Original Research Article

Augmented Decision-Making in Wound Care: Evaluating the Clinical Utility of a Deep-Learning Model for Pressure Injury Staging

Jemin Kim[a*], Changyoon Lee[b*], Sungchul Choi[b], Da-In Sung[b], Jeonga Seo[b], Yun Na Lee[c], Joo Hee Lee[c], Eun Jin Han[d], Ah Young Kim[d], Hyun Suk Park[d], Hye Jeong Jung[d], Jong Hoon Kim[e], and Ju Hee Lee[c†]

[a]*Department of Dermatology, Yongin Severance Hospital, Yonsei University College of Medicine, Gyeonggi-do, Korea*

[b]*Department of Medicine, Yonsei University College of Medicine, Seoul, Korea*

[c]*Department of Dermatology and Cutaneous Biology Research Institute, Severance Hospital, Yonsei University College of Medicine, Seoul, Korea*

[d]*Department of Nursing, Severance Hospital, Seoul, Korea*

[e]*Department of Dermatology and Cutaneous Biology Research Institute, Gangnam Severance Hospital, Yonsei University College of Medicine, Seoul, Korea*

[*]These authors contributed equally to this work.

[†]**Corresponding author**

Ju Hee Lee, M.D., Ph.D.

Department of Dermatology and Cutaneous Biology Research Institute,

Yonsei University College of Medicine, 50-1 Yonsei Ro, Seodaemun-gu, Seoul 03722, Korea

E-mail: juhee@yuhs.ac

Tel: 82-2-2228-2080

Fax: 82-2-393-6947

Manuscript word count: 2,531; **Figure Count**: 4; **Table Count**: 2

Abstract

Background: Precise categorization of pressure injury (PI) stages is critical in determining the appropriate treatment for wound care. However, the expertise necessary for PI staging is frequently unavailable in residential care settings.

Objective: This study aimed to develop a convolutional neural network (CNN) model for classifying PIs and investigate whether its implementation can allow physicians to make better decisions for PI staging.

Methods: Using 3,098 clinical images (2,614 and 484 from internal and external datasets, respectively), a CNN was trained and validated to classify PIs and other related dermatoses. A two-part survey was conducted with 24 dermatology residents, ward nurses, and medical students to determine whether the implementation of the CNN improved initial PI classification decisions.

Results: The top-1 accuracy of the model was 0.793 (95% confidence interval [CI], 0.778–0.808) and 0.717 (95% CI, 0.676–0.758) over the internal and external testing sets, respectively. The accuracy of PI staging among participants was 0.501 (95% CI, 0.487–0.515) in Part I, improving by 17.1% to 0.672 (95% CI, 0.660–0.684) in Part II. Furthermore, the concordance between participants increased significantly with the use of the CNN model, with Fleiss' κ of 0.414 (95% CI, 0.410–0.417) and 0.641 (95% CI, 0.638–0.644) in Parts I and II, respectively.

Conclusions: The proposed CNN model can help classify PIs and relevant dermatoses. In addition, augmented decision-making can improve consultation accuracy while ensuring concordance between the clinical decisions made by a diverse group of health professionals.

Keywords: pressure injury staging, wound care, convolutional neural network, augmented decision-making

1. Introduction

Pressure injuries (PIs) occur on the skin and underlying tissue, typically over a bony prominence, and may present as an open wound or intact skin caused by intense or prolonged pressure or a combination of pressure and shear [1]. Despite a considerable focus on treating PIs in hospitals and long-term care facilities, they remain a significant problem and have emerged as a healthcare burden, especially for older adults with limited mobility [2, 3]. Since 1989, the National Pressure Injury Advisory Panel (NPUAP) has developed a staging system that classifies and describes PIs; this system has been widely adopted among healthcare providers [4]. Although proper assessment of PIs is crucial for treatment planning, first-line caregivers, especially non-specialists, typically find accurate PI staging challenging [5].

Recently, automated classification systems have been developed using convolutional neural networks (CNNs) to analyze medical images, focusing on skin cancers [6-8]. Augmented decision-making via the implementation of CNNs has been reported to improve the diagnostic performance of physicians in detecting malignant skin diseases [9, 10]. Deep-learning techniques have been used for wound detection, segmentation, tissue type classification, and severity grading of PIs [3, 11-15]. Nevertheless, research on the potential of artificial intelligence (AI) in improving the assessment of PIs by healthcare providers, especially non-specialists, remains limited.

Therefore, to address this knowledge gap, this study aimed to develop a CNN model that can classify the grading of PIs using medical images. In addition, we examined the effectiveness of this model in enhancing the accuracy of PI classification by clinical practitioners.

2. Methods

2.1 Dataset and preprocessing

This retrospective analysis involved clinical images of PIs and relevant dermatoses collected from January 2017 to September 2022 at two referral hospitals (Severance Hospital and Gangnam Severance Hospital) in Seoul, South Korea, with cases of suspected PIs or relevant dermatoses in risk-prone areas (e.g., sacrum, greater trochanter, heel, and occiput) diagnosed by the dermatology and plastic surgery departments or wound nurses included in the analysis. This study was approved with waivers of informed consent by the Institutional Review Board of both institutions (approval numbers 4-2021-1076 and 3-2022-0411). Images were obtained not in strictly controlled conditions but rather from a plethora of clinical settings encountered in inpatient care. Hence, various devices including cellphones, tablets, and digital cameras were used to capture the images, which accentuates the model's practical applicability in varied clinical and informal settings.

A total of 2,614 images from 493 subjects were included in the main dataset, whereas the external dataset consisted of 484 images from 137 subjects. We compiled the data from Severance Hospital as the main dataset, allocating 80% for model training, 10% for validation, and 10% for internal testing. All data (100%) from Gangnam Severance Hospital were considered part of the external testing set (Table S1). The images were labeled according to the NPUAP criteria presented in 2016 [1], which includes six stages (Stage I, Stage II, Stage III, Stage IV, deep tissue injury (DTI), or unstageable), with an additional category of "others," which encompasses relevant dermatoses of PIs, such as folliculitis, herpes simplex, superficial fungal infection, or incontinence-associated dermatitis. Four certified wound nurses with more than ten years of experience and one board-certified dermatologist independently labeled the entire dataset, with the majority vote designated as the "gold standard" label of the image. In cases where the voting results were divided, the

professionals convened and reviewed the relevant image and consultation report to arrive at a consensus and assign a single label.

2.2.CNN model and training

We used the SE-ResNext101 architecture pretrained on the ImageNET dataset and achieved top-1 and top-5 accuracies of 83.6% and 96.69%, respectively, for 1000 classes [16, 17]. Supplementary text S1 and Figure S1 describe the architecture and processes of the CNN model. The images were manually cropped to include only one lesion per image before being resized to 512×512 and normalized to a pixel intensity range of [0,1] to be used as the model input. Images showing multiple lesions were cropped separately and used as distinct images with corresponding labels. Data augmentation techniques, including rotation from -30° to 30°, translation (shifting), and addition of random Gaussian noise to the images, were applied to the input images to improve the robustness of the model and to address imbalances in dataset representation, especially for Stage I and III images. In addition, a stratified K-fold cross-validation approach was implemented to maintain the proportionality of samples for each class across every fold, thus ensuring a more balanced training and validation set.

2.3 Human evaluators and decision study

Twenty-four participants, comprising eight dermatology residents, eight ward nurses with less than five years of experience, and eight final-year medical students, were recruited to investigate the ability of healthcare providers in classifying PIs using clinical images and potential performance improvement with the assistance of a CNN model. We randomly selected 250 images from the internal testing dataset, presented them as original-resolution photographs, and asked the participants to select the most appropriate classification. An anonymous online-based questionnaire was conducted in two parts over four-week intervals using Google Survey (Fig. S2). In Part I, the participants were asked to make their

classification decisions based solely on the clinical images. In Part II, the participants were provided with the top-3 predictions of ulcer staging and confidence scores predicted by the CNN model. Participant performance was assessed by comparing their predictions and associated decisions with the gold standard label; case sequences were shuffled to ensure unbiased responses in each part, with the reference diagnosis for each case and participant scores not disclosed until the end of the study.

2.4 Outcome measurement and statistics

Ten-fold stratified cross-validation was performed to verify the robustness of the best-fit model, with CNN performance evaluated in terms of the top-1 accuracy, sensitivity, specificity, and F1 score. Receiver operating characteristic (ROC) curves were drawn using the sensitivity and specificity for each threshold, and areas under the curve (AUCs) were calculated. The 95% confidence intervals (CIs) were calculated through bootstrap resampling of the test dataset with the replacement N = 1000 times.[18] The accuracy, sensitivity, and specificity of the participants were determined for each part, with their performances compared by conducting the McNemar test (before vs. after assistance of the algorithm). Fleiss' kappa (κ) values [19] were used to evaluate the agreement among participants' responses, with a heatmap generated using hierarchical agglomerative clustering used to visualize the inter-participant agreement rate. The statistical analyses were performed using Python version 3.9.0 and R version 4.2.2, and p-values less than 0.05 were considered statistically significant.

3. **Results**

3.1 Model performance

Table 1 summarizes the sensitivity, specificity, F1 score, ROC-AUC, and top-1 accuracy values of the deep neural network model. The overall model accuracy was 0.793 (95% CI:

0.778–0.808) over the internal testing set, decreasing slightly to 0.717 (95% CI: 0.676–0.758) over the external testing set. The sensitivity, specificity, F1 score, and ROC-AUC of each severity class exhibited considerable variations. In particular, the sensitivities of Stage IV (0.884 [95% CI: 0.853–0.916]) and others (0.992 [95% CI: 0.984–0.998]) were higher than those of Stage I, Stage III, and DTI, with values of 0.7 over the internal testing set. The sensitivity of Stage III over the external testing set was approximately 30% lower than that for the internal testing set. Specifically, its sensitivity decreased from 0.463 (95% CI: 0.383–0.547) to 0.167 (95% CI: 0.000–0.389). In contrast to the sensitivity, the per-class prediction specificity for all PI classes was uniformly high, exceeding 0.90 for both the internal and external sets.

3.2 Performance of human evaluators

The performance of 24 participants in classifying 250 images of PIs is summarized in Table 2. Figure 1 shows the ROC curves and plots for each PI class. In Part I, without assistance from the CNN model, the participants demonstrated an accuracy of 0.501 (95% CI: 0.487–0.515), a sensitivity of 0.500 (95% CI: 0.488–0.513), and a specificity of 0.927 (95% CI: 0.925–0.929). With few exceptions, the deep learning model outperformed most of the participants in classifying each stage of the PIs.

In Part II, the participants were provided with the top-3 predictions of the neural network model and the corresponding probability. The accuracy, sensitivity, and specificity values for Part II were 0.672 (95% CI: 0.660–0.684), 0.672 (95% CI: 0.661–0.684), and 0.944 (95% CI: 0.941–0.946), respectively. Compared with Part I, Part II was characterized by a significant increase in the accuracy and sensitivity by 17.2%p (95% CI: 15.8–18.6%p; $p<0.001$) and 17.2%p (95% CI: 15.7–18.7%p; $p<0.001$), respectively. The specificity increased slightly by 1.7%p (95% CI: 1.4–2.0%p; $p<0.001$). The improvement was significant for the relatively inexperienced groups of final-year medical students and hospital nurses, as indicated by the

noticeable increase in their accuracy after receiving assistance from the CNN model (18.8%p [95% CI: 16.1–21.5%p] and 19.3%p [95% CI: 16.9–21.9%p], respectively).

Table 1. Performance of the classification model according to pressure injury staging

Model/Class	Sensitivity (95% CI)	Specificity (95% CI)	F1-score (95% CI)	ROC-AUC (95% CI)
Internal testing set				
Stage I	0.659 (0.590-0.727)	0.973 (0.967-0.979)	0.648 (0.591-0.702)	0.944 (0.927-0.959)
Stage II	0.769 (0.728-0.808)	0.943 (0.933-0.952)	0.743 (0.709-0.777)	0.937 (0.926-0.948)
Stage III	0.463 (0.383-0.547)	0.983 (0.977-0.987)	0.527 (0.453-0.597)	0.909 (0.883-0.932)
Stage IV	0.884 (0.853-0.916)	0.972 (0.965-0.979)	0.868 (0.841-0.891)	0.980 (0.973-0.987)
DTI	0.643 (0.585-0.706)	0.967 (0.959-0.973)	0.657 (0.604-0.704)	0.936 (0.921-0.950)
Unstageable	0.728 (0.693-0.762)	0.925 (0.913-0.936)	0.736 (0.707-0.763)	0.912 (0.899-0.925)
Others	0.992 (0.984-0.998)	0.992 (0.988-0.996)	0.984 (0.977-0.991)	0.998 (0.996-0.999)
Average[1]	0.793 (0.778-0.809)	0.962 (0.958-0.965)	0.791 (0.776-0.807)	0.951 (0.945-0.956)
External testing set				
Stage I	0.623 (0.491-0.738)	0.941 (0.918-0.964)	0.613 (0.514-0.710)	0.892 (0.848-0.933)
Stage II	0.691 (0.624-0.758)	0.928 (0.894-0.955)	0.767 (0.714-0.815)	0.919 (0.893-0.943)
Stage III	0.167 (0.000-0.389)	0.970 (0.953-0.983)	0.171 (0.000-0.343)	0.840 (0.764-0.907)
Stage IV	0.970 (0.893-1.000)	0.969 (0.952-0.985)	0.810 (0.704-0.898)	0.995 (0.990-0.999)
DTI	0.543 (0.395-0.689)	0.961 (0.941-0.977)	0.568 (0.452-0.686)	0.898 (0.847-0.949)
Unstageable	0.700 (0.578-0.811)	0.927 (0.901-0.952)	0.632 (0.524-0.727)	0.910 (0.861-0.949)
Others	1.000 (1.000-1.000)	0.963 (0.946-0.980)	0.909 (0.857-0.955)	0.999 (0.999-1.000)
Average[1]	0.717 (0.674-0.754)	0.943 (0.930-0.954)	0.715 (0.674- 0.755)	0.927 (0.908-0.944)

[1]*Calculated by the weighted-average value of each severity class for the given model, with bootstrap resampling (N=1000) of the test dataset.*

Abbreviations: CI: confidence interval; DTI: deep tissue injury; ROC-AUC: area under the receiver operating characteristic curve.

Table 2. Overall performance of human evaluators

Outcome/Participant type	Performance (95% CI)		Difference (95% CI)	
	Part I (Image only)	Part II (Image + AI assist)	Part II – Part I	P-value[1]
Accuracy				
All evaluators (*n* = 24)	0.501 (0.487-0.515)	0.672 (0.660-0.684)	0.172 (0.158-0.186)	<0.001
Dermatology residents (*n* = 8)	0.531 (0.508-0.554)	0.666 (0.646-0.686)	0.135 (0.109-0.159)	<0.001
Final-year medical students (*n* = 8)	0.489 (0.467-0.509)	0.677 (0.658-0.696)	0.188 (0.161-0.215)	<0.001
Hospital nurses (*n* = 8)	0.482 (0.462-0.504)	0.675 (0.656-0.694)	0.193 (0.169-0.219)	<0.001
Sensitivity				
All evaluators (*n* = 24)	0.500 (0.488-0.513)	0.672 (0.661-0.684)	0.172 (0.157-0.187)	<0.001
Dermatology residents (*n* = 8)	0.531 (0.510-0.552)	0.666 (0.645-0.687)	0.135 (0.109-0.162)	<0.001
Final-year medical students (*n* = 8)	0.489 (0.467-0.513)	0.676 (0.655-0.696)	0.187 (0.163-0.214)	<0.001
Hospital nurses (*n* = 8)	0.481 (0.460-0.505)	0.674 (0.653-0.696)	0.193 (0.167-0.217)	<0.001
Specificity				
All evaluators (*n* = 24)	0.927 (0.925-0.929)	0.944 (0.941-0.946)	0.017 (0.014-0.020)	<0.001
Dermatology residents (*n* = 8)	0.933 (0.929-0.937)	0.944 (0.941-0.948)	0.011 (0.007-0.016)	<0.001
Final-year medical students (*n* = 8)	0.927 (0.923-0.931)	0.946 (0.942-0.950)	0.019 (0.013-0.023)	<0.001
Hospital nurses (*n* = 8)	0.920 (0.916-0.925)	0.941 (0.937-0.945)	0.021 (0.016-0.025)	<0.001

[1]*McNemar test was performed for the paired values (Part I vs. Part II).*

Abbreviations: CI: confidence interval; AI: artificial intelligence

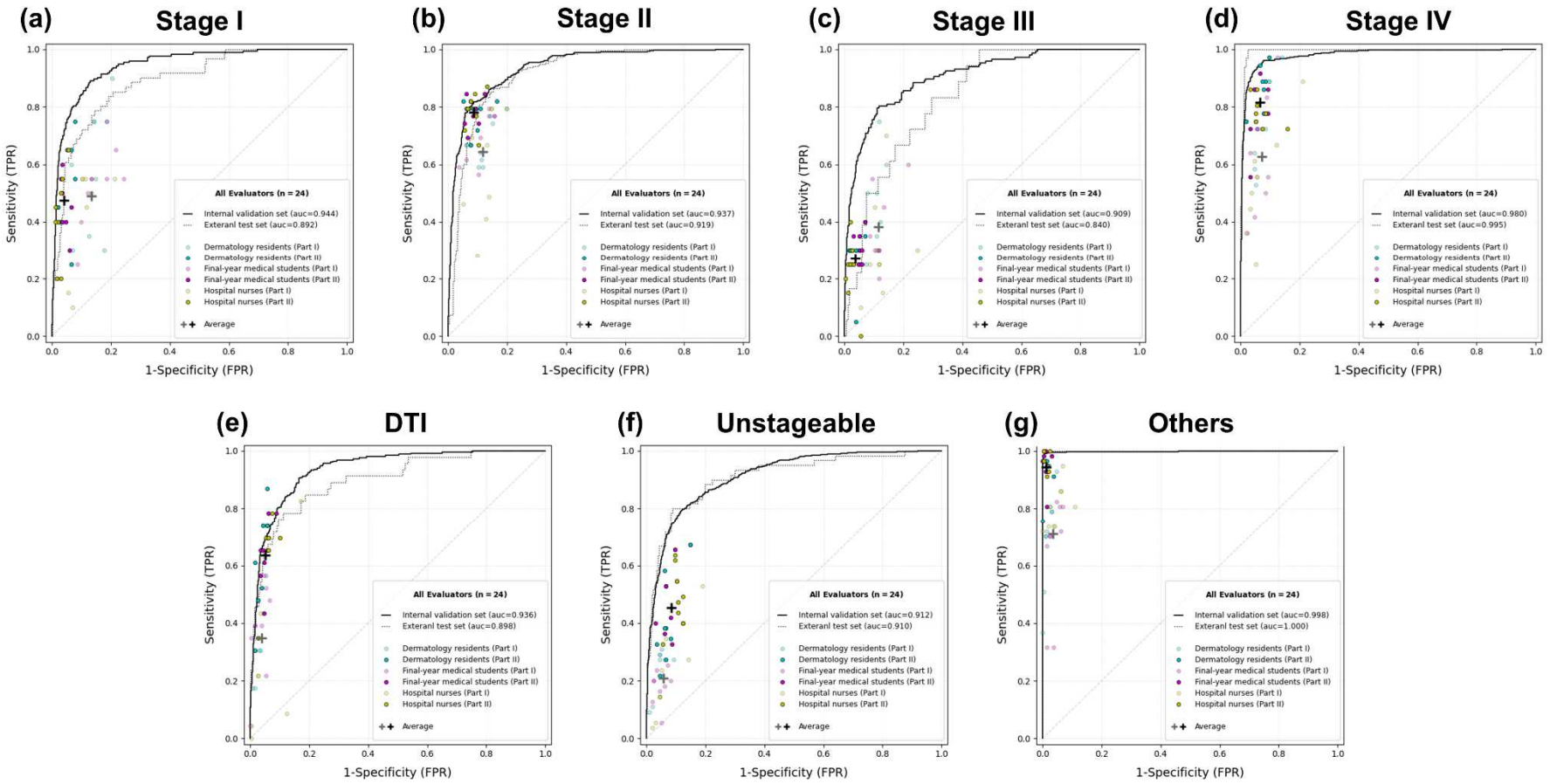

Figure 1. Receiver operating characteristic (ROC) curves depicting the performance metrics of the CNN model and 24 human evaluators. ROC curves for each class were drawn for the internal validation (black curve) and external test set (dotted curve). The dots represent the sensitivity (TPR, true positive rate) and 1-specificity (FPR, false positive rate) of the human evaluators in each part. (a) Stage I, (b) Stage II, (c) Stage III, (d) Stage IV, (e) Deep tissue injury (DTI). (f) Unstageable, (g) Others.

3.3 Effect of AI predictions on participant responses

Figure 2 presents the inter-rater agreement rates between the participant responses, CNN model predictions, and gold standard label. In Part I, significant disagreements (Fig. 2A, Fleiss' κ of 0.414 [95% CI, 0.410–0.417]) and performance gaps (Fig. 2C) were observed among the participants. In contrast, in Part II, the inter-rater agreement among participants significantly improved (Fig. 2B, Fleiss' κ of 0.641 [95% CI, 0.638–0.644]), with reduced performance gap between participants (Fig. 2D). Hospital nurses showed a remarkable increase in concordance, with Fleiss' κ increasing from 0.345 (95% CI, 0.335–0.354) to 0.668 (95% CI, 0.657–0.678) with the assistance of the CNN model. A comparison of the responses in Parts I and II indicated that 47.7% of the participants (2862/6000) changed their responses as a direct result of augmented decision-making, among which 56.5% (1617/2862) resulted in increased accuracy. Nonetheless, the accuracy decreased by 16.7% in instances where the top-3 predictions from the CNN model were incorrect (*N*=336). Similarly, a drop in accuracy of 13.2% was recorded when the top-1 predictions of the CNN model were incorrect (*N*=1,512).

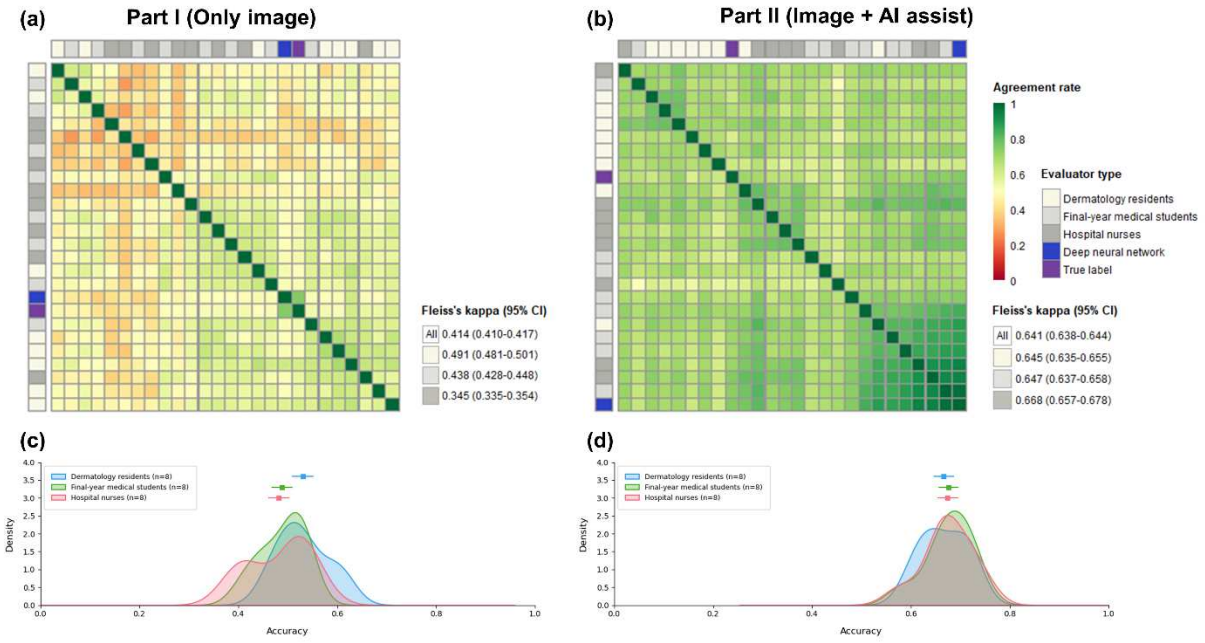

Figure 2. Inter-rater agreement and reliability among human evaluators. The heatmaps

visualize the inter-rater agreement rate across the responses of 24 human participants, model prediction, and true label of (a) Part I and (b) Part II. The kernel density plots show the accuracy distributions of each participant type for (c) Part I and (d) Part II.

3.4 Visualization of the explanatory model

We applied gradient-weighted class activation mapping (Grad-CAM) and t-distributed stochastic neighbor embedding (t-SNE) techniques to visualize the neural network model. In Figure 3, the heatmaps from class activation mapping represent pixel areas activated by the deep neural network, indicating its ability to distinguish PI lesions from the surrounding skin and detect characteristic features corresponding to each PI grade. Conversely, Figure S3 demonstrates instances where the model made incorrect predictions, displaying a tendency to inaccurately focus on lesion margins or areas of hyperpigmentation outside the ulcers. Figure 4 shows the two-dimensional representation of the internal features extracted from the CNN model. The CNN extracted distinct features for PI classification, with each class occupying relative regions in the two-dimensional map corresponding to clinical features. Despite some overlap between the classes, the model successfully distinguished the “others” skin lesions from PIs, with the severity order of PI grading reflected in the cluster distributions.

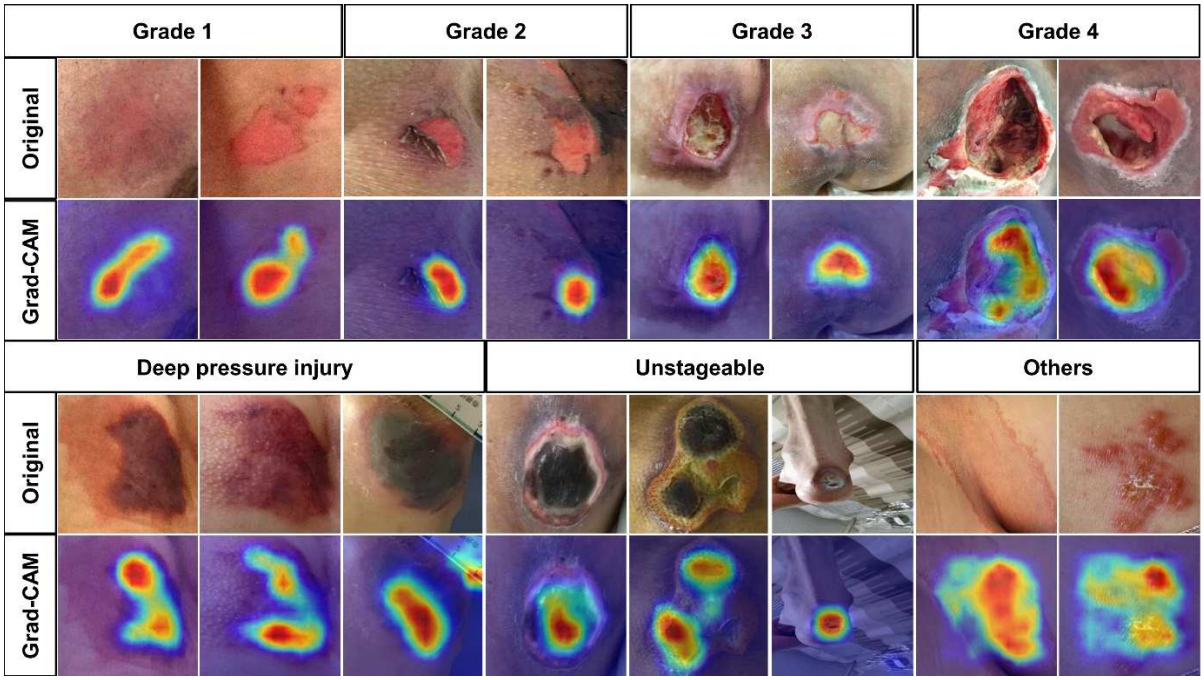

Figure 3. Visual explanations of pressure injury grading model via class activation mapping. Clinical images of each pressure injury grading and corresponding heatmaps via gradient-based localization (Grad-CAM), with activation focused on skin defects caused by pressure injury.

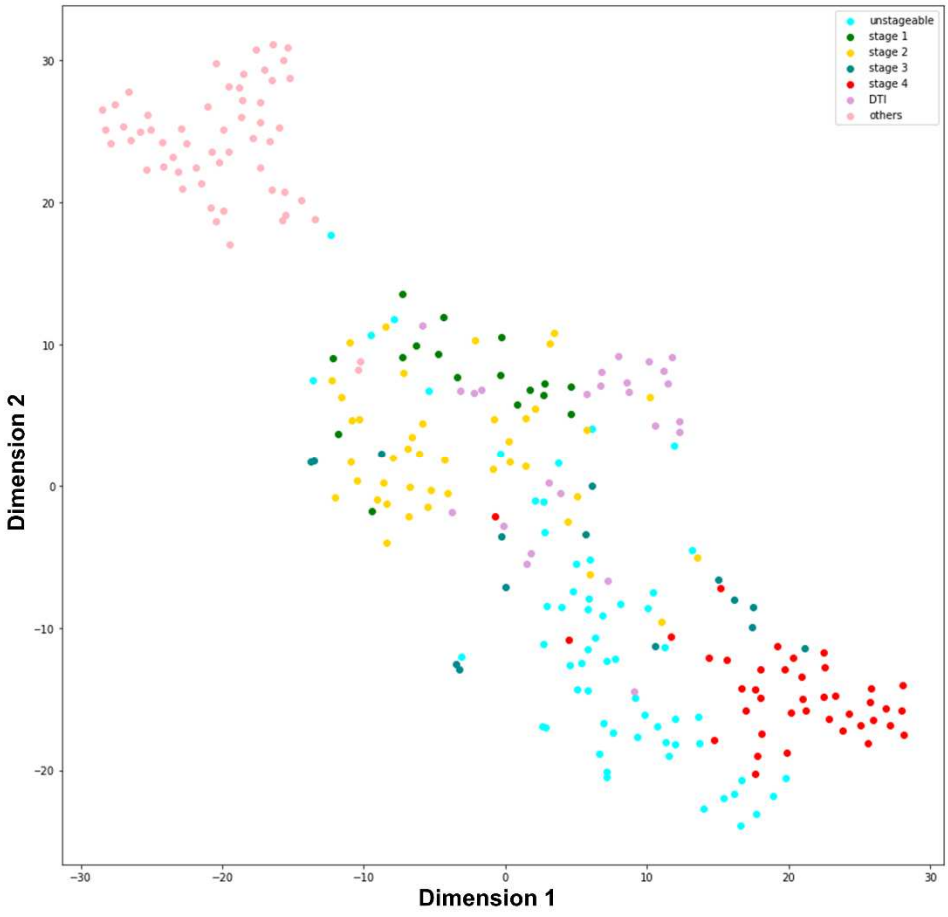

Figure 4. T-distributed stochastic neighbor embedding (t-SNE) visualization of the last hidden layer representations in the neural network model. The output of the last hidden layer of the neural network is projected onto a two-dimensional map using the t-SNE method. Colored point clouds represent different severity classifications, showing how the algorithm clusters the pressure injury grading. DTI denotes deep tissue injury.

4. Discussion

This study aimed to develop a CNN model for categorizing PI stages and assess its clinical value in improving the precision of healthcare providers. The deep neural network model based on SE-ResNext101 outperformed most human evaluators. However, the accuracy, sensitivity, and specificity of the participant responses improved when aided by the CNN model. This improvement was especially pronounced for novice medical students and

hospital nurses. Furthermore, the inter-rater agreement among participants improved with the support of the neural network model. Nonetheless, the accuracy was considerably reduced when the top-3 or top-1 predictions of the model were incorrect.

Several researchers have attempted to classify PI images using neural network models based on the clinical staging system. However, most of these studies relied on public datasets containing a limited number of images per class, typically only a few dozen [12, 14, 20]. As a result, it is recommended that more than 150 training images must be used per class to achieve reasonable classification accuracy when resources are limited [21]. Therefore, it is necessary to construct a sufficiently large dataset with reliable labeling of PI grades. Although two sets of researchers incorporated high-quality PI images with a sample size of over 1,000, all the grades of the NPUAP criteria, including unstageable and DTI, were not considered [22, 23]. These studies were also limited by the lack of external validation owing to the single-center image collection.

In this study, we used 3,098 images from two independent hospitals to develop a CNN model capable of classifying all stages of the NPUAP guidelines and frequent dermatoses in injury-prone areas. Our CNN model, SE-ResNext101, exhibited a performance (ROC-AUC of 0.951) comparable with those of the best-fit models used in previous studies (ResNet-152 [23] and EfficientNet-B4 [22] with ROC-AUC values of 0.93 and 0.976, respectively). Although different models were used in each study, the performance trends for each ulcer stage were similar, with Stage III consistently corresponding to the poorest performance, consistent with the observation that the visual diagnosis of Stage III PIs was the most controversial among healthcare professionals [24-26]. In general, relying solely on image-based inspection without actual physical examination makes it quite challenging to differentiate partial-thickness skin loss (Stage II) from full-thickness skin loss (Stage III) [25].

To date, several researchers have investigated the effect of AI assistance on the diagnostic

accuracy of non-expert physicians for various skin conditions, focusing on skin neoplasms [8, 10, 27-29]. Despite the distinct disease characteristics of skin cancer and PIs, the effect of AI assistance on decision-making among the survey participants was similar. Therefore, augmented decision-making based on CNNs improved the diagnostic performance of healthcare providers and narrowed the practice gap of non-experts in classifying the PI grades [10]. Additionally, less experienced clinicians derived the largest benefit from CNN support, consistent with findings from a recent prospective diagnostic study [29].

Although multiple studies have demonstrated the positive impact of adequate health professional education on the accuracy of pressure ulcer classification [30-32], we expect that the proposed CNN model will help enhance the ability of inexperienced healthcare providers in pressure ulcer grading. To ensure widespread utilization, we are currently developing a mobile web application system named “Amulda” (meaning “heal” in Korean), which can enable real-time pressure ulcer staging based on photographs and provide guidance on appropriate prevention and management strategies (Fig. S4). A key consideration is that human participants are susceptible to erroneous information provided by the CNN model, which may lead to a performance loss when the AI support is flawed, consistent with other comparable analyses [27, 28].

The limitations of this study can be summarized as follows: first, the retrospective nature of this study led to an imbalanced dataset and possible selection bias, which could constrain the relevance of the findings to a wider population. The notable variation in sensitivity, specificity, F1 score, and ROC-AUC across different severity classes could be attributed to this imbalance, potentially affecting model performance and robustness when classifying rarer PI stages. Second, because the data were sourced from two referral hospitals in Seoul, South Korea, the dataset included only patients with Fitzpatrick skin types III and IV, possibly limiting the model’s generalization ability to various populations and clinical

settings worldwide. Third, the performance of the human participants may have been underestimated in the experimental setting, considering they were not given access to patient clinical information, which is expected to be available in real-world clinical practice. The underestimation may also be attributable to the fact that participants classified PI stages based solely on images lacking anatomical markings. Also, the images based on two-dimensional representations lack three-dimensional details such as ulcer depth, which are critical for accurate pressure ulcer staging in a real clinical scenario, posing further challenges in making precise assessments.

5. Conclusions

A deep neural network model was developed for categorizing PI stages, with its potential clinical value in improving PI diagnoses by healthcare providers demonstrated. The CNN model outperformed most of the human evaluators, and AI assistance enhanced the performance and concordance of decisions among a group of 24 healthcare providers. Nevertheless, while our model shows promise, it's pivotal to acknowledge the inherent limitations of solely relying on images for PI staging and progression predictions. Combinatory approach considering both clinical variables known to influence PI development [33] along with correlated image data would yield more holistic and accurate insights. Further prospective, integrative studies in larger, multicenter settings and diverse regions and ethnicities must be performed to assess the effectiveness of the proposed model in real-world applications.

Summary Table

- The National Pressure Injury Advisory Panel developed a staging system that can classify and describe pressure injuries (PIs).
- Proper assessment of PIs is crucial for treatment planning, but accurate PI staging remains challenging for first-line caregivers, especially non-specialists.
- Deep neural networks can facilitate accurate PI staging and improve PI diagnoses by healthcare providers.

Acknowledgments: We thank the all the healthcare providers who participated in the online survey. Moreover, we are indebted to the Digital Health Center at the Yonsei University Health System and Yonsei University College of Medicine Synapse Center for their assistance in project management and technical support.

Declaration of Interest: None

Ethics Statement: This study was approved with waivers of informed consent by the Institutional Review Board of Severance Hospital (approval number 4-2021-1076) and Gangnam Severance Hospital (approval number 3-2022-0411). The patients referenced in this manuscript have provided written informed consent for the publication of their case details.

Funding sources: This work was supported by the Korea Medical Device Development Fund grant funded by the Korea government (the Ministry of Science and ICT; Ministry of Trade, Industry and Energy; Ministry of Health & Welfare; and Ministry of Food and Drug Safety) (Project Number: 1711179407, RS-2022-KD141479). Also, this research was supported by a grant of the Medical data-driven hospital support project through the Korea Health Information Service (KHIS), funded by the Ministry of Health & Welfare, Republic of Korea.

Data Availability Statement: The data that support the findings of this study are available from the corresponding author upon reasonable request. Also, the relevant source code for developing and validating the neural network and all pixel-wise annotations were published in our public repository (https://github.com/yunvleetp/Pressure-Ulcer-Staging-pytorch).

References

[1] L.E. Edsberg, J.M. Black, M. Goldberg, L. McNichol, L. Moore, M. Sieggreen, Revised national pressure ulcer advisory panel pressure injury staging system: revised pressure injury staging system, Journal of Wound, Ostomy, and Continence Nursing, 43 (2016) 585.

[2] Z. Li, F. Lin, L. Thalib, W. Chaboyer, Global prevalence and incidence of pressure injuries in hospitalised adult patients: A systematic review and meta-analysis, Int. J. Nurs. Stud., 105 (2020) 103546.

[3] T.J. Liu, M. Christian, Y.-C. Chu, Y.-C. Chen, C.-W. Chang, F. Lai, H.-C. Tai, A pressure ulcers assessment system for diagnosis and decision making using convolutional neural networks, J. Formosan Med. Assoc., 121 (2022) 2227-2236.

[4] J.S. Mervis, T.J. Phillips, Pressure ulcers: Pathophysiology, epidemiology, risk factors, and presentation, J. Am. Acad. Dermatol., 81 (2019) 881-890.

[5] I. Cho, H.-A. Park, E. Chung, Exploring practice variation in preventive pressure-ulcer care using data from a clinical data repository, Int. J. Med. Inf., 80 (2011) 47-55.

[6] S.S. Han, M.S. Kim, W. Lim, G.H. Park, I. Park, S.E. Chang, Classification of the Clinical Images for Benign and Malignant Cutaneous Tumors Using a Deep Learning Algorithm, Journal of Investigative Dermatology, 138 (2018) 1529-1538.

[7] P. Puri, N. Comfere, L.A. Drage, H. Shamim, S.A. Bezalel, M.R. Pittelkow, M.D. Davis, M. Wang, A.R. Mangold, M.M. Tollefson, Deep learning for dermatologists: Part II. Current applications, J. Am. Acad. Dermatol., (2020).

[8] P. Tschandl, C. Rosendahl, B.N. Akay, G. Argenziano, A. Blum, R.P. Braun, H. Cabo, J.-Y. Gourhant, J. Kreusch, A. Lallas, Expert-level diagnosis of nonpigmented skin cancer by combined convolutional neural networks, JAMA Dermatol, 155 (2019) 58-65.

[9] S. Lee, J.W. Lee, S.J. Choe, S. Yang, S.B. Koh, Y.S. Ahn, W.S. Lee, Clinically Applicable Deep Learning Framework for Measurement of the Extent of Hair Loss in Patients With Alopecia Areata, JAMA Dermatol, 156 (2020) 1018-1020.

[10] S.S. Han, I.J. Moon, W. Lim, I.S. Suh, S.Y. Lee, J.-I. Na, S.H. Kim, S.E. Chang, Keratinocytic Skin Cancer Detection on the Face Using Region-Based Convolutional Neural Network, JAMA Dermatology, 156 (2020) 29-37.

[11] C.W. Chang, M. Christian, D.H. Chang, F. Lai, T.J. Liu, Y.S. Chen, W.J. Chen, Deep learning approach based on superpixel segmentation assisted labeling for automatic pressure ulcer diagnosis, PLoS One, 17 (2022) e0264139.

[12] P. Fergus, C. Chalmers, W. Henderson, D. Roberts, A. Waraich, Pressure Ulcer Categorisation

using Deep Learning: A Clinical Trial to Evaluate Model Performance, arXiv preprint arXiv:2203.06248, (2022).

[13] B. García-Zapirain, M. Elmogy, A. El-Baz, A.S. Elmaghraby, Classification of pressure ulcer tissues with 3D convolutional neural network, Med Biol Eng Comput, 56 (2018) 2245-2258.

[14] C.H. Lau, K.H.-O. Yu, T.F. Yip, L.Y.F. Luk, A.K.C. Wai, T.-Y. Sit, J.Y.-H. Wong, J.W.K. Ho, An artificial intelligence-enabled smartphone app for real-time pressure injury assessment, Frontiers in Medical Technology, 4 (2022) 905074.

[15] S. Zahia, D. Sierra-Sosa, B. Garcia-Zapirain, A. Elmaghraby, Tissue classification and segmentation of pressure injuries using convolutional neural networks, Comput. Methods Programs Biomed., 159 (2018) 51-58.

[16] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7132-7141.

[17] S. Zhao, L. Zhou, W. Wang, D. Cai, T.L. Lam, Y. Xu, Toward Better Accuracy-Efficiency Trade-Offs: Divide and Co-Training, IEEE Transactions on Image Processing, 31 (2022) 5869-5880.

[18] B. Sanchez-Lengeling, J.N. Wei, B.K. Lee, R.C. Gerkin, A. Aspuru-Guzik, A.B. Wiltschko, Machine learning for scent: Learning generalizable perceptual representations of small molecules, arXiv preprint arXiv:1910.10685, (2019).

[19] J.L. Fleiss, Measuring nominal scale agreement among many raters, Psychological bulletin, 76 (1971) 378.

[20] B. Yilmaz, E. Atagün, F.Ö. Demırcan, İ. Yücedağ, Classification of pressure ulcer images with logistic regression, 2021 International Conference on INnovations in Intelligent SysTems and Applications (INISTA), IEEE, 2021, pp. 1-6.

[21] S. Shahinfar, P. Meek, G. Falzon, "How many images do I need?" Understanding how sample size per class affects deep learning model performance metrics for balanced designs in autonomous wildlife monitoring, Ecol. Inform., 57 (2020) 101085.

[22] S. Seo, J. Kang, I.H. Eom, H. Song, J.H. Park, Y. Lee, H. Lee, Visual classification of pressure injury stages for nurses: A deep learning model applying modern convolutional neural networks, J. Adv. Nursing, (2023).

[23] B. Ay, B. Tasar, Z. Utlu, K. Ay, G. Aydin, Deep transfer learning-based visual classification of pressure injuries stages, Neural Computing and Applications, 34 (2022) 16157-16168.

[24] J. Stausberg, N. Lehmann, K. Kröger, I. Maier, W. Niebel, Reliability and validity of pressure ulcer diagnosis and grading: an image-based survey, Int. J. Nurs. Stud., 44 (2007) 1316-1323.

[25] D. Beeckman, L. Schoonhoven, J. Fletcher, K. Furtado, L. Gunningberg, H. Heyman, C. Lindholm, L. Paquay, J. Verdu, T. Defloor, EPUAP classification system for pressure ulcers: European reliability study, J. Adv. Nursing, 60 (2007) 682-691.

[26] T. Defloor, L. Schoonhoven, Inter-rater reliability of the EPUAP pressure ulcer classification system using photographs, J. Clin. Nursing, 13 (2004) 952-959.
[27] S.S. Han, Y.J. Kim, I.J. Moon, J.M. Jung, M.Y. Lee, W.J. Lee, C.H. Won, M.W. Lee, S.H. Kim, C. Navarrete-Dechent, Evaluation of Artificial Intelligence–Assisted Diagnosis of Skin Neoplasms: A Single-Center, Paralleled, Unmasked, Randomized Controlled Trial, J. Invest. Dermatol., 142 (2022) 2353-2362. e2352.
[28] P. Tschandl, C. Rinner, Z. Apalla, G. Argenziano, N. Codella, A. Halpern, M. Janda, A. Lallas, C. Longo, J. Malvehy, Human–computer collaboration for skin cancer recognition, Nat. Med., 26 (2020) 1229-1234.
[29] J.K. Winkler, A. Blum, K. Kommoss, A. Enk, F. Toberer, A. Rosenberger, H.A. Haenssle, Assessment of Diagnostic Performance of Dermatologists Cooperating With a Convolutional Neural Network in a Prospective Clinical Study: Human With Machine, JAMA Dermatol, (2023).
[30] D.L. Young, N. Estocado, M.R. Landers, J. Black, A pilot study providing evidence for the validity of a new tool to improve assignment of national pressure ulcer advisory panel stage to pressure ulcers, Adv. Skin Wound Care, 24 (2011) 168-175.
[31] W.H. Ham, L. Schoonhoven, M.J. Schuurmans, R. Veugelers, L.P. Leenen, Pressure ulcer education improves interrater reliability, identification, and classification skills by emergency nurses and physicians, J. Emerg. Nurs., 41 (2015) 43-51.
[32] Y.J. Lee, J.Y. Kim, K.A.o.W.O.C. Nurses, Effects of pressure ulcer classification system education programme on knowledge and visual differential diagnostic ability of pressure ulcer classification and incontinence-associated dermatitis for clinical nurses in Korea, Int. Wound J., 13 (2016) 26-32.
[33] C. Ji-Yu, Z. Man-Li, S. Yi-Ping, C. Hong-Lin, Predicting the development of surgery-related pressure injury using a machine learning algorithm model, The Journal of Nursing Research, 29 (2021) e135.