocak B, Klontzas ME, Stanzione A, et al. Evaluation metrics in medical imaging AI: fundamentals, pitfalls, misapplications, and recommendations. Eur J Radiol Artif Intell. 2025;3:100030.
34. Vanbelle S. A new interpretation of the weighted kappa coefficients. Psychometrika. 2016;81(2):399-410.
35. Fayyaz AM, Abdulkadir SJ, Talpur N, Al-Selwi SM, Hassan SU, Sumiea EH. Grad-CAM (Gradient-weighted Class Activation Mapping): a systematic literature review. Comput Biol Med. 2025;198(Pt B):111200.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementary.docx