



저작자표시-동일조건변경허락 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.
- 이차적 저작물을 작성할 수 있습니다.
- 이 저작물을 영리 목적으로 이용할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



동일조건변경허락. 귀하가 이 저작물을 개작, 변형 또는 가공했을 경우에는, 이 저작물과 동일한 이용허락조건하에서만 배포할 수 있습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

Development of an artificial intelligence
model to evaluate and predict the severity of
postoperative scars

Jemin Kim

Department of Medicine

The Graduate School, Yonsei University

Development of an artificial intelligence
model to evaluate and predict the severity of
postoperative scars

Directed by Professor JuHee Lee

The Doctoral Dissertation
submitted to the Department of Medicine,
the Graduate School of Yonsei University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Medical Science

Jemin Kim

December 2022

This certifies that the Doctoral Dissertation of
Jemin Kim is approved.

[Signature]

Thesis Supervisor : Ju Hee Lee

[Signature]

Thesis Committee Member#1 : Sang Ho Oh

[Signature]

Thesis Committee Member#2 : Jae-woo Kim

[Signature]

Thesis Committee Member#3: Mi Youn Park

[Signature]

Thesis Committee Member#4: Hyo Hyun Ahn

The Graduate School
Yonsei University

December 2022

ACKNOWLEDGEMENTS

The final outcome of this study required much guidance and inspiration from Professor Ju Hee Lee. I would like to deeply thank her for all his support and encouragement along the completion of this study.

I also owe my profound gratitude to Professor Sang Ho Oh, Professor Jae-woo Kim, Professor Hyo Hyun Ahn, and Professor Mi Youn Park who took keen interest on this study and offered invaluable professional advices and guidance.

I want to give my special thanks to Dr. Inrok Oh from LG Chem Ltd. for collaborating on the artificial intelligence model and providing technical support for this study.

I also thank my colleagues in the clinic and lab members for their kind support and assistances. Last but not least, I thank my family for their unchanging presence in my life.

<TABLE OF CONTENTS>

ABSTRACT.....	iv
I. INTRODUCTION	1
II. MATERIALS AND METHODS	2
1. Study design and participants.....	2
A. Patient cohorts.....	2
B. Data acquisition and preprocessing	3
2. Neural network structure and training.....	3
3. Evaluation of algorithm performance	6
4. Statistics	6
III. RESULTS	8
1. Patients and clinical characteristics.....	8
2. Performance of the model.....	8
3. Comparison between the neural network vs. dermatologists.....	15
4. Visualization of the explanatory model	15
IV. DISCUSSION	20
V. CONCLUSION	25
REFERENCES	26
ABSTRACT(IN KOREAN)	30

LIST OF FIGURES

Figure 1. Description and representative images of scar severity classification according to morphologic features.....	5
Figure 2. Illustration of severity prediction model's pipeline.....	7
Figure 3. Receiver operating characteristic (ROC) curves for the three classification models	13
Figure 4. A Bland-Altman plot shows the association between the measured and predicted VSS score in the regression model.....	16
Figure 5. Scar severity classification performance of the CNN and dermatologists.....	17
Figure 6. Confusion matrix comparison between prediction models and dermatologists.....	18
Figure 7. t-SNE visualization of the last hidden layer representations in the image-based prediction model.....	19
Figure 8. Visual explanations of postoperative scar cases via class activation mapping	21
Figure 9. Interpretation of the clinical-data-based model via SHAP analysis.....	22

LIST OF TABLES

Table 1. Characteristics of main and external dataset.....	4
Table 2. Baseline patient characteristics and comparison of features stratified by scar severity groups	9
Table 3. Multinomial logistic regression analysis by scar severity groups	10
Table 4. Performance of severity prediction models.....	12
Table 5. Performance of prediction model according to each severity class in the internal testing set	14

ABSTRACT

Development of an artificial intelligence model to evaluate and predict the severity of postoperative scars

Jemin Kim

*Department of Medicine
The Graduate School, Yonsei University*

(Directed by Professor Ju Hee Lee)

Although most postoperative scars are inevitable sequelae after the surgical procedures, they cause significant cosmetic problems and functional impairments. The appropriate evaluation of the severity of the scar is crucial for determining the proper treatment modalities, yet there is no gold standard in assessing the scars. Our objective of the study was to develop and evaluate an artificial intelligence (AI) model using the image and clinical data to predict the severity of postoperative scars.

Deep neural network models were trained and validated using images and clinical data from 1,283 patients (main dataset: 1,043, external dataset: 240) with post-thyroidectomy scars. The Model's performance in classifying the scar severity was externally validated on patients of another hospital and tested against 16 dermatologists. With the internal test set, the area under the receiver operating characteristic curve (ROC-AUC) of the image-based model was 0.931 (95% CI 0.910-0.949), and increased to 0.938 (0.916-0.955) when combined with clinical data. With the external test set, the ROC-AUC of the image-based and combined prediction model were 0.896 (0.874-0.916) and 0.912 (0.892-0.932), respectively. The tested algorithm performance with images of the internal test set was comparable to that of 16 dermatologists. Regression model for VSS score prediction showed the mean absolute error of 1.075 (95% CI 0.960-1.184) in the internal testing set,

and 1.183 (95% CI 1.080-1.283) in the external testing set.

In conclusion, this study revealed that deep neural network model derived from image and clinical data could predict the severity of postoperative scar. We anticipate that the proposed AI model may utilize in the clinical practice of scar management, especially for deciding severity and treatment initiation.

Key words : postoperative scar, hypertrophic scar, dataset, artificial intelligence, deep learning

Development of an artificial intelligence model to evaluate and predict the severity of postoperative scars

Jemin Kim

*Department of Medicine
The Graduate School, Yonsei University*

(Directed by Professor Ju Hee Lee)

I. INTRODUCTION

Scars are one of the most common medical problems that affect a patient not only cosmetically but also cause functional impairment and psychosocial burdens. Significantly, after surgical procedures, hypertrophic scars and even keloids frequently develop. The incidence of hypertrophic scars after a surgical procedure is estimated to be 40-70% in the absence of adequate management¹, and they significantly impair the quality of life (QoL) of patients². The post-thyroidectomy scar is particularly problematic because of the location (exposed area of neck), relatively young age of the affected patients, and rapidly increasing incidence of thyroid cancer³. Furthermore, since the underlying molecular mechanism of wound healing and scarring formation is quite complex⁴, predisposing factors or prognostic markers for hypertrophic scarring are also not wholly understood⁵. For post-thyroidectomy scar, several clinical risk factors related to hypertrophic scarring were reported, such as young age, high body mass index (BMI), scar-related symptoms, incision site near the sternal notch, prominent sternocleidomastoid muscles, and history of abnormal wound healing or pathologic scarring^{3,5,6}.

In the era of artificial intelligence (AI), a convolutional neural network (CNN) has been successfully introduced and formed the basis for various emerging applications in the field

of dermatology⁷. Current studies using CNN in dermatology mainly focused on classifying skin diseases, especially skin cancer⁸⁻¹¹, or lesion identification and quantification via segmentation algorithm^{12,13}. However, recent studies in radiology revealed that implementing a deep learning model combining imaging and clinical data could predict disease severity, risk of progression, or treatment response¹⁴⁻¹⁶.

In this study, we aimed to develop an artificial intelligence model that predicts the severity of postoperative scars using medical images and clinical data. We also aimed to compare the performance of the AI model to that of dermatologists.

II. MATERIALS AND METHODS

1. Study design and participants

A. Patient cohorts

We did a retrospective study and identified patients with a post-thyroidectomy scar who presented to the scar laser and plastic surgery center within the Yonsei cancer hospital, Seoul, Republic of Korea, between September 2015 and December 2021. The study was approved by the institutional review board of Yonsei University Severance Hospital (approval number 4-2022-0741). In this main dataset, we randomly assigned these patients to model training, validation, and internal testing datasets (7:1:2). Also, we independently collected the post-thyroidectomy patients who presented to the department of dermatology at Severance Hospital, Seoul, Republic of Korea, between December 2010 and July 2015 who were assigned to the external testing dataset. All patients underwent conventional thyroidectomy, minimally invasive thyroidectomy (MIS), modified radical neck dissection (MRND), or transaxillary robotic thyroidectomy and were referred to the dermatology clinic for scar minimization treatments. Medical images of the anterior neck or axilla were taken with high-resolution (≥ 6 million pixels) digital cameras at the initial visit, 3, 6, and 12 months of follow-up. We additionally collected the photographs of patients without scars in the anterior neck region at the same intuition to set as a control ('normal') group. In total, 2,727 images from 1,043 patients were included in the main dataset, and 234

images from 185 patients were obtained from the external dataset (Table 1).

B. Data acquisition and preprocessing

For each patient's visits, clinical data were collected, including age, sex, BMI, date after surgery (scar age), past keloid history, operation site, clinical scar characteristics (itching, pain, adhesion, tightening, induration, or edema), number of treatment sessions, and treatment response (for follow-up visits). The digital images of the anterior neck or axilla included in the study were de-identified and minimally cropped to contain adjacent anatomical structure around scar; for example, we cut off the photos of the anterior neck to include from the adam's apple to the sternal notch. Then these images were independently scored for scar severity by three board-certified dermatologists who specialized in scar treatment, using the Vancouver scar scale (VSS)¹⁷. Based on the VSS score and the required scar treatment modalities judged by the scar-specialized dermatologists, we classified the scars into four categories by their severity: normal, mild, moderate, and severe (Figure 1)¹⁸. Treatment response was defined as $\geq 50\%$ of VSS score or ≥ 2 decrements of severity grade, compared to the initial visit.

2. Neural network structure and training

For the image-based severity prediction model, we adopted the convolutional block attention module (CBAM) integrated with a ResNet-50. CBAM consists of a channel and spatial attention submodules, which allows focusing on meaningful features and suppressing unwanted ones¹⁹. All images were resized to 224 x 224 pixels and normalized to the range from -1 to 1 for training. Also, to adjust the data imbalance between classes and avoid overfitting, data augmentation techniques such as random image cropping and white balance were adopted²⁰. The cross entropy loss was used to train the network, and softmax operation was applied to model output. A stochastic gradient descent (SGD) optimizer was used with a learning rate of 0.01 and batch size 16. For the clinical-data-based severity prediction, a multilayer perceptron (MLP) model was trained to distinguish each severity class based on 11 collected clinical variables. Finally, the combined model

Table 1. Summary of the main and external dataset and the corresponding demographic information

Characteristics	Dataset	
	Main	External
Data collection period	2015. 9 - 2021. 12	2010. 12 - 2015. 7
Location (Hospital)	Scar laser and plastic surgery center, Yonsei cancer hospital	Department of dermatology, Severance Hospital
Dataset allocation	Training (70%) Validation (20%) (Internal) Testing (10%)	External testing (100%)
Patient demographics		
Unique individuals, n	1043	240
Female sex, %	88.7	87.9
Age at diagnosis, mean \pm SD	40.5 \pm 11.5	42.2 \pm 11.2
Number of images, n	2727	374
Normal	332	50
Mild	688	74
Moderate	1289	201
Severe	418	49

Abbreviations: SD; standard deviation





Class	Description	Required treatments	Representative figures
Normal	Flat, soft, normal color (VSS 0)	No treatment required	
Mild	Height < 1mm, supple, light to dark pink color (VSS 1-4)	Topical applications (silicon gel/sheet, onion extract gel)	
Moderate	Height < 2mm, yielding, dark pink to dark red color (VSS 5-7)	Intralesional triamcinolone injection ± fractional ablative lasers	
Severe	Height > 2mm, firm to banding, red to brown color (VSS ≥ 8)	Combined laser treatments (lasers, injections, superficial cryotherapy), scar revision surgery, etc.	

Figure 1. Description and representative images of scar severity classification according to morphologic features and treatment requirements.

Abbreviations : VSS, vancouver scar scale

of severity prediction was obtained from the 6:4 ratio of the weighted sum of the image-based and clinical-data-based prediction models. Furthermore, we developed an image-based regression model to estimate the VSS score based on the score labeled on each image (Figure 2). The neural network structure was implemented in Python, using Pytorch backend (Python 3.9.0, Pytorch 1.9.0).

3. Evaluation of algorithm performance

The trained model was evaluated using the test dataset, both from the internal and external testing datasets. For the best-fitted model, five-fold stratified cross-validation was performed to verify the robustness of the model. Then, the classification performance of the image-based severity prediction model was compared against that of eight board-certified dermatologists and eight dermatology residents. We randomly selected 240 images from the internal test dataset (60 images from each severity class), presented them as original resolution photographs, and asked them to choose the best appropriate classification (single choice). Furthermore, a class activation map (Grad-CAM and Guided Grad-CAM), which allows visualizing the important features via gradient-based localization²¹, was implemented to understand the prediction made by the deep network model qualitatively.

4. Statistics

The performance of each model was calculated by Top-1 accuracy, sensitivity, specificity, and F1 score. Receiver operating characteristic (ROC) curves were drawn via sensitivity and specificity for each threshold, with areas under the curve (AUC) calculated. 95% Confidence intervals (CI) were calculated using the bootstrap resampling of the test dataset with the replacement $N = 1000$ times²².

Also, we examined the internal features learned by the models using t-distributed stochastic neighbor embedding (t-SNE), which reduces the 2048-dimensional vectors obtained using the classification models to a 2-dimensional map.

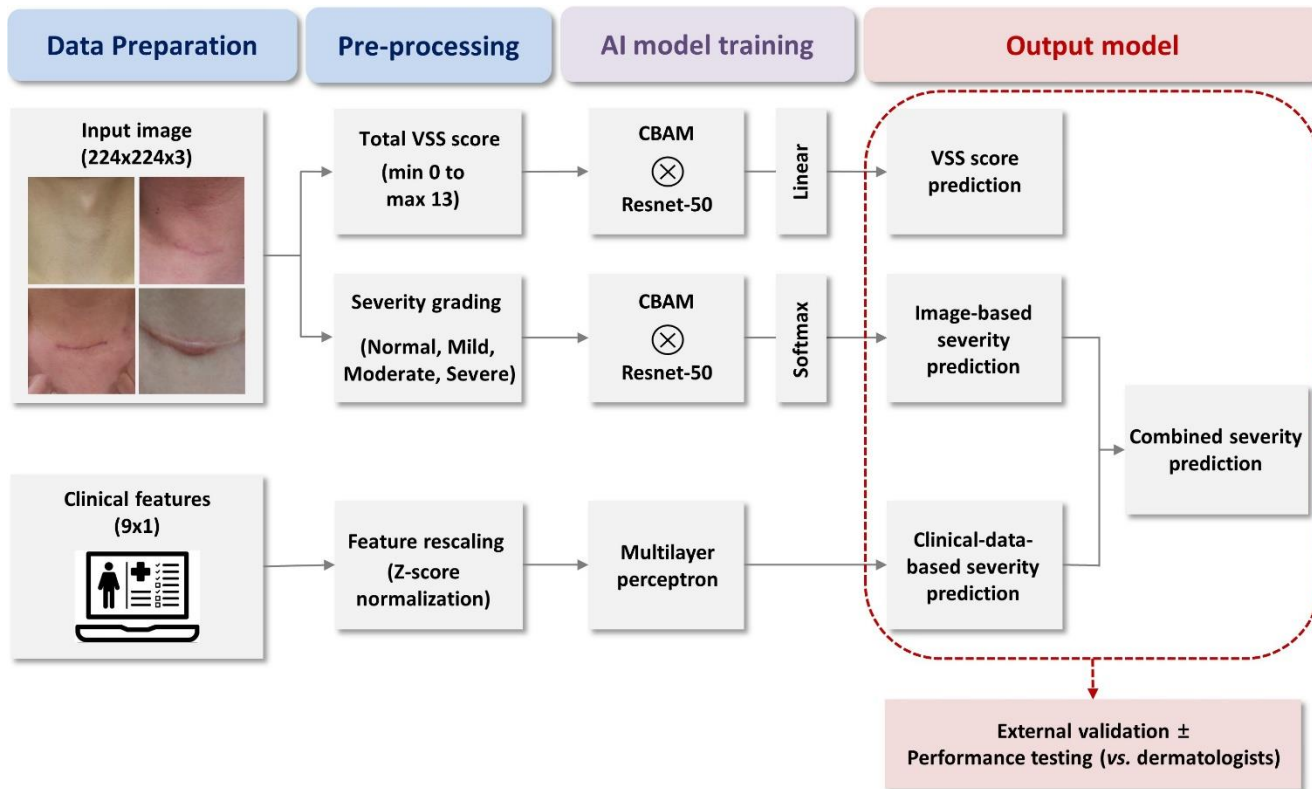


Figure 2. Illustration of severity prediction model's pipeline.

Abbreviations: CBAM, convolutional block attention module; CNN, convolutional neural network; MLP, multilayer perceptron

Clinical characteristics of the three severity groups (mild, moderate, severe) were compared using the Fisher's exact or chi-square test with adjusted residuals if variables were in 2x3 categorical tables. Otherwise, one-way analysis of variance (ANOVA) test was used to compare continuous variables. All variables with $P < 0.10$ in the analyses were selected for multinomial logistic regression analysis (reference group: moderate severity). Statistical analyses were performed using Python version 3.9.0, and all P values are two-sided and less than .05 were considered statistically significant.

III. RESULTS

1. Patients and clinical characteristics

A total of 1043 patients were included in the main dataset, and 109 (10.5%) had mild, 705 (67.6%) had moderate, and 229 (22.0%) had severe degrees of scar severity, according to the initial clinical presentation. When comparing clinical variables between these severity groups, the factors listed below showed significant differences; BMI, date after surgery, minimally invasive thyroidectomy (MIT), modified radical neck dissection (MRND), transaxillary approach, itching/pain, adhesion/tightening, and induration/edema (Table 2).

To identify predictive factors associated with scar severity, we performed multinomial logistic regression with the significant variables ($P < 0.10$) shown in Table 2, setting the moderate group as the reference group. In the multivariate model, MIT (OR: 2.18, 95% CI: 1.32-3.60) and the date after surgery (OR: 1.04, 95% CI: 1.03-1.06) were positively correlated with the mild scar severity. Transaxillary approach (OR: 3.11, 95% CI: 1.75-5.50), date after surgery (OR: 1.07, 95% CI: 1.05-1.09), and itching/pain (OR: 1.52, 95% CI: 1.03-2.24) were positively correlated with the severe scar severity, yet adhesion/tightening (OR: 0.69, 95% CI: 0.50-0.97) and induration/edema (OR: 0.55, 95% CI: 0.34-0.89) were negatively associated with the severe group (Table 3).

2. Performance of the model

Table 2. Baseline patient characteristics and comparison of features stratified by initial scar severity

Feature	Mild (N=109)	Moderate (N=705)	Severe (N=229)	Total (N=1043)	P value
Female sex	102 (93.6)	625 (88.7)	198 (86.5)	925 (88.7)	0.15
Age at diagnosis	42.2 ± 12.1	40.6 ± 11.2	39.2 ± 12.2	40.5 ± 11.5	0.077
Body mass index (BMI)	22.8 ± 3.50	23.2 ± 3.76	23.8 ± 4.36	23.3 ± 3.89	0.046*
Date after surgery (months)	10.9 ± 18.9	4.14 ± 7.59	10.7 ± 13.8	6.29 ± 11.3	<0.001*
Past keloid history	1 (0.9)	17 (2.4)	8 (3.5)	26 (2.5)	0.38
Location of surgery					
Conventional	68 (62.4)	473 (67.1)	149 (65.1)	690 (66.2)	0.59
MIT	29 (26.6) [†]	105 (14.9)	24 (10.5)	158 (15.1)	0.001*
MRND	6 (5.5)	92 (13.0)	34 (14.8)	132 (12.7)	0.047*
Transaxillary	6 (5.5)	34 (4.8)	27 (11.8) [†]	67 (6.4)	0.001*
Clinical scar characteristics					
Itching/pain	19 (17.4)	123 (17.4)	60 (26.2) [†]	202 (19.4)	0.012*
Adhesion/Tightening	54 (49.5)	346 (49.1) [†]	84 (36.7)	484 (46.4)	0.004*
Induration/Edema	17 (15.6)	171 (24.3) [†]	26 (11.4)	214 (20.5)	<0.001*

*Statistically significant P values (<0.05)

[†]Statistically significant adjusted standardized residuals (>2.1)

Abbreviations: MIT; Minimally invasive thyroidectomy, MRND; modified radical neck dissection

Table 3. Multinomial logistic regression analysis by scar severity groups¹

Independent variables	Mild		Severe	
	OR (95% CI)	P value	OR (95% CI)	P value
Age at diagnosis	1.02 (0.99-1.04)	0.077	0.99 (0.98-1.01)	0.63
Body mass index (BMI)	0.97 (0.91-1.03)	0.34	1.04 (0.99-1.08)	0.064
Date after surgery (months)	1.04 (1.03-1.06)	<0.001*	1.07 (1.05-1.09)	<0.001*
Location of surgery				
Conventional	Ref		Ref	-
MIT	2.18 (1.32-3.60)	0.002*	0.69 (0.42-1.16)	0.16
MRND	0.41 (0.16-1.04)	0.061	1.31 (0.81-2.13)	0.27
Transaxillary	1.31 (0.51-3.36)	0.58	3.11 (1.75-5.50)	<0.001*
Clinical scar characteristics				
Itching/pain	1.10 (0.63-1.92)	0.74	1.52 (1.03-2.24)	0.034*
Adhesion/Tightening	1.10 (0.71-1.69)	0.67	0.69 (0.50-0.97)	0.032*
Induration/Edema	0.65 (0.37-1.17)	0.15	0.55 (0.34-0.89)	0.014*

*Statistically significant P values (<0.05)

¹Individual effect sizes (ORs) and 95% CIs refer to the comparison of the mild and severe severity group with the moderate scar severity group as reference group for the outcome

We developed and validated three severity prediction models and one VSS score regression model: (i) image-based severity prediction model, which integrated CBAM with CNN architecture, (ii) clinical-data-based severity prediction model which uses MLP model with clinical variables, (iii) combined severity prediction model, which is derived from the weighted sum of the model (i) and (ii), and (iv) image-based regression model to predict VSS score. The results for the sensitivity, specificity, ROC-AUC, and Top-1 accuracy of the severity prediction models are listed in Table 4. In the internal test dataset, the image-based model had a ROC-AUC of 0.931 (95% CI 0.910-0.949), the clinical-data-based model had a ROC-AUC of 0.905 (95% CI 0.877-0.928), and combination of these two model yields ROC-AUC of 0.938 (0.916-0.955). The combined severity prediction model significantly improved ($P = 0.042$) compared with the clinical-data-based model but was statistically insignificant compared to the image-based model ($P = 0.633$). Trends were similar in the external test dataset, yet slightly lower ROC-AUC and Top-1 accuracy was noted than the corresponding values in the internal test set (Figure 3).

The sensitivity, specificity, F1-score, and ROC-AUC of each severity class in the internal testing set was shown in Table 5. ROC-AUC was highest in the normal (0.998, 95% CI 0.994-0.999), followed by severe (0.925, 95% CI 0.884-0.954), mild (0.919, 95% CI 0.880-0.951), and moderate (0.833, 95% CI 0.781-0.878) class in the image-based model. The trend and value was similar in the combined severity prediction model; highest in the normal (0.996, 95% CI 0.988-1.000) and lowest in the moderate (0.834, 95% CI 0.781-0.882) group. In combined model, specificity of each classification were highest at normal grade (0.983, 95% CI 0.962-1.000), followed by severe (0.950, 95% CI 0.914-0.979), mild (0.944, 95% CI 0.909-0.977), and moderate (0.761, 95% CI 0.699-0.823) grade. The sensitivity of each grade tends to be lower than those of specificity, especially in the mild (0.650, 95% CI 0.518-0.762) and severe (0.617, 95% CI 0.500-0.731) groups.

In the case of the regression model for VSS score prediction, we obtained the mean absolute error (MAE), root mean square error (RMSE), and the Bland-Altman plot depicting the association between the predicted and measured VSS. The MAE of the

Table 4. Performance of severity prediction models¹

Model (class)	Sensitivity (95% CI)	Specificity (95% CI)	ROC-AUC (95% CI)	Top-1 accuracy (95% CI)	P value²
Internal testing set					
Image-based model	0.725 (0.672-0.774)	0.908 (0.888-0.926)	0.931 (0.910-0.949)	0.725 (0.667-0.780)	0.633
Clinical-data-based model	0.692 (0.638-0.750)	0.897 (0.879-0.917)	0.905 (0.877-0.928)	0.692 (0.638-0.750)	0.042
Combined model	0.730 (0.675-0.783)	0.910 (0.892-0.928)	0.938 (0.916-0.955)	0.729 (0.675-0.783)	ref
External testing set					
Image-based model	0.695 (0.652-0.741)	0.898 (0.884-0.914)	0.896 (0.874-0.916)	0.695 (0.652-0.741)	0.260
Clinical-data-based model	0.658 (0.610-0.706)	0.886 (0.870-0.902)	0.875 (0.848-0.899)	0.658 (0.610-0.706)	0.023
Combined model	0.733 (0.687-0.775)	0.911 (0.896-0.925)	0.912 (0.892-0.932)	0.733 (0.687-0.775)	ref

¹Calculated by the micro-averaged value of each severity class for the given model, using bootstrap resampling (N=1000) of the test dataset.

²The p-value from the binomial test measures the difference in performance between the combined model and image- or clinical-data-based model in terms of ROC-AUC.

Abbreviations: ROC-AUC; area under the receiver operating characteristic curve, Ref; reference model.

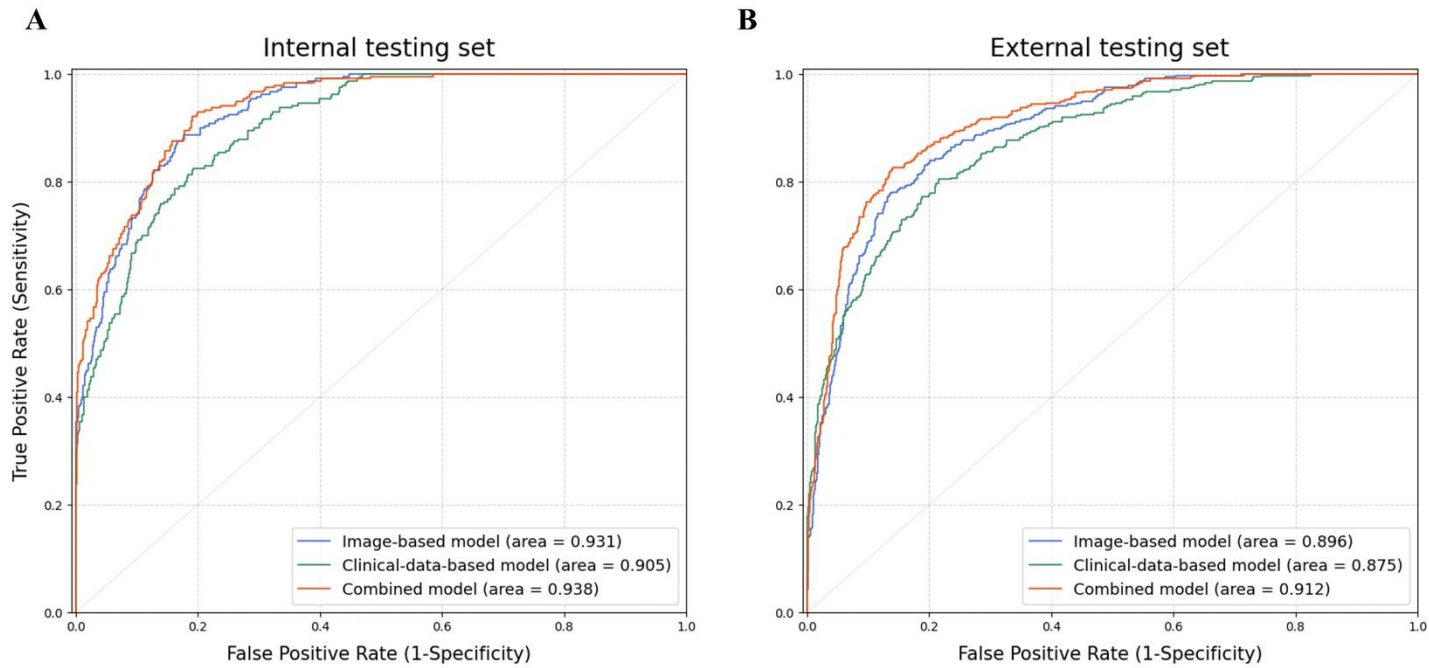


Figure 3. Receiver operating characteristic (ROC) curves for the three classification models in (A) internal testing set and (B) external testing set. Blue curve: image-based model by convolutional block attention module (CBAM) integrated Resnet-50, Green curve: clinical-data-based model by multilayer perceptron (MLP), Red curve: combined model from the weighted sum of the image-based and clinical-data-based models.

Table 5. Performance of prediction model according to each severity class in the internal testing set

Model (class)	Sensitivity (95% CI)	Specificity (95% CI)	F1-score (95% CI)	ROC-AUC (95% CI)
Image-based model				
Normal	0.950 (0.889-1.000)	0.983 (0.961-1.000)	0.950 (0.906-0.984)	0.998 (0.994-0.999)
Mild	0.617 (0.500-0.729)	0.944 (0.904-0.976)	0.692 (0.593-0.781)	0.919 (0.880-0.951)
Moderate	0.817 (0.712-0.914)	0.728 (0.667-0.788)	0.620 (0.525-0.703)	0.833 (0.781-0.878)
Severe	0.517 (0.390-0.638)	0.978 (0.954-0.995)	0.653 (0.530-0.750)	0.925 (0.884-0.954)
Clinical-data-based model				
Normal	1.000 (1.000-1.000)	1.000 (1.000-1.000)	1.000 (1.000-1.000)	1.000 (0.999-1.000)
Mild	0.817 (0.714-0.911)	0.811 (0.757-0.865)	0.685 (0.590-0.770)	0.889 (0.840-0.928)
Moderate	0.467 (0.333-0.593)	0.900 (0.853-0.944)	0.528 (0.404-0.644)	0.808 (0.745-0.862)
Severe	0.483 (0.362-0.614)	0.878 (0.828-0.922)	0.522 (0.409-0.632)	0.805 (0.747-0.858)
Combined model				
Normal	0.967 (0.918-1.000)	0.983 (0.962-1.000)	0.959 (0.917-0.991)	0.996 (0.988-1.000)
Mild	0.650 (0.518-0.762)	0.944(0.909-0.977)	0.716 (0.608-0.797)	0.932 (0.894-0.962)
Moderate	0.683 (0.571-0.797)	0.761 (0.699-0.823)	0.569 (0.470-0.671)	0.834 (0.781-0.882)
Severe	0.617 (0.500-0.731)	0.950 (0.914-0.979)	0.698 (0.600-0.790)	0.928 (0.895-0.959)

internal testing set was 1.075 (CI 0.960-1.184), and RMSE was 1.418 (CI 1.269-1.563). These values were slightly higher in the external testing set, 1.183 (CI 1.080-1.283) for MAE and 1.561 (CI 1.431-1.680) for RMSE. Bland-Altman plot showed a positive linear slope, indicating a positive proportional bias (Figure 4).

Five-fold stratified cross-validation was done, and the Top-1 accuracy of the image-based and combined model fluctuated in the range of $\pm 1.6\%$ and $\pm 4.0\%$, respectively, showing the robustness of the models.

3. Comparison between the neural network vs. dermatologists

Our model is tested against eight board-certified dermatologists and eight dermatology residents to compare performances. The overall Top-1 accuracy of the board-certified dermatologist and dermatology resident was 0.746 and 0.729, respectively. Both image-based and combined models were able to classify four scar severity groups with a level of competence comparable to that of dermatologists (Figure 5). Confusion matrices of neural network models and dermatologists over the four severity classes were shown in Figure 6. Both AI models and the dermatologists significantly confused mild and moderate scar lesions with each other; The model had a slight higher rate of misclassifying mild severity as moderate (7.9% vs. 4.4%) than human, whereas humans had a higher rate of misclassifying moderate as mild (7.9% vs. 3.5%). Also, both models and dermatologists tended to misclassify severe lesions into moderate class (10.8% and 11.0%, respectively).

4. Visualization of the explanatory model

We adopted two visualization methods for image-based model, dimensionality reduction via t-SNE and class activation mapping (Grad-CAM). The two-dimensional expression of the internal features extracted from the image-based classification model is shown in Figure 7. Neural network model could extract distinct features for scar severity classification, and the cluster represented in each class occupied relative regions in the two-

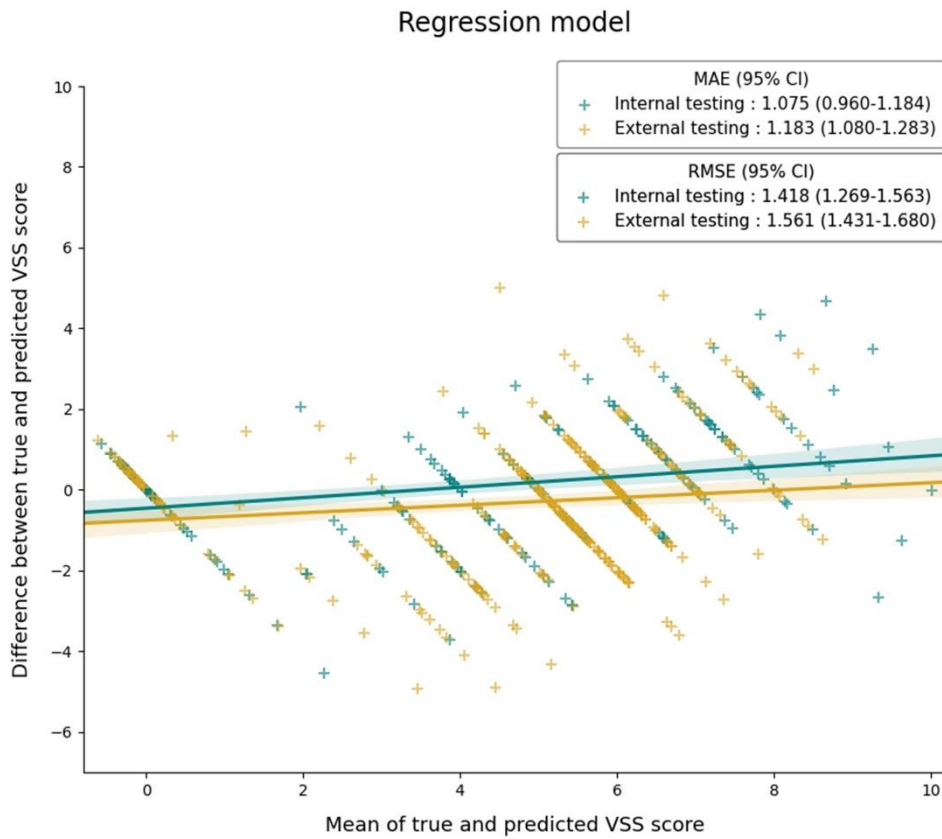


Figure 4. A Bland-Altman plot shows the association between the measured and predicted VSS score in the regression model. The shaded areas correspond to 95% confidence intervals. MAE, mean absolute error; RMSE, root mean square error.

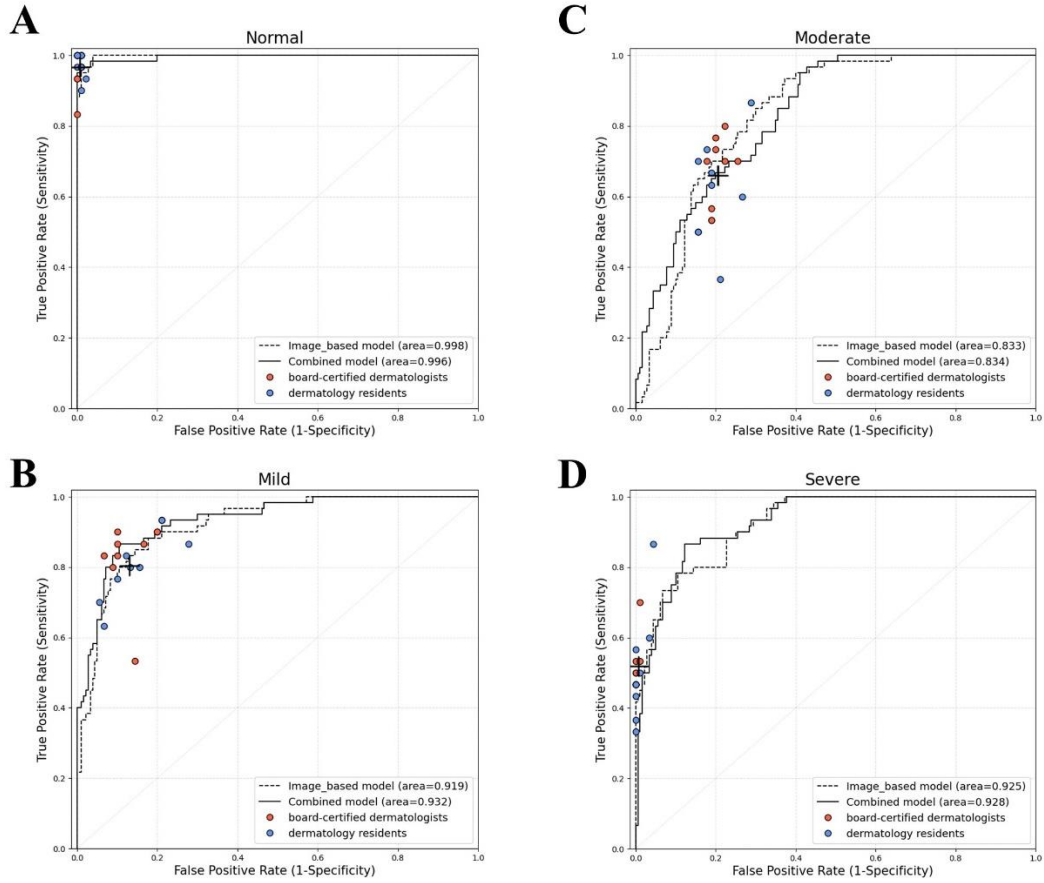


Figure 5. Scar severity classification performance of the CNN and dermatologists. ROC (receiver operating characteristic) curves for each severity class were drawn for the image-based (dotted curve) and combined prediction model (black curve). Also, the prediction value of the 16 dermatologists was plotted; red dot = 8 board-certified dermatologists; blue dot = 8 dermatology residents; black cross = average value of 16 dermatologists.

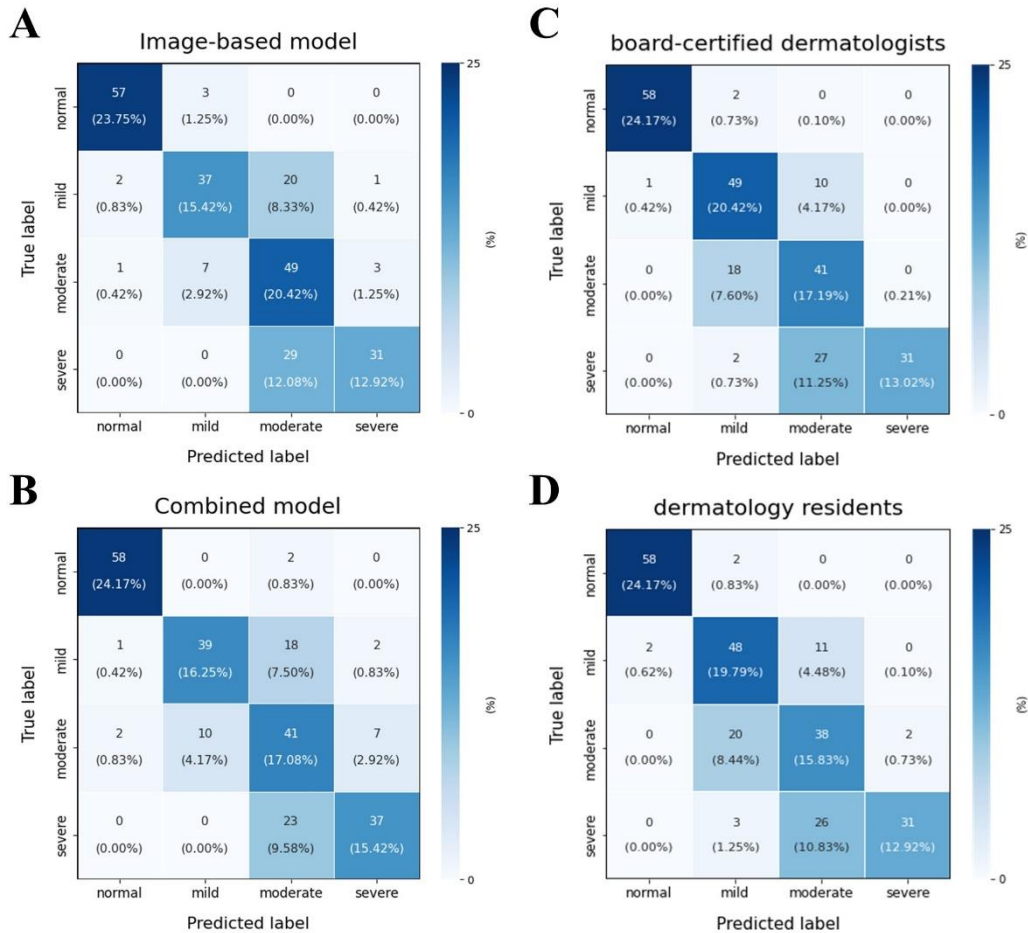


Figure 6. Confusion matrix comparison between prediction models and dermatologists. Confusion matrices were drawn for (A) image-based model, (B) combined prediction model, (C) board-certified dermatologists, and (D) dermatology residents. All matrices are computed using the 240 images from the internal test set. True label in y axis refers to ground truth label, and predicted label in x axis refers to classification output by AI models or dermatologists.

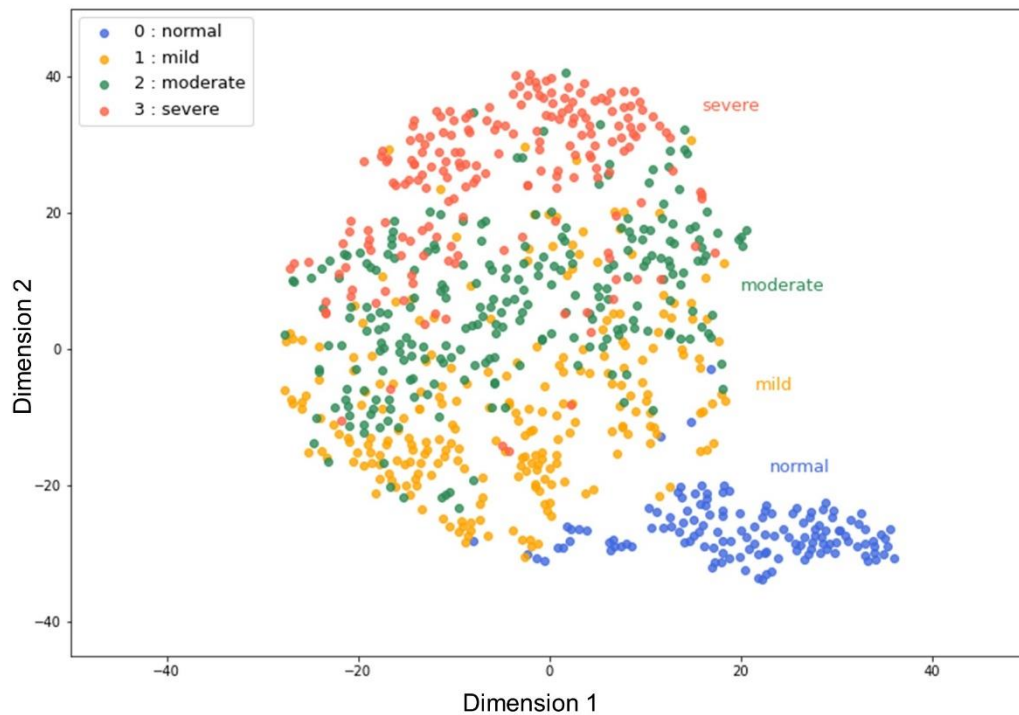


Figure 7. t-SNE visualization of the last hidden layer representations in the image-based prediction model. The output of the neural network's last hidden layer projected onto a 2-dimensional map using the t-distributed Stochastic Neighbor Embedding (t-SNE) method. Colored point clouds represent different severity classifications, showing how the algorithm clusters the postoperative scars.

dimensional map corresponding to clinical features. For example, the cluster of the mild class is located between normal and moderate severity, and the moderate class is sandwiched between mild and severe classes with some overlapping.

Figure 8 shows the results from the class activation mapping, in which heatmaps represent the pixel areas activated by the deep neural network. The CBAM integrated CNN model successfully distinguished postoperative scars from the wrinkles of the surrounding skin. Also, it can detect coarse and hypertrophic portions of the lesion in the moderate or severe degree scar.

Also, to elucidate significant variables in predicting the outcome of the clinical-data-based model, we introduced the SHAP (SHapley Additive exPlanations) method for visualizing the importance ranking of the features²³. Figure 9 shows the importance ranking of all variables used in the clinical-data-based model evaluated by the average absolute SHAP value. Operation site, induration/edema, date after surgery, BMI, and itching/pain were considered the Top-5 dominating features for predicting the severity of the scar.

IV. DISCUSSION

All undesirable scars are undesirable in different ways²⁴; thus, it is hard to differentiate 'undesirable' scars on a clinical basis easily. Various scar assessment scales for clinicians have been emerged to assist the evaluation of scar severity, progression, and treatment response, yet a “gold standard” scar scale still does not exist²⁵. In this study, we aimed to evaluate the postoperative scars using the deep neural network models by the scar severity. Using the AI model based on the patients' digital images and clinical information, we successfully classified the postoperative scars according to their severity, and the performance of the models was comparable to those of dermatologists.

It should be noted that we intentionally collected and cropped digital images to include not only scar but also adjacent skin structure and even artifacts like clothes or ruler (Figure 1). Intensive preprocessing, including resizing and cropping the clinical image to include

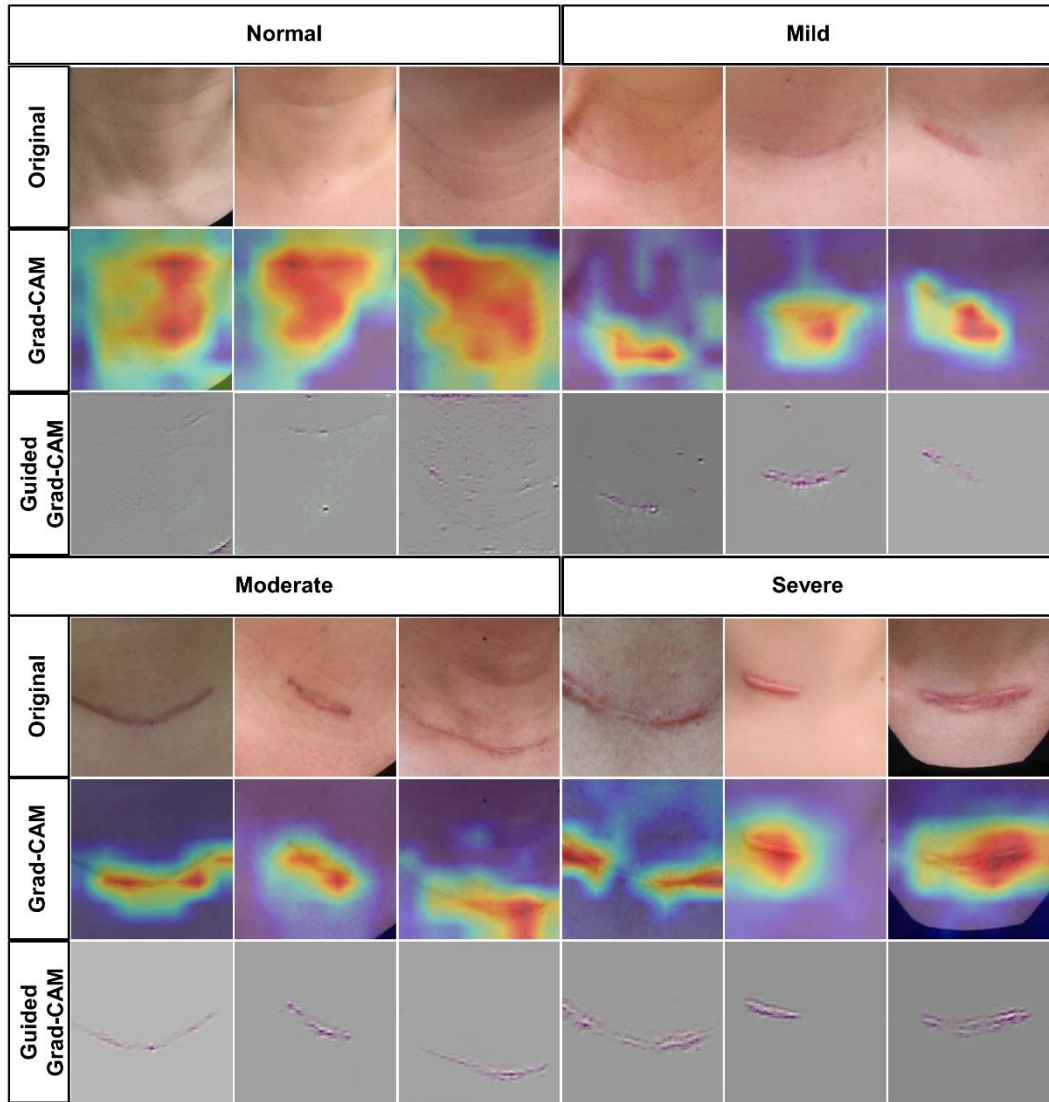


Figure 8. Visual explanations of postoperative scar cases via class activation mapping. Clinical images of each scar severity grade and corresponding heatmaps via gradient-based localization (Grad-CAM). The activation was focused on the hypertrophied region of the scar.

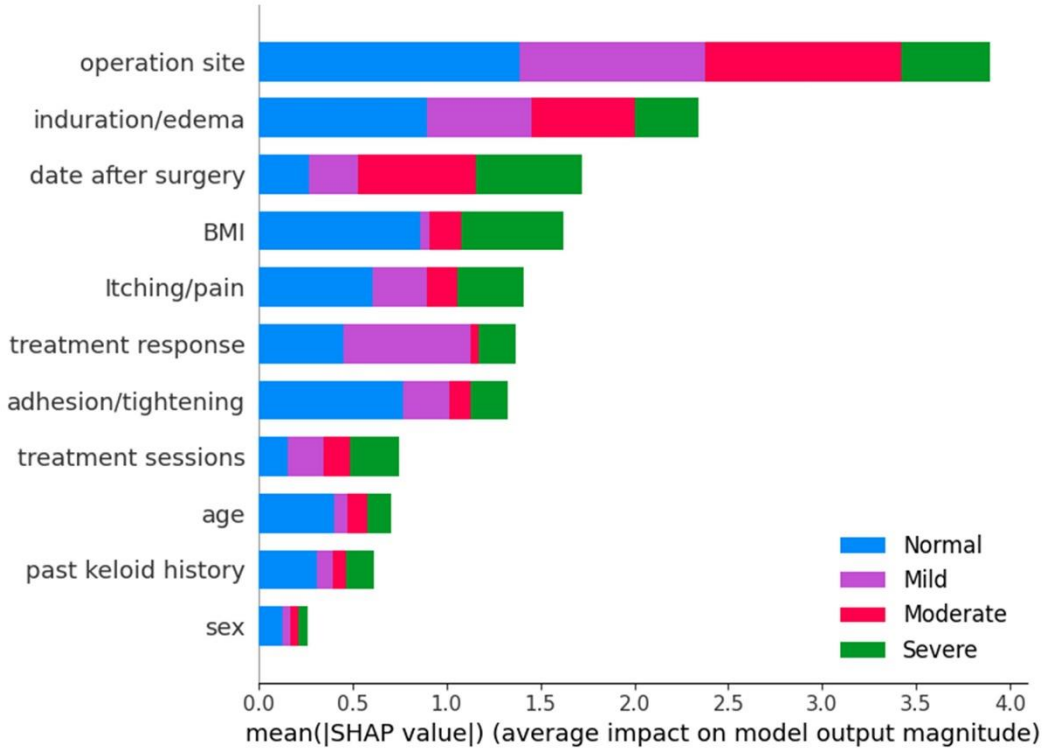


Figure 9. Interpretation of the clinical-data-based model via SHAP analysis. The importance ranking of variables used in clinical-date-based model according to the mean (|SHAP value|).

only lesions of interest for analysis, may help to improve classification performance. However, this is not only such a laborious and exhausting task but also far from the actual clinician's viewpoint of scar, which usually incorporates broader adjacent anatomical structures^{6,26}. Thus, we integrated the convolutional block attention module (CBAM) into the CNN architecture, which selectively and automatically focuses on the salient lesion like the human visual perception mechanism^{19,27}. Therefore, our image-based model successfully classified scar severity while appropriately concentrating on the lesion of interest (Figure 7) without direct human labeling or cropping of the scar lesion.

To construct image-based AI model, we classified the postoperative scar into four classes mainly based on the Vancouver scar scale (VSS), which was the first validated and most widely used scar scale to date¹⁷. The VSS consisted of four parameters related to scar characteristics : height, pliability, pigmentation, and vascularity, to generate a semi-quantitative score ranging from 0 - 13 points²⁸. However, the VSS has a significant limitation: it does not reflect various factors that determine scar severity other than the morphological characteristics of the scar^{17,25}. Therefore, we planned to develop a neural network model trained with eleven clinical variables related to postoperative scars, including patient's demographic factors, subject symptoms, local complications, and scar age. The AI model based on clinical variables showed considerable performance in predicting the severity of post-operative scars; however, significantly lower than those with the combination of clinical variables and medical images. These results imply the importance of utilizing both scar-related clinical characteristics and morphological features when predicting the severity of the postoperative scar. Furthermore, we adopted SHAP analysis to clarify the influential clinical features for predicting the severity of the postoperative scar, and to give a plausible interpretation of the model's decision-making process. The SHAP method showed the most critical risk factors for post-operative hypertrophic scarring, such as the location of the scar, increased BMI, and presence of subjective symptoms. These results are in line with the multinomial logistic regression analysis and previous studies of risk factors of post-operative scars^{3,6,29}.

AI has performed at least equal to or superior to dermatologists for the diagnosis or classification tasks for various skin diseases^{8,9,30,31}. Our deep neural network model also showed comparable performance to board-certified dermatologists or dermatology residents for classifying postoperative scars by their severity. We need to take into account the nature of the classification task in this study; not to distinguish the different diseases, but to grade the severity of the same disorder. Considering the semi-quantitative, rater-dependent, and subjective nature of current scar-grading system²⁴, significant ambiguity and overlap were expected between the classification classes used in this study. Interestingly, the confusion matrices revealed striking similarities in misclassification between human and neural network models. Both AI models and dermatologists tend to misclassify mild or severe classes into moderate severity. One plausible reason for this phenomenon is insufficiently distinctive features of an intermediate grade than other severity groups³², the other lies in the central tendency bias of visual perception, which is likely to estimate towards the mean of the stimuli³³.

There are several limitations in our study. First, the AI model in our study showed decreased performance on the external testing set, compared to the internal testing set. This could have been due to different image acquisition settings between the different hospitals. Since the VSS has two components directly related to the color of the image (pigmentation and vascularity), slight differences in input in the color (RGB) channels by individual camera settings might create substantial changes to the output of the model³⁴. Second, due to the study's retrospective design, data imbalance in the training dataset and possible selection bias might disturb the application of this study to the broader general population who have the postoperative scar. Moreover, although several studies assess scar scales with a photograph-based examination by scar-specialized clinicians^{24,35,36}, some components of VSS (i.e., pliability or height) might have difficulty being evaluated only by the clinical images without examination of live scars.

A third limitation is that we included the post-operative scar at two different anatomical sites, anterior neck and axillary area. Since anatomical location is one of the

major predisposing factors for hypertrophic scarring²⁶, though axillary scars were included as small as 6.4% of the main dataset, it could affect the model's overall performance as a confounding factor. However, we believe that the involvement of different anatomical backgrounds and directions of scar could prevent the overfitting of the model and enhance generalization to other types of scar.

Lastly, our study cohorts exclusively included Korean patients; hence subjects only with Fitzpatrick skin type III and IV were included in the dataset. Since darker skin type is known to be one of the predisposing factors for hypertrophic scar²⁶, future studies with larger-scale datasets from different ethnic groups with various etiology of the scar will be warranted.

V. CONCLUSION

In conclusion, artificial intelligence model based on image and clinical data could predict the severity of postoperative scar. Though our neural network models were trained with a relatively small (< 5000) number of images, they efficiently classified the severity of postoperative scar lesions with the comparable performance of dermatologists. These models could aid clinicians, both specialized or not-specialized in scar management, in determining the severity of scars and setting the treatment decision. Moreover, our established dataset of post-operative scar is expected to extend to the other types of scars (i.e., burn, trauma, post-infectious, etc.) in the future study.

REFERENCES

1. Lewis WH, Sun KK. Hypertrophic scar: a genetic hypothesis. *Burns*. 1990;16(3):176-178.
2. Balci DD, Inandi T, Dogramaci CA, Celik E. DLQI scores in patients with keloids and hypertrophic scars: a prospective case control study. *J Dtsch Dermatol Ges*. 2009;7(8):688-692.
3. Shin JU, Park JH, Oh SH, et al. Early intervention in thyroidectomy scars: demographics, symptoms, and prevention. *J Wound Care*. 2015;24(4):163-164, 166-168, 170-161.
4. Ogawa R. Keloid and Hypertrophic Scars Are the Result of Chronic Inflammation in the Reticular Dermis. *Int J Mol Sci*. 2017;18(3).
5. Xie H, Xiang Y, Yang E, Zhang H. Factors Influencing Hypertrophic Scarring after Thyroidectomy. *Adv Skin Wound Care*. 2021;34(10):1-6.
6. Kim JH, Sung JY, Kim YH, et al. Risk factors for hypertrophic surgical scar development after thyroidectomy. *Wound Repair Regen*. 2012;20(3):304-310.
7. Puri P, Comfere N, Drage LA, et al. Deep learning for dermatologists: Part II. Current applications. *J Am Acad Dermatol*. 2020.
8. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542(7639):115-118.
9. Han SS, Kim MS, Lim W, Park GH, Park I, Chang SE. Classification of the Clinical Images for Benign and Malignant Cutaneous Tumors Using a Deep Learning Algorithm. *Journal of Investigative Dermatology*. 2018;138(7):1529-1538.
10. Han SS, Moon IJ, Lim W, et al. Keratinocytic Skin Cancer Detection on the Face Using Region-Based Convolutional Neural Network. *JAMA Dermatology*. 2020;156(1):29-37.
11. Tschandl P, Rosendahl C, Akay BN, et al. Expert-Level Diagnosis of Nonpigmented Skin Cancer by Combined Convolutional Neural Networks. *JAMA*

- Dermatology*. 2019;155(1):58-65.
12. Lee S, Lee JW, Choe SJ, et al. Clinically Applicable Deep Learning Framework for Measurement of the Extent of Hair Loss in Patients With Alopecia Areata. *JAMA Dermatol*. 2020;156(9):1018-1020.
 13. McNeil A, Parks K, Liu X, et al. Artificial intelligence recognition of cutaneous chronic graft-versus-host disease by a deep learning neural network. *British Journal of Haematology*. 2022;197(6):e69-e72.
 14. Wu JT-y, de la Hoz MÁA, Kuo P-C, et al. Developing and Validating Multi-Modal Models for Mortality Prediction in COVID-19 Patients: a Multi-center Retrospective Study. *Journal of Digital Imaging*. 2022.
 15. Xu Y, Hosny A, Zeleznik R, et al. Deep Learning Predicts Lung Cancer Treatment Response from Serial Medical Imaging. *Clin Cancer Res*. 2019;25(11):3266-3275.
 16. Jiao Z, Choi JW, Halsey K, et al. Prognostication of patients with COVID-19 using artificial intelligence based on chest x-rays and clinical data: a retrospective study. *The Lancet Digital Health*. 2021;3(5):e286-e294.
 17. Park JW, Koh YG, Shin SH, et al. Review of Scar Assessment Scales. *Medical Lasers*. 2022;11(1):1-7.
 18. Signorini M, Clementoni MT. Clinical evaluation of a new self-drying silicone gel in the treatment of scars: a preliminary report. *Aesthetic Plast Surg*. 2007;31(2):183-187.
 19. Woo S, Park J, Lee J-Y, Kweon IS. Cbam: Convolutional block attention module. Paper presented at: Proceedings of the European conference on computer vision (ECCV)2018.
 20. Takahashi R, Matsubara T, Uehara K. Data augmentation using random image cropping and patching for deep CNNs. *IEEE Transactions on Circuits and Systems for Video Technology*. 2019;30(9):2917-2931.
 21. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization. Paper

- presented at: Proceedings of the IEEE international conference on computer vision2017.
22. Sanchez-Lengeling B, Wei JN, Lee BK, Gerkin RC, Aspuru-Guzik A, Wiltchko AB. Machine learning for scent: Learning generalizable perceptual representations of small molecules. *arXiv preprint arXiv:191010685*. 2019.
 23. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. *Advances in neural information processing systems*. 2017;30.
 24. Kantor J. Reliability and Photographic Equivalency of the Scar Cosmesis Assessment and Rating (SCAR) Scale, an Outcome Measure for Postoperative Scars. *JAMA Dermatol*. 2017;153(1):55-60.
 25. Nguyen TA, Feldstein SI, Shumaker PR, Krakowski AC. A review of scar assessment scales. *Semin Cutan Med Surg*. 2015;34(1):28-36.
 26. Nabai L, Pourghadiri A, Ghahary A. Hypertrophic Scarring: Current Knowledge of Predisposing Factors, Cellular and Molecular Mechanisms. *J Burn Care Res*. 2020;41(1):48-56.
 27. Larochelle H, Hinton GE. Learning to combine foveal glimpses with a third-order Boltzmann machine. *Advances in neural information processing systems*. 2010;23.
 28. Sullivan T, Smith J, Kermode J, McIver E, Courtemanche DJ. Rating the burn scar. *J Burn Care Rehabil*. 1990;11(3):256-260.
 29. On HR, Lee SH, Lee YS, Chang HS, Park C, Roh MR. Evaluating hypertrophic thyroidectomy scar outcomes after treatment with triamcinolone injections and copper bromide laser therapy. *Lasers Surg Med*. 2015;47(6):479-484.
 30. Han SS, Park GH, Lim W, et al. Deep neural networks show an equivalent and often superior performance to dermatologists in onychomycosis diagnosis: Automatic construction of onychomycosis datasets by region-based convolutional deep neural network. *PLoS One*. 2018;13(1):e0191493.
 31. Fujisawa Y, Otomo Y, Ogata Y, et al. Deep-learning-based, computer-aided classifier developed with a small dataset of clinical images surpasses board-

- certified dermatologists in skin tumour diagnosis. *Br J Dermatol*. 2019;180(2):373-381.
32. Lim ZV, Akram F, Ngo CP, et al. Automated grading of acne vulgaris by deep learning with convolutional neural networks. *Skin Research and Technology*. 2020;26(2):187-192.
33. Aston S, Negen J, Nardini M, Beierholm U. Central tendency biases must be accounted for to consistently capture Bayesian cue combination in continuous response data. *Behavior Research Methods*. 2022;54(1):508-521.
34. Cha D, Pae C, Seong S-B, Choi JY, Park H-J. Automated diagnosis of ear disease using ensemble deep learning with a big otoendoscopy image database. *EBioMedicine*. 2019;45:606-614.
35. Thompson CM, Sood RF, Honari S, Carrougher GJ, Gibran NS. What score on the Vancouver Scar Scale constitutes a hypertrophic scar? Results from a survey of North American burn-care providers. *Burns*. 2015;41(7):1442-1448.
36. Lee YI, Kim J, Yang CE, Hong JW, Lee WJ, Lee JH. Combined Therapeutic Strategies for Keloid Treatment. *Dermatol Surg*. 2019;45(6):802-810.

ABSTRACT(IN KOREAN)

흉터의 중증도 평가 및 예후를 예측하는 인공지능모델 개발

<지도교수 이 주 희>

연세대학교 대학원 의학과

김 제 민

대부분의 수술 후 흉터는 수술 이후의 불가피한 휴유증이지만 심각한 미용적 문제와 기능적 손상을 유발할 수 있다. 흉터 중증도에 대한 적절한 평가가 적절한 치료 방법을 결정하는 데 중요하지만, 흉터를 평가하는 데 있어 표준적인 방법은 없는 실정이다. 따라서 본 연구는 수술 후 흉터의 중증도를 예측하기 위하여 이미지 및 임상 데이터를 기반으로 한 인공지능 모델을 개발하고 평가하고자 하였다.

갑상선 절제술 후 흉터가 있어 내원한 1,283명의 환자 (주 데이터셋: 1,043, 외부 데이터셋: 240)로부터 사진 및 임상 정보를 수집하여 심층 신경망 모델 (deep neural network) 을 훈련하고 검증하였다. 흉터 중증도를 분류하는 인공지능 모델의 성능은 다른 병원 환자군의 자료를 대상으로 외부적으로 검증하였으며, 16명의 피부과 의사에게 흉터 사진을 평가받아 성능을 비교하였다. 내부 테스트 세트 (external test set) 에서 이미지 기반 모델의 ROC-AUC (Receiver Operating Characteristic Curve) 면적은 0.931 (95% 신뢰구간 0.910-0.949) 이었고, 임상 데이터와 결합하면 0.938 (0.916-0.955)으로 증가하였다. 외부 테스트 세트 (external test set) 에서 이미지 기반 및 결합 예측 모델의 ROC-AUC 면적은 각각 0.896 (0.874-0.916) 및 0.912 (0.892-0.932)으로 계산되었다. 내부 테스트 세트의 이미지를 기반으로 평가한 알고리즘의 성능은 16명의 피부과 의사와 비교하였을 때 유사하였다. 밴쿠버 흉터 지수 (Vancouver scar scale)를 기준으로 한 회귀 예측 모델의 경우 평균 절대 오차 (mean

absolute error) 가 내부 테스트 세트에서 1.075 (95% 신뢰구간 0.960-1.184), 내부 테스트 세트에서 1.183 (95% CI 1.080-1.283)으로 측정되었다.

결론적으로, 본 연구는 이미지 및 임상 자료로부터 도출된 심층 신경망 모델이 수술 후 흉터의 중증도를 예측할 수 있음을 보여주었다. 본 연구에서 제안하는 인공지능 모델은 추후 흉터 관리의 임상 분야, 특히 흉터 중증도를 평가하여 치료 시점을 결정하는 데 활용할 수 있을 것으로 기대한다.

핵심되는 말 : 수술 후 흉터, 비후성 흉터, 데이터 세트, 인공지능, 딥러닝 (deep learning)