

Grading and Monitoring Pressure Injury Using Machine Learning Approach and a Web-Based Application

Author

Yu Cheng Lai¹; Peng Yuan Chen, MSc²; Cheng Yang Liu, MSc²; Shih Chung Kang, PhD^{2,3}; Pei-Hung Liao, RN, PhD⁴; Chia-Ching Chou, PhD¹

Affiliations:

¹Institute of Applied Mechanics, National Taiwan University, Taipei City, Taiwan

²Jubo Co., Taipei City, Taiwan

³Department of Civil & Environmental Engineering, University of Alberta, Edmonton, Alberta, Canada

⁴National Taipei University of Nursing and Health Sciences, Taipei City, Taiwan

Corresponding authors:

Chia-Ching Chou, Ph.D. Institute of Applied Mechanics, National Taiwan University,
No.1, Sect. 4, Roosevelt Rd., Da'an Dist., Taipei City 10617, Taiwan

ccchou@iam.ntu.edu.tw

ORCID:

Yu Cheng Lai: 0000-0002-4861-0604

Abstract

Background:

Classification of pressure injuries (PI) is a critical task for long-term care organizations. Incorrect classification may lead to inappropriate treatment and cause serious infection due to the rapid development of the ulcer. In addition, the changing standard of the classification system over the years requires caregivers to adapt to the new standard fast and accurately. The machine learning approach provides a promising opportunity to help and improve the efficiency and precision of ulcer classification in long-term care. To enhance accessibility, smartphone devices and online services are merged with our deep learning model.

Objective:

We aim to develop a deep learning model that learns and generates patterns from expert labeling to help caregivers evaluate PI images and make decisions on follow-up treatment. Furthermore, we take advantage of the multifunctional smartphone as the platform of the deep learning model so that a photo can be uploaded and graded through the smartphone devices of users.

Methods:

505 pieces of PI photos were collected from patients admitted to 40 long-term care organizations from 2018 to 2020. The four stages of pressure injury in the photos were classified by three registered nurses and taken as ground truth. Deep learning CNN models, VGG16, are used to predict the stage of pressure injury. Performance of our model was evaluated using accuracy, area under the receiver operating characteristic curve (AUROC), confusion matrix, and heatmap. The web-based application was further developed and integrated with deep learning models. An external validation was carried out in the same organizations from 10/2020 to 05/2021 by caregivers to check whether our system performs well in clinical applications.

Results:

The highest accuracy in 5-fold stratified cross-validation is reported as 86.67%, and the accuracy of external validation is calculated as 85.16%. Confusion matrix, AUROC, and heatmaps are evaluated to analyze the performance of our model. The AUROC of the four stages are calculated to be 0.93, 0.91, 0.85, and 0.95. Highlighted areas in heatmaps show that the perspective of our model is in close relationship with clinical knowledge.

Conclusions:

We developed a deep learning model to classify PI images into four stages with an accuracy of 85.16%. The model integrated with the web-based application can help caregivers evaluate patient injuries and provide appropriate treatment in time.

Keywords: Supervised Machine Learning; Smartphone; Mobile Phone; Wounds and Injuries; Pressure Injuries; Monitoring

Introduction

Background

Pressure injuries (PI) are a common problem in all health care settings, mainly developed by inducing constant pressure on the capillary. Most PI, observed in an immobilized patient, develop rapidly in 2 to 6 hours [1]. If the ulcer remains untreated, the resulting necrosis tissue can lead to serious infection and could threaten the patient's life. Therefore, the classification of a PI is critical for all caregivers. Furthermore, the classification system continues to be refined throughout the years due to a deeper understanding of this injury[2-4]. In other words, the ability to quickly adapt to a new standard is critical for caregivers, or we can use the powerful learning ability of computers, machine learning (ML), to perform the job.

ML gives computers the ability to automatically learn from input data and improve from experience without being explicitly programmed [5]. Some fundamental ML methods, for example artificial neural network (ANN) [6], decision trees (DT), or support vector machine (SVM) [7], have achieved outstanding achievements in various aspects of engineering in the past few years. When the number of layers is increased, the accuracy and performance of our model can be improved. This method is called 'deep learning(DL)' [8]. Applications based on deep learning include image classification [9], image segmentation [10], natural language processing [11], autonomous driving [12], and image generation [13].

Medical imaging analysis is one of the aspects where ML has achieved great success [14-16]. Among these applications, PI classification is one of the most difficult tasks due to heterogeneous colorations in PI images, which are related to the skin color of the patient and other various anomalies that may appear in the images, such as erythema and skin striations. In addition, the boundaries between the ulcer and the unwanted surrounding skin are hard to determine, making the segmentation task inaccurate and complex.

Prior work

Many tests have been performed in the past to achieve tissue segmentation and ulcer measurement of a PI image [17-20]. However, the direct classification of the stage of a PI image based on any known standard is limited due to the complexity of an image. The closest study was conducted in 2007 by Dimitrios I. Kosmopoulos et al. [21] with the SVM method, which is a powerful but improvable method for image analysis. Other research focuses on the binary classification of a wound type. Varun Shenoy et al. [22] proposed a convolutional neural network (CNN) model to classify different types of wound and achieve an accuracy over 70% in all cases. However, their model was designed to be a multi-binary classification system, which is different from our multi-class classification task. Moreover, Mengyao Jiang et al. point out that most machine learning-based predictive models lack external validation [23], thus the stability of these models could remain unknown when in practical use. In a study of mobile applications used for PI, Janine Koepp et al. [24] observed that the applications remain mainly in the initial phase. The prescription of an application for the identification, evaluation, treatment, or prevention of PI in adults still leave room for improvement. According to

Janine Koepp et al. [24], there are few studies of application that combine PI image classification, patient management, and treatment tracking.

The Objective of this Study

In this study, our aim is to develop a DL model that can classify the four stages of PI by a single photograph and implement this model into a web-based application. Through this study, caregivers and patients with PI may both benefit from the convenience and accuracy that this application brings.

Methods

Datasets

In the classification task, 527 images are used to train and validate the model. They are collected by caregivers in 40 long-term care organizations from 2018 to 2020 through smartphones. These images satisfy one restriction: The length and width of an image must be larger than 224 pixels. After collection, they are classified into four stages with the agreement of 3 professional nurses. Our classification standard is adapted from the grading standard set by the National Pressure Injury Advisory Panel (NPIAP) in 2016 [2]. The result of our dataset is presented in the next section.

Deep Learning Model and Training Strategy

Among various CNN models, VGG16 [25] is selected as the baseline model. Vgg16 is a 16-depth weight layer network, including 13 convolutional layers and 3 fully connected layers. The weights learnt from ImageNet are induced to improve the task of classifying the nature image. This technique is often called transfer learning [26]. The following 2 fully-connected layers which are on top of the convolutional layers are not freeze, and their sizes are reduced from 4096 to 1024 to prevent early overfitting. Batch normalization is added before the sigmoid function to prevent the output values from falling in the saturation region of the activation function. A dropout layer with a rate of 0.5 is added to prevent overfitting. Finally, a dense layer of size 4 along with the SoftMax function is used for classification. The detailed structure is shown in Table 1.

We perform 5 folds in stratified k-fold cross validation[27] with a training validation ratio of 4:1. The batch size is set to 16. Adam optimizer [28] is used with an initial learning rate of 0.001. The learning rate will be reduced by 0.5 if the validation loss does not decrease in 5 epochs and will stop decreasing until it reaches 10E-13. Categorical cross-entropy is selected as our loss function. We set 300 epochs in each fold and save the model only when the validation accuracy increases. Each model outputs four values, and each value means the probability of an image being classified into a corresponding stage. The stage corresponding to the highest probability will be the final stage of the input image.

Table 1. Architecture of the modified vgg16 model in our classification task.

Layer name	Layer details ^a	Output size
Input	random size	(224,224,3)
Block1_conv	Conv3-64 Conv3-64	(224,224,64)
Block1_pool	Maxpooling (2D)	(112,112,64)
Block2_conv	Conv3-128 Conv3-128	(112,112,128)
Block2_pool	Maxpooling (2D)	(56,56,128)
Block3_conv	Conv3-256 Conv3-256 Conv3-256	(56,56,256)
Block3_pool	Maxpooling (2D)	(28,28,256)
Block4_conv	Conv3-512 Conv3-512 Conv3-512	(28,28,512)
Block4_pool	Maxpooling (2D)	(14,14,512)
Block5_conv	Conv3-512 Conv3-512 Conv3-512	(14,14,512)
Block5_pool	Maxpooling (2D)	(7,7,512)
Sequential	FC-1024	(4)
	BN + Dropout (0.5) +Sigmoid	
	FC-1024	
	BN + Dropout (0.5) +Sigmoid	
Output	FC-4 + Softmax	(4)

^a 'Conv' refers to 'convolutional layer' and is denoted as “conv<receptive field size> - <number of channels>”. 'FC' refers to 'fully connected layer' and is denoted as “FC-<numbers of output>”. BN refers to 'batch normalization'. Dropout rates are in parentheses.

Data Preprocessing and Adjustment for Imbalance Data

The 'Caffe' mode is used as our preprocess method: images were converted from RGB to BGR and then zero-centered each color channel with respect to our own dataset by subtracting the mean of each channel. In validation, a stratified K-fold is performed so that the images of each stage can be uniformly distributed in each fold. Finally, imbalance adjustments were implemented to prevent unequal training due to the unequal distribution of each stage by simply oversampling the lesser input data. No other augmentation is applied to the input data.

Web-based application and external validation

After the training and validation process, we develop a web-based application by implementing the trained model. This application will be released to our cooperating medical facilities to assist with the long-term care recording work. Users may upload an image from their smartphone to our application and receive a predicted stage from the system. If the result is correct, one may press the save button. If not, one may correct the result, and at the same time, the error prediction will be recorded to improve our model in future version.

Evaluation and Performance Metrics

To analyze the result of external validation, the uploaded PI images will be labeled by 3 professional nurses under the same protocol in training and validation dataset. Accuracy, confusion matrix, area under the receiver operating characteristic curve (AUROC), and heatmap[29] are used to evaluate the performance of our model.

Confusion matrix, which consists of ground truth axis and prediction axis, is used to measure the effectiveness of a model. Recall, precision, specificity, and accuracy may be calculated to understand the performance of a model. AUROC represents the probability that a random positive data is ranked as negative data. For a robust model, the difference of positive and negative predictions should be large; therefore, the value of AUROC will be close to 1. Heatmaps explain which regions provide strong responses that make their image fall into the corresponding category. Through this visualization analysis, we discover the relationships between clinical knowledge and responses to our approach.

Results

Dataset and Ground-Truth Classification Result

527 pieces of PI images are collected from 2018 to 2020 from 40 cooperating long-care organizations. These images are then sorted into 4 stages by 3 professional nurses according to the NPIAP definition announced in 2016. The result is shown in Table 2, and the sample images are shown in Figure 1. From Table 2, there are the least number of stage 1 PI images in our dataset. The reason is that the pixel of stage 1 PI images' length rarely exceeds 225 pixels, and thus will be picked out from the collected images. According to NPIAP's definition, stage 1 PI is described as a localized area of nonblanchable erythema on intact skin. However, the early phase of erythema usually appears to be a relatively small area. Therefore, after picking out images that do not meet our standard, stage 1 PI has the fewest images. Stage 3 PI also has a relatively small number of images, while stage 2 PI has the highest number of images, which is more than three times the number of stage 1 PI images. Overall, in our dataset, the number of images in the four stages is imbalance.

Table 2. Dataset details in classification task

Stage ^a	Classification
1	60
2	207
3	94
4	144
Total	505

^aThe staging system refers to the standard of NPIAP set in 2016 [2].



Figure 1. Sample images for the input data of classification task

Training and Validation Result

The results of our training and validation process are plotted as loss curves and accuracy curves, as shown in Figure 2. In the training process, it is shown that our model converges rapidly, reaching a stable accuracy of 95% after 25 epochs. The same trend can be observed in the validation process but with a larger bias. In the accuracy curve, the accuracy of validation ranges from 71.03% to 86.67%. In the loss curve, the loss of validation also oscillates with a larger noise. The reason being that in validation process, we usually report a smaller batch of data with respect to training dataset. Therefore, it is normal for validation process to be noisier. Moreover, we do not observe overfitting in training and validation process since our validation accuracy is smaller than training accuracy and our loss in validation does not increase rapidly in all folds. To obtain the final model, we save our model and its parameters once the model reaches a better accuracy in each fold. The model with a validation accuracy of 86.67% is recorded from the fourth fold and will be used for application development and testing.

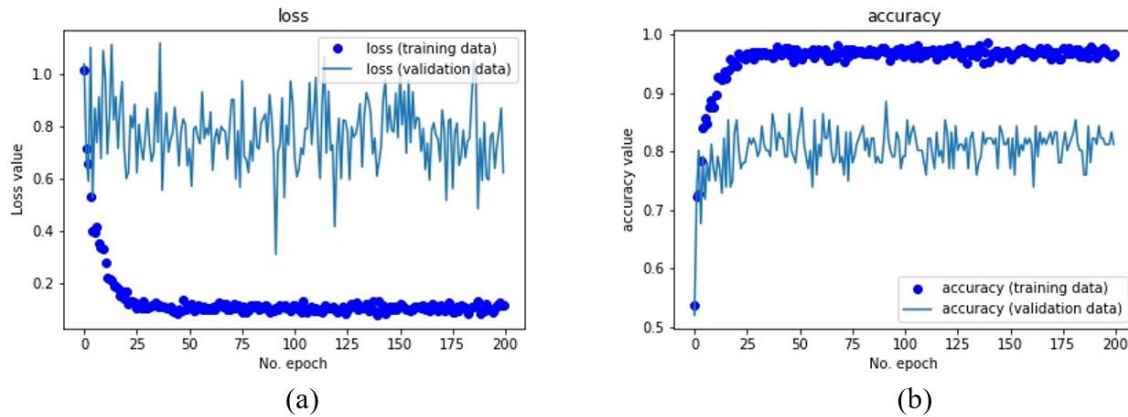


Figure 2. (a) Loss curve and (b) accuracy curve of training and validation process in fold 4. In this fold, the best model is obtained by reaching the highest accuracy of 86.67% in (b).

Web-Based Application and External Validation

After obtaining the model with the best accuracy, we implement this model into a comprehensive web-based application, which helps medical workers to classify PI stages by uploading images. Snapshots of the user interface of this application are shown in Figure 3. Figure 3-(a) is the default interface of our application. In addition to the stage of the PI image, the size, color, and other properties of a PI image are also listed for the user to select. Figure 3-(b) is the interface once the users upload their PI image onto our application. In this figure, the stage of the PI image is predicted and recorded by our application.

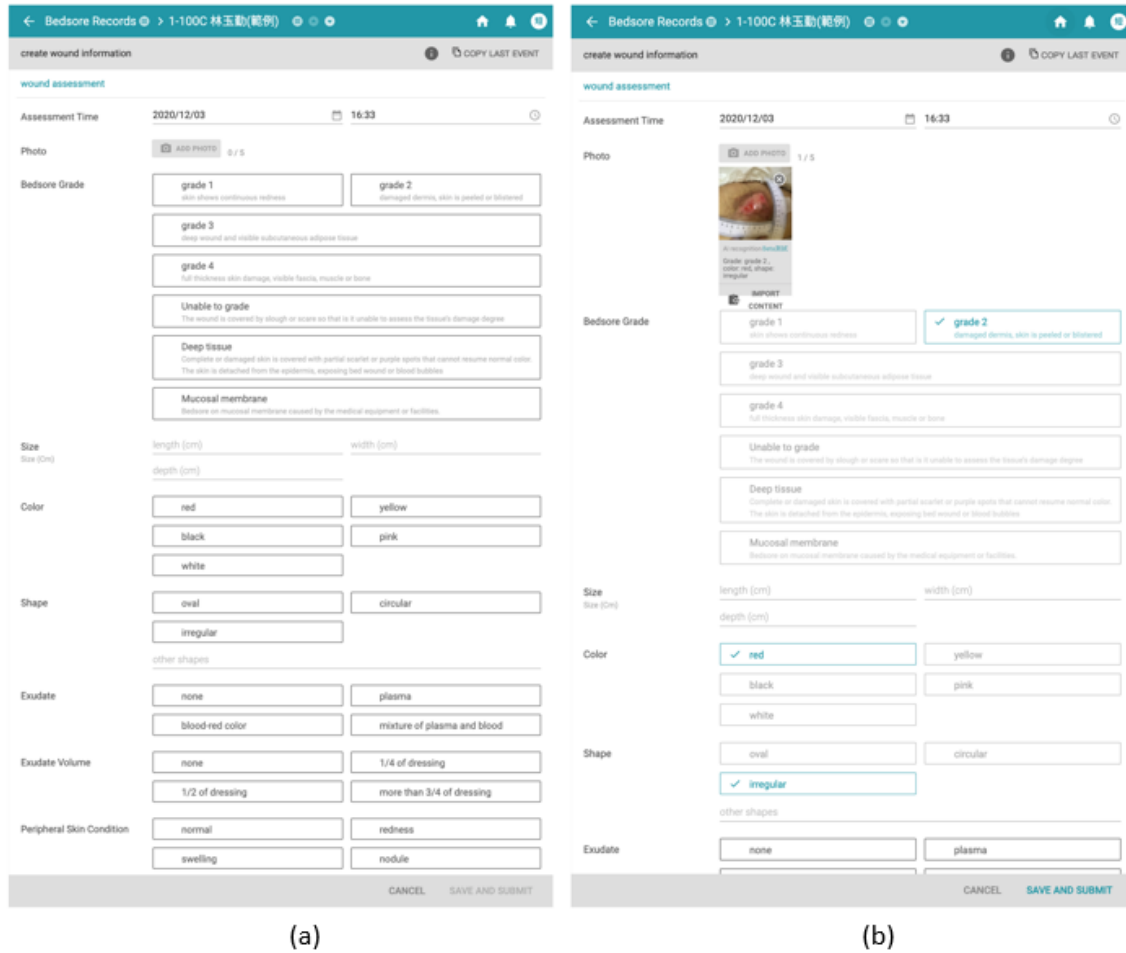


Figure 3. Screenshots of our web-based application. (a) is the default interface of our application, and (b) is the snapshot after the users uploaded their images.

There are two main services in our application. First, the application is used as a rapid classifier for users to obtain the stage of PI images. Second, this application serves as a recorder for medical workers to record the PI status of patients and their treatment. To verify our work, we collaborate with the same long-term care organizations from which we gather training data and perform an external validation.

This external validation was carried out from 10/2020 to 5/2021. Within this time period, caregivers in long-term care organizations used our application to record and predict the stage of PI images. At the same time, ground truth of PI images is labeled by 3 professional nurses under the same standard as our dataset. The number of PI images in each stage used in external validation is listed in Table 3.

Table 3. Dataset details of external validation

Stage ^a	External validation
1	17
2	111
3	53
4	156
Total	337

^aThe staging system is based on the NPIAP standard set in 2016 [2]

External Validation Result

The results of external validation, performed from 10/2020 to 05/2021, are evaluated as confusion matrix and AUROC in Figure 4. According to the confusion matrix, most of the predictions fall in the main diagonal of the matrix, and the accuracy of our result 85.16%. We can see that the stage 3 PI images are the most difficult to distinguish among all stages, with the recall of only 52.83%, indicating that the performance in stage 3 PI images differs little from a random guess. From the result of the prediction, it is observed that stage 3 PI images are easily predicted as stage 2 or stage 4. Another reason is that the number of stage 3 PI images in training dataset is insufficient compared to stage 2 and stage 4 PI images. We believe that by increasing our training dataset, the model will be able to distinguish the PI images of different stages more accurately.

Similar results can be observed from AUROC. Evidently, our model has the lowest ability to distinguish the stage 3 PI images compared to other stages (AUROC = 0.85). In other words, it is more likely to mispredict stage 3 PI images and categorize them as stage 2 or stage 4. As for other stages, all of their AUROC values reach more than 0.91, demonstrating its ability to correctly classify stage 1, 2 and 4 PI images. Among all stages, stage 4 PI has the best measure of separability with its highest AUROC of 0.95.

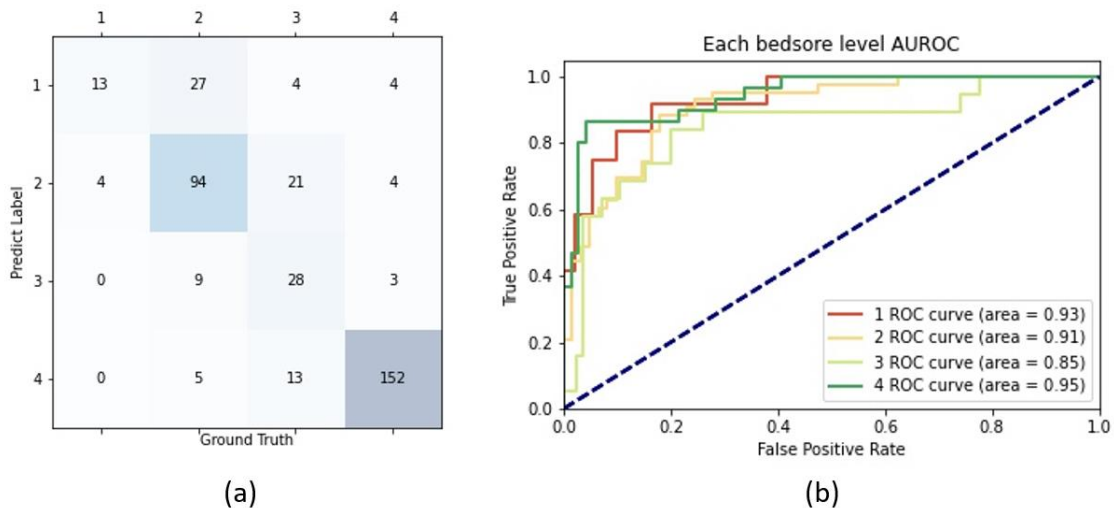


Figure 4. The confusion matrix (a) and AUROC curve (b) using transfer learning model. AUC: area under the curve; ROC: receiver operating characteristic.

Heatmap Analysis

For further analysis, PI input images and their heatmaps are plotted and compared in Figure 5. Although we do not explicitly label the attributes of each type of PI images when training our model, it is shown that our classifying system highlights the representative hot spots. For example, muscle tissue and fat tissue, which are the characteristics of stage 3 and 4 PI images, are generally considered highly relevant and labeled as red areas in the corresponding heatmaps. For the less important regions, the mask is colored blue.

In stage 1, some regions are labeled red. In stages 2, 3, and 4, attention is clearly drawn to fat tissue, necrotic tissue or tissue with depth and deeper colors. It is also observed in the heatmaps of stage 3 and stage 4, some surrounding regions are misinterpreted as important areas. These areas usually have deeper colors or are located near the boundaries of clothing and skin. This implies that our model is sensitive to such features in an image and should be avoided when entering data. Overall, through heatmap analysis, we discover close relationships between clinical knowledge and the perspective of our classifier.

By combining the confusion matrix and heatmap, we may explain some of the incorrect PI testing images. First, images that were predicted into higher stages usually have a deeper color (e.g. black) around the injury, possibly caused by shadows, medicine treatment, or clothing. For example, a stage 2 PI image in Figure 6 is predicted as stage 4 due to the black area on top of the wound. In contrast, images that were misclassified into lower stages tend to have lower resolution or poor lighting. These images may be the result of photos taken from difficult positions by caregivers (Figure 6).





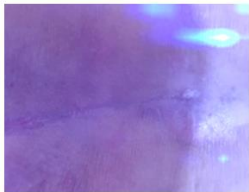
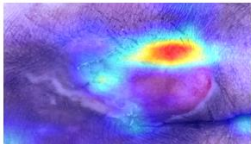

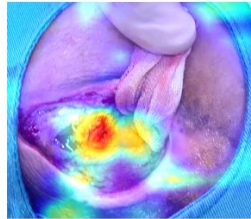
	Stage 1	Stage 2	Stage 3	Stage 4
Input image				
Heatmap				

Figure 5. Images and their heatmaps.



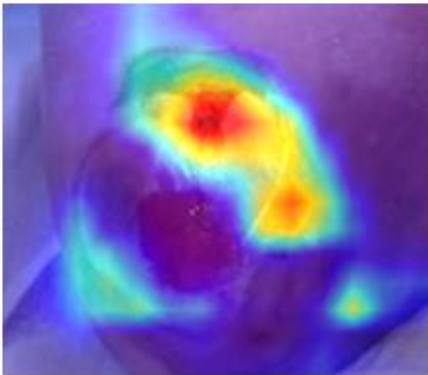
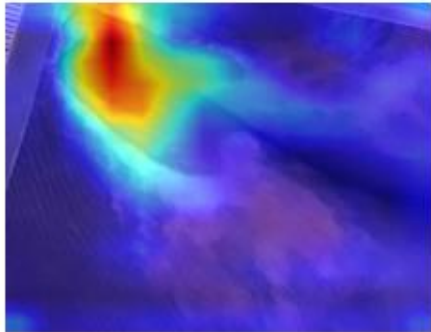
	Stage 2 predict into Stage 3	Stage 4 predict into Stage 2
Input image		
Heatmap		

Figure 6. Mispredicted Images and their heatmaps

Discussion

Principal Results

In this study, we develop a web-based application that implements deep learning models to perform the classification of pressure injury. The loss curve and accuracy curve show a stable result in training process, implying that our input data may be able to obtain a robust classification model. The highest accuracy in 5-fold stratified cross-validation is reported as 86.67%. In external validation, the accuracy is calculated as 85.16%. Confusion matrix, AUROC, and heatmaps are evaluated to analysis the performance of our model. The AUROC of the four stages is calculated to be 0.93, 0.91, 0.85, and 0.95. Highlighted areas in heatmaps prove that the perspective of our model is in close relationship with clinical knowledge. These results confirm that the model should be able to assist the caregivers in long-term care organizations to perform recording work.

Limitations

This dataset was provided by Taiwanese patients; therefore, it is possible that the model may not reach the same accuracy with the same parameters or setup for patients of other races. In addition, images with insufficient lighting, strange angles, or interfering objects are shown to be distracting to our classification model. Therefore, it is critical for users to follow the restrictions of our model.

Conclusions

This study demonstrates that the DL approach could be used as an efficient PI stage classifier whose classification standard is based on the NPIAP standard set in 2016 [2]. Not only does this application predict the stage of PI, it also serves as a recording application for caregivers to track their treatment or PI development of different patients.

Acknowledgements

Please include all authors' contributions, funding information, financial disclosures, role of sponsors, and other acknowledgements here. This description should include participation, if any, in the review and approval of the manuscript for publication and the role of the sponsors. Omit if not applicable.

Abbreviations

AUROC: area under the receiver operating characteristic curve

CNN: Convolutional neural network

DL: Deep learning

ML: machine learning

NPIAP: National Pressure Injury Advisory Panel

PI: Pressure injury

ROC: receiver operating characteristic curve

SVM: support vector machine

References

1. Lyder, C.H., *Pressure Ulcer Prevention and Management*. JAMA, 2003. **289**(2): p. 223-226.
2. Edsberg, L.E., et al., *Revised National Pressure Ulcer Advisory Panel Pressure Injury Staging System: Revised Pressure Injury Staging System*. J Wound Ostomy Continence Nurs, 2016. **43**(6): p. 585-597.
3. Nixon, J., et al., *Reliability of pressure ulcer classification and diagnosis*. Journal of Advanced Nursing, 2005. **50**(6): p. 613-623.
4. Yarkony, G.M., et al., *Classification of Pressure Ulcers*. Archives of Dermatology, 1990. **126**(9): p. 1218-1219.
5. Shalev-Shwartz, S. and S. Ben-David, *Understanding machine learning: From theory to algorithms*. 2014: Cambridge university press.
6. Stephen, I., *Perceptron-based learning algorithms*. IEEE Transactions on neural networks, 1990. **50**(2): p. 179.
7. Chang, C.-C. and C.-J. Lin, *LIBSVM: A library for support vector machines*. ACM transactions on intelligent systems and technology (TIST), 2011. **2**(3): p. 1-27.
8. Goodfellow, I., et al., *Deep learning*. Vol. 1. 2016: MIT press Cambridge.
9. Rawat, W. and Z. Wang, *Deep convolutional neural networks for image classification: A comprehensive review*. Neural computation, 2017. **29**(9): p. 2352-2449.

10. Pal, N.R. and S.K. Pal, *A review on image segmentation techniques*. Pattern recognition, 1993. **26**(9): p. 1277-1294.
11. Cambria, E. and B. White, *Jumping NLP curves: A review of natural language processing research*. IEEE Computational intelligence magazine, 2014. **9**(2): p. 48-57.
12. Yurtsever, E., et al., *A survey of autonomous driving: Common practices and emerging technologies*. IEEE Access, 2020. **8**: p. 58443-58469.
13. Goodfellow, I., et al., *Generative adversarial nets*. Advances in neural information processing systems, 2014. **27**: p. 2672-2680.
14. Wang, X., et al. *Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.
15. Abbas, A., M.M. Abdelsamea, and M.M. Gaber, *Classification of COVID-19 in chest X-ray images using DeTraC deep convolutional neural network*. arXiv preprint arXiv:2003.13815, 2020.
16. Rajpurkar, P., et al., *CheXNet: radiologist-level pneumonia detection on chest x-rays with deep learning*. 2017. arXiv preprint arXiv:1711.05225, 2020.
17. Veredas, F., H. Mesa, and L. Morente, *Binary tissue classification on wound images with neural networks and bayesian classifiers*. IEEE transactions on medical imaging, 2009. **29**(2): p. 410-427.
18. Veredas, F.J., et al., *Wound image evaluation with machine learning*. Neurocomputing, 2015. **164**: p. 112-122.
19. Noguchi, H., et al. *Clustering and classification of local image of wound blotting for assessment of pressure ulcer*. in *2014 World Automation Congress (WAC)*. 2014. IEEE.
20. Zahia, S., et al., *Tissue classification and segmentation of pressure injuries using convolutional neural networks*. Computer methods and programs in biomedicine, 2018. **159**: p. 51-58.
21. Kosmopoulos, D.I. and F.L. Tzevelekou, *Automated pressure ulcer lesion diagnosis for telemedicine systems*. IEEE Engineering in Medicine and Biology Magazine, 2007. **26**(5): p. 18-22.
22. Shenoy, V.N., et al. *Deepwound: automated postoperative wound assessment and surgical site surveillance through convolutional neural networks*. in *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 2018. IEEE.
23. Jiang, M., et al., *Using Machine Learning Technologies in Pressure Injury Management: Systematic Review*. JMIR medical informatics, 2021. **9**(3): p. e25704.
24. Koepp, J., et al., *The Quality of Mobile Apps Used for the Identification of Pressure Ulcers in Adults: Systematic Survey and Review of Apps in App Stores*. JMIR mHealth and uHealth, 2020. **8**(6): p. e14266.
25. Simonyan, K. and A. Zisserman, *Very deep convolutional networks for large-scale image recognition*. arXiv preprint arXiv:1409.1556, 2014.
26. Pan, S.J. and Q. Yang, *A survey on transfer learning*. IEEE Transactions on knowledge and data engineering, 2009. **22**(10): p. 1345-1359.

27. Raschka, S., *Model evaluation, model selection, and algorithm selection in machine learning*. arXiv preprint arXiv:1811.12808, 2018.
28. Kingma, D.P. and J. Ba, *Adam: A method for stochastic optimization*. arXiv preprint arXiv:1412.6980, 2014.
29. Selvaraju, R.R., et al. *Grad-cam: Visual explanations from deep networks via gradient-based localization*. in *Proceedings of the IEEE international conference on computer vision*. 2017.