Original Paper

Fine-Grained Classification of Pressure Ulcers and Incontinence-Associated Dermatitis Using Multimodal Deep Learning: Algorithm Development and Validation Study

Alexander Brehmer[1], MSc; Constantin Seibold[1], PhD; Jan Egger[1,2,3], Prof Dr; Khalid Majjouti[4], MSc; Michaela Tapp-Herrenbrück[4], BSc; Hannah Pinnekamp[5], MSc; Vanessa Priester[5], MA; Michael Aleithe[6], PhD; Uli Fischer[5], Prof Dr; Bernadette Hosters[4], MSc; Jens Kleesiek[1,3], Prof Dr, Prof Dr med

[1]Institute for Artificial Intelligence in Medicine, Essen University Hospital, Essen, Germany
[2]Center for Virtual and Extended Reality in Medicine, University Medicine Essen, Essen, Germany
[3]Faculty of Computer Science, University of Duisburg-Essen, Essen, Germany
[4]Department of Nursing Development and Nursing Research, University Hospital Essen, Essen, Germany
[5]Department of Clinical Nursing Research and Quality Management, Hospital of the Ludwig Maximilian University, Munich, Germany
[6]Sciendis GmbH, Leipzig, Germany

Corresponding Author:

Alexander Brehmer, MSc
Institute for Artificial Intelligence in Medicine
Essen University Hospital
Girardetstr. 2
Essen, 45131
Germany
Phone: 0201 72377829
Email: alexander.brehmer@uk-essen.de

Abstract

Background: Pressure ulcers (PUs) and incontinence-associated dermatitis (IAD) are prevalent conditions in clinical settings, posing significant challenges due to their similar presentations but differing treatment needs. Accurate differentiation between PUs and IAD is essential for appropriate patient care, yet it remains a burden for nursing staff and wound care experts.

Objective: This study aims to develop and introduce a robust multimodal deep learning framework for the classification of PUs and IAD, along with the fine-grained categorization of their respective wound severities, to enhance diagnostic accuracy and support clinical decision-making.

Methods: We collected and annotated a dataset of 1555 wound images, achieving consensus among 4 wound experts. Our framework integrates wound images with categorical patient data to improve classification performance. We evaluated 4 models—2 convolutional neural networks and 2 transformer-based architectures—each with approximately 25 million parameters. Various data preprocessing strategies, augmentation techniques, training methods (including multimodal data integration, synthetic data generation, and sampling), and postprocessing approaches (including ensembling and test-time augmentation) were systematically tested to optimize model performance.

Results: The transformer-based TinyViT model achieved the highest performance in binary classification of PU and IAD, with an F1-score (harmonic mean of precision and recall) of 93.23%, outperforming wound care experts and nursing staff on the test dataset. In fine-grained classification of wound categories, the TinyViT model also performed best for PU categories with an F1-score of 75.43%, while ConvNeXtV2 showed superior performance in IAD category classification with an F1-score of 53.20%. Incorporating multimodal data improved performance in binary classification but had less impact on fine-grained categorization. Augmentation strategies and training techniques significantly influenced model performance, with ensembling enhancing accuracy across all tasks.

Conclusions: Our multimodal deep learning framework effectively differentiates between PUs and IAD, achieving high accuracy and outperforming human wound care experts. By integrating wound images with categorical patient data, the model enhances diagnostic precision, offering a valuable decision-support tool for health care professionals. This advancement has the potential to reduce diagnostic uncertainty, optimize treatment pathways, and alleviate the burden on medical staff, leading to faster interventions and improved patient outcomes. The framework's strong performance suggests practical applications in clinical settings, such as integration into hospital electronic health record systems or mobile applications for bedside diagnostics. Future work should focus on validating real-world implementation, expanding dataset diversity, and refining fine-grained classification capabilities to further enhance clinical utility.

JMIR AI 2025;4:e67356; doi: 10.2196/67356

Keywords: computer vision; image classification; wound classification; deep learning; pressure ulcer; incontinence-associated dermatitis; multi modal data; synthetic image generation

Introduction

Background

Pressure ulcers (PUs) and incontinence-associated dermatitis (IAD) are significant challenges in clinical settings due to their prevalence and impact on patient health and well-being. The global prevalence of PUs is estimated to be 12.8% [1], while studies have estimated the IAD prevalence to be between 5.6% and 50% [2]. These wounds not only cause physical discomfort but also pose risks of infection and prolonged hospital stays, increasing health care costs and diminishing the quality of life for affected individuals.

Accurately distinguishing between PUs and IAD poses a considerable challenge for health care providers and wound care experts. Both conditions share similar presentations, yet their underlying causes and optimal treatment approaches differ vastly. This ambiguity not only complicates diagnosis but also delays appropriate interventions, potentially exacerbating patient discomfort and prolonging healing times [3].

To address this challenge, the KIADEKU project [4] was initiated to develop an innovative artificial intelligence (AI) system capable of distinguishing between PUs and IAD using wound image data and key patient information.

Goal of This Study

The goal of this study is to advance wound care by developing a robust multimodal deep learning framework for the fine-grained classification of PUs and IAD. By integrating wound images with categorical patient data, we aim to enhance diagnostic accuracy in distinguishing between these conditions and in categorizing their respective wound severities. We conduct extensive benchmarking of state-of-the-art convolutional and transformer-based models, emphasizing optimal performance while ensuring computational efficiency for practical deployment in clinical settings. The optimized model addresses the challenging task of accurately classifying PU and IAD wounds, providing valuable insights and tools to support clinical decision-making and guide future research in wound classification.

Related Work

Deep learning has significantly advanced wound classification, including PUs and other wound types. Various studies have explored different deep learning architectures and techniques to improve diagnostic accuracy and efficiency. Table 1 summarizes key contributions in this domain.

While previous studies have demonstrated the effectiveness of deep learning for wound classification, they predominantly rely on image data alone. However, accurate wound diagnosis often depends on both visual appearance and key clinical factors, such as wound location, patient mobility, and incontinence severity. To our knowledge, no existing study rigorously integrates multimodal data fusion, combining wound images with categorical patient information. Our approach leverages this additional patient context, allowing the model to capture clinically relevant patterns that purely image-based models may overlook, thereby significantly improving diagnostic precision and decision support. Furthermore, our approach involves extensive benchmarking of state-of-the-art convolutional and transformer-based models, as well as various training techniques, augmentations, and postprocessing methods to enhance performance. This comprehensive evaluation sets our method apart in both scope and effectiveness, contributing to a novel multimodal framework for fine-grained wound classification that can support clinical decision-making and guide future research in this domain.

Table 1. Summary of related work in wound classification and pressure ulcer classification.

Authors	Method	Key contributions
Pressure ulcer classification		
Aldughayfiq et al [5]	YOLOv5-based classification	Classified pressure ulcers into 4 stages and non-pressure ulcer categories with real-time detection capabilities.
Chang et al [6]	Superpixel segmentation	Used superpixel techniques for automatic pressure ulcer diagnosis, enhancing wound segmentation and classification accuracy.
Seo et al [7]	CNN[a]-based classification	Developed a deep learning model to visually classify pressure injury stages, aiding nurses in diagnostic accuracy.
García-Zapirain et al [8]	3D CNNs	Explored 3D CNNs for classifying pressure ulcer tissues, capturing spatial features for precise tissue type classification.

Authors	Method	Key contributions
Liu et al [9]	CNN-based assessment system	Introduced a system to aid in pressure ulcer diagnosis and clinical decision-making, enhancing speed and accuracy.
Lau et al [10]	AI[b]-enabled smartphone app	Developed an app for real-time pressure injury assessment using advanced AI algorithms.
Kim et al [11]	Deep learning model for staging	Assessed a deep-learning model's clinical utility for pressure injury staging, enhancing decision-making in wound care.
Swerdlow et al [12]	Mask R-CNN	Proposed simultaneous segmentation and classification of pressure injury images, improving diagnostic efficiency.
Zahia et al [13]	CNNs for classification	Focused on classification and segmentation of pressure injury tissues, identifying different tissue types accurately.
Pandey et al [14]	Thermal imaging classification	Developed and validated a deep learning-based thermal imaging framework to automatically stage pressure ulcers
Wound classification		
Huang et al [15]	CNN-based tool	Developed a tool for automatic classification of various wound types, supporting accurate diagnoses.
Oura et al [16]	Deep learning in forensic analysis	Applied deep learning for gunshot wound interpretation, demonstrating versatility in wound classification contexts.
Rostami et al [17]	Ensemble CNN classifier	Explored multiclass wound image classification using ensemble methods to enhance accuracy.
Patel et al [18]	Integrated image and location analysis	Incorporated visual and locational data for wound classification, highlighting multimodal data integration's importance.
Liu et al [19]	EfficientNet models	Applied EfficientNet to classify diabetic foot ulcer ischemia and infection, handling complex wound classification tasks.
Lee et al [20]	Ultrasound imaging with deep learning	Developed a model for burn depth classification using ultrasound images for non-invasive assessment.
Afza et al [21]	Hybrid deep features selection	Investigated skin lesion classification using deep features and extreme learning machines, enhancing medical image analysis.
Cheng et al [22]	ConvNext Tiny, Gun Shot Classification	Pioneers the application of deep learning in forensic pathology by demonstrating that AI can reliably differentiate between entrance and exit gunshot wounds using digital color images.
Odame et al [23]	CLAHE-enhanced images, DWT, FixCaps	Developed a multi-wound classification framework that integrates image enhancement (using CLAHE and DWT) with deep learning

[a] CNN: convolutional neural network.
[b] AI: artificial intelligence.

Methods

Dataset

In this study, we use a new wound dataset collected and annotated over a 2-year period as part of the KIADEKU project. The data originate from the project partners Ludwig Maximilian University University Hospital and Essen University Hospital and were annotated by 4 wound experts with extensive clinical experience in wound management using the Label Studio Software [24]. Considering the difficulty of the task, we enforced a strong ground truth by having all images annotated by 3 wound experts and only used images where 2 wound experts reached consensus in their annotations.

The annotators categorized each image as either IAD, PU, invalid, or borderline case (both wounds present) and assessed the categorization of each wound type. For PU classification, we followed the *International Classification of Diseases*-10 standard [25], which defines 4 degrees (1-4) of PU wounds. Similarly, for IAD classification, we used the Ghent Global IAD Categorization Tool (GLOBIAD) [26], which categorizes IAD wounds into 4 distinct categories: 1A, 1B, 2A, and 2B. Figure 1 shows an exemplary annotation interface.

Employing the described annotation protocol, a dataset of 1555 images was annotated, from which 1514 images received consensus validation among the annotators. Analysis of the data revealed a generally balanced distribution between the 2 principal wound types under study, PUs and IAD, as depicted in Figure 2. The dataset comprised 763 images of PU and 339 images of IAD.

Of the 763 images categorized as PUs, consensus was achieved for 742 images regarding their specific PU category. The distribution of these categories, as illustrated in Figure 2, reveals a significant class imbalance. Notably, categories 1 and 4 are markedly underrepresented, containing only 25 and 27 images, respectively, compared to 187 images in category 2 and 503 in category 3. This pronounced disparity in class sizes is a critical factor that must be considered when interpreting the training results.

Of the 339 images initially classified as IAD, a consensus on the specific IAD category was reached for 327 images. The class distribution within these categories, as depicted in Figure 2, is relatively balanced compared to the distribution observed in PU categories. Category 2B is the most represented, with 120 images, followed by category 2A with 105 images, 1B with 57 images, and 1A with 45 images. Although this distribution is less skewed than that observed in the PU categories, the smaller sample size overall remains a significant consideration for model training and validation.

Figure 1. Dataset composition with distribution of wound types and categories. PU: pressure ulcer; IAD: incontinence-associated dermatitis.

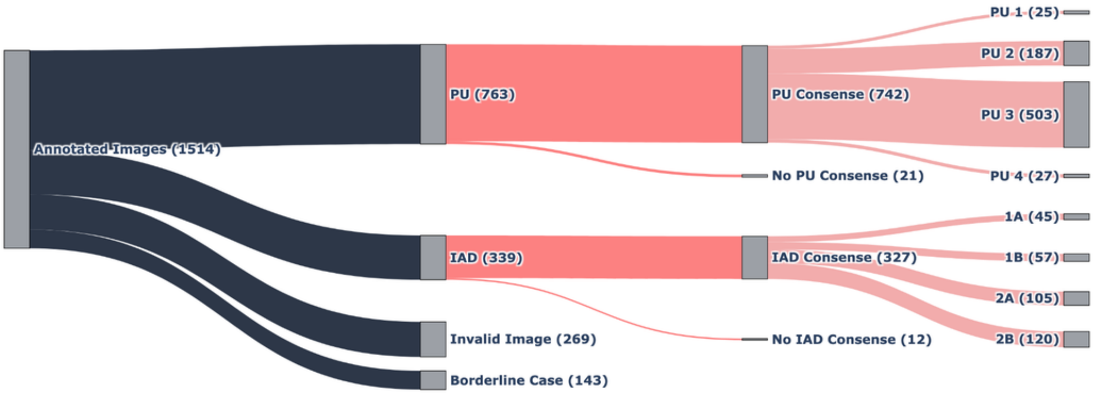
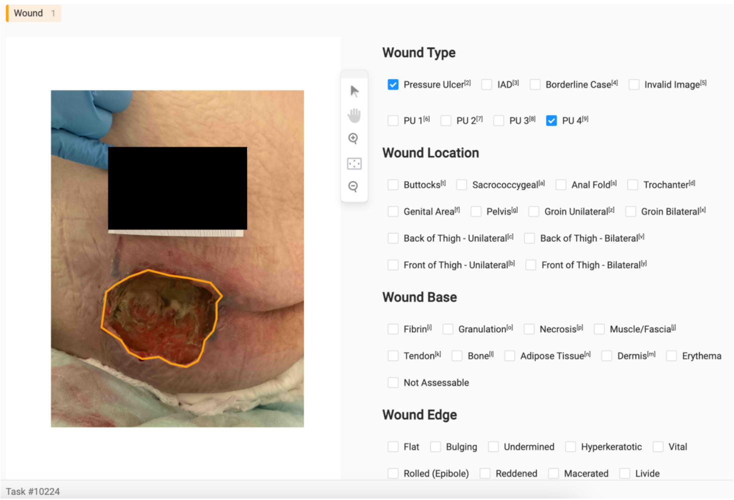


Figure 2. Exemplary annotation process in Label Studio. PU: pressure ulcer; IAD: incontinence-associated dermatitis.



[redacted]

Wound Type

☑ Pressure Ulcer[2] □ IAD[3] □ Borderline Case[4] □ Invalid Image[5]

□ PU 1[6] □ PU 2[7] □ PU 3[8] ☑ PU 4[9]

Wound Location

□ Buttocks[t] □ Sacrococcygeal[a] □ Anal Fold[s] □ Trochanter[d]

□ Genital Area[f] □ Pelvis[g] □ Groin Unilateral[z] □ Groin Bilateral[x]

□ Back of Thigh - Unilateral[c] □ Back of Thigh - Bilateral[v]

□ Front of Thigh - Unilateral[b] □ Front of Thigh - Bilateral[y]

Wound Base

□ Fibrin[i] □ Granulation[o] □ Necrosis[p] □ Muscle/Fascia[j]

□ Tendon[k] □ Bone[l] □ Adipose Tissue[n] □ Dermis[m] □ Erythema

□ Not Assessable

Wound Edge

□ Flat □ Bulging □ Undermined □ Hyperkeratotic □ Vital

□ Rolled (Epibole) □ Reddened □ Macerated □ Livide

Task #10224

Methodology

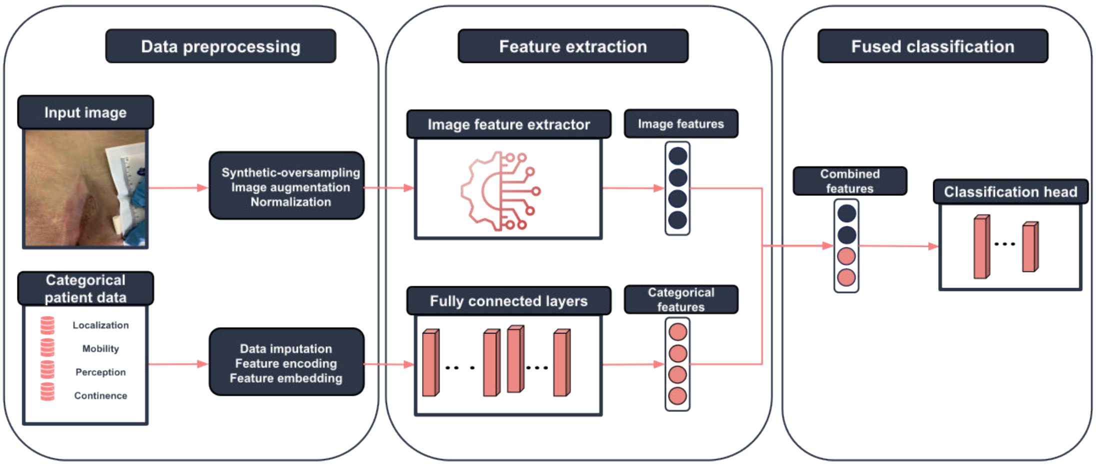
Our proposed classification framework is specifically designed to handle and classify both images and categorical data effectively, as shown in Figure 3.

Initially, the original images undergo several preprocessing steps. These steps include image augmentations and normalization to standardize the input data, alongside the generation of synthetic data points by fine-tuning a stable diffusion model and using these synthetic samples to oversample the minority classes across tasks. We then compare the performance of this approach with traditional oversampling techniques that rely on original data points. For categorical patient data (eg, wound location, mobility, perception ability, and continence status), missing values were addressed using mode imputation, where the most frequent value for each feature was assigned. In cases where missing values exceeded 20% of the dataset for a particular feature, the affected samples were excluded to prevent bias. Additionally, images with conflicting expert annotations (ie, cases where consensus was not reached) were removed to maintain ground truth integrity. After data preprocessing, we extract features from each modality. Image features are extracted using various feature extractors from the Timm [27] library, renowned for their robustness and efficiency; in parallel, categorical features are derived using a simple feed-forward neural network designed to capture the essential characteristics of the embedded categorical data.

In the final stage of our framework, we employ 3 distinct modality fusion techniques to integrate image and categorical features before classification. In the concatenation-based fusion, features from both modalities are directly concatenated to form a comprehensive feature set, which is then passed to a classification head. In the cross-attention-based fusion, categorical features are projected into the image feature space, and a multihead attention mechanism is applied to capture their interactions. In the gated fusion, a gating mechanism adaptively balances the contributions of both modalities, allowing the model to learn the optimal weighting before classification. Each approach ensures effective multimodal integration while leveraging different fusion strategies. The combined feature set is then fed into a final classification head, which is tasked with making the final prediction based on the integrated data.

This setup facilitates a systematic examination and evaluation of various data preprocessing strategies, training techniques, and postprocessing approaches, both independently and in combination. This rigorous methodology allows for a comprehensive comparison and assessment of their efficacy in various combinations across our designated tasks.

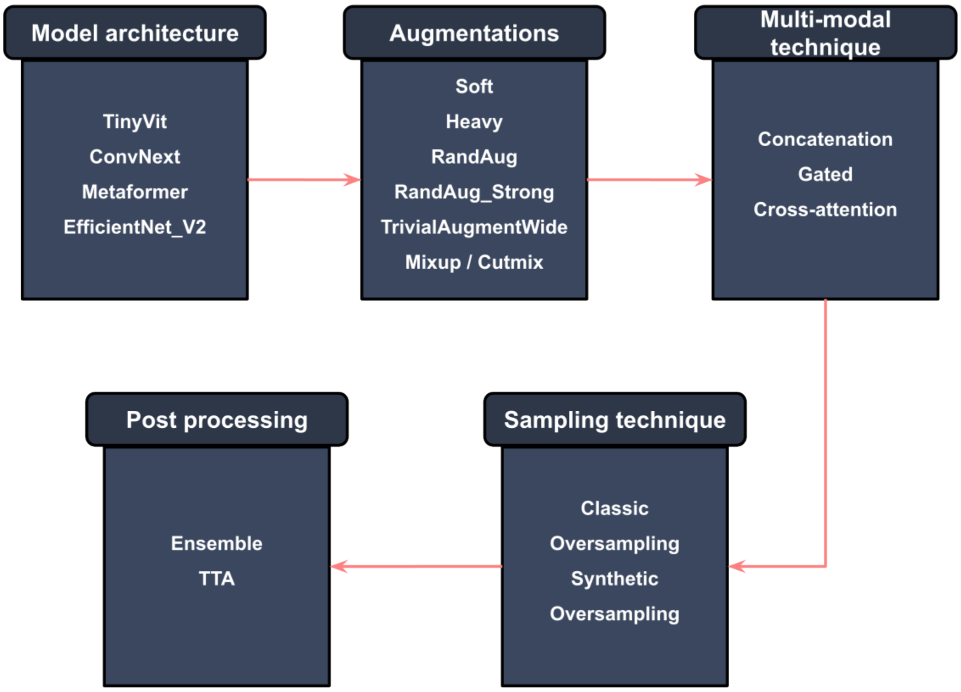
Figure 3. Multimodal architecture visualization illustrating the use of data preprocessing and feature extraction, before fusing the features for the final classification.



Experimental Setup

To maintain a manageable number of experiments, we did not evaluate every possible combination. Instead, we benchmarked various components individually and sequentially integrated the optimal variations for subsequent tests. Specifically, we first identify the model architecture that achieves the highest average rank across our metrics and use this model as the basis for testing different augmentation techniques. The best performing augmentation variation, as determined by the average metric rank, is then used to assess different training techniques. Finally, the best combination of model, augmentation, and training technique is used to benchmark the most effective postprocessing strategy. For a visual representation of this benchmarking flow, refer to Figure 4.

Figure 4. Visualization of experiment setup and benchmark flow. TTA: test-time augmentation.



Training

The general training procedure involves setting the learning rate to 0.0001 and resizing the input images to 384x384 pixels. We use a batch size of 64, with the AdamW optimizer to manage weight decay, and the CrossEntropyLoss loss function for training. The learning rate is adjusted using a CosineAnnealingWarmRestarts scheduler, starting with a cycle length of 10 epochs and a minimum learning rate of 1e-6. Training is performed on an NVIDIA A100 graphics processing unit, with early stopping enabled and a patience of 15 epochs to prevent overfitting. The dataset is initially split into an 80:20 ratio for training and testing. The training set is further divided into 5 equal parts (folds) for cross-validation to enable robust model evaluation.

Models

To evaluate and identify the best possible model for the binary classification task of IAD and PU, as well as the fine-grained wound category classification within these wound types, we selected 4 models with approximately 25 million parameters to ensure a fair comparison and fast inference speed. Using transfer learning, we employed pretrained models from the Timm library, which were originally trained on ImageNet [27]. Our selection includes 2 convolution-based models and 2 transformer-based models, chosen for their exceptional performance relative to their parameter count, as evidenced by Timm's test results on the ImageNet benchmark. For the convolution-based models, we selected a pretrained ConvNeXtV2 model [28,29] and a pretrained EfficientNetV2 model [30,31]. These models are chosen for their state-of-the-art performance and efficiency, making them highly suitable for a wide range of computer vision tasks. The ConvNeXtV2 incorporates advanced architectural enhancements, while EfficientNetV2 uses a novel scaling approach for optimal accuracy and computational efficiency. For the transformer models, we included the MetaFormer [32] and TinyViT [33,34]. The MetaFormer is selected for its innovative design that enhances transformer capabilities, while TinyViT, a distilled vision transformer, is designed to retain high accuracy with fewer parameters and computational resources, making it suitable for resource-constrained environments.

Augmentations

We evaluate 6 distinct augmentation techniques comprising 3 randomized methods and 2 custom-designed variants and the use of CutMix/MixUp [35]. Initially, we test the RandAugment [36] method using its PyTorch implementation with default settings. To explore more robust options, we employ an intensified version of RandAugment, increasing the augmentation count to 4 and the magnitude to 12. Additionally, we assess the PyTorch implementation of TrivialAugmentWide [37] with default parameters, a straightforward approach that applies a single, random augmentation to each image. Moreover, we introduce 2 proprietary augmentations developed for exploratory purposes. The first, a mild augmentation set, incorporates random affine transformations, perspective adjustments, and rotations. The second, a more intensive augmentation suite, applies random flips, rotations, color jittering, affine transformations, perspective adjustments, and Gaussian blurring, all implemented using the torchvision transformation library. Finally, we evaluate CutMix and MixUp augmentation using the PyTorch v2 implementation, where images are randomly augmented with either CutMix or MixUp using a random selection strategy, ensuring diverse augmentation during training.

Training Techniques and Postprocessing

Next, we explore various training variations and postprocessing techniques used in this study. Initially, we incorporate multimodal data in our training, which includes both patient images and tabular data detailing wound location, mobility, perception ability, and urinary plus fecal continence. Each factor of mobility, perception, and continence is quantified on a scale from 0 to 4. A joint fusion approach is adopted for multimodal classification, where image embeddings are combined with tabular data embeddings, followed by a final classification head.

Concerning sampling strategies, we address the low sample size in certain classes by employing oversampling techniques to balance class distributions. In addition to classic oversampling, we introduce a synthetic data generation approach by fine-tuning a stable diffusion model to generate artificial images for the minority classes. This allows us to augment underrepresented categories with high-quality synthetic samples. We compare the performance of this approach against traditional oversampling methods to assess its effectiveness in mitigating class imbalance. In terms of postprocessing, we implement ensembling to enhance model performance and robustness by averaging predictions from all 5 folds. Furthermore, test time augmentation is employed by averaging predictions of the original image with 4 additional variants that have undergone mild augmentations such as random flips, rotations, and slight color jitter.

Evaluation Metrics

To assess the performance of the various models and training strategies, we employ several key metrics. The evaluation metrics used in this study include F1 score, area under the receiver operating characteristic curve (AUROC), and average precision (AP). All metrics were implemented using the torchmetrics library [38]. These metrics were chosen based on informed estimations and insights from Maier-Hein et al [39] recommendations.

Ethical Considerations

Ethical approval for this study was granted by the Ethics Committee of the Medical Faculty of the University of Duisburg-Essen on October 4, 2022 (ref number: 22-10905-BO). The study involved retrospective analysis of de-identified image data, and no direct contact with participants occurred. As such, informed consent was not required. All data were processed in compliance with applicable privacy and data protection regulations. In addition, the overall KIADEKU project is registered with the German Clinical Trials Register (Deutsches Register Klinischer Studien (DRKS)) under the registration number DRKS00029961.

Results

Overview

Table 2 presents the performance metrics of our best models across the 3 classification tasks. For the binary classification between PU and IAD, the model achieved an F1-score of 93.23%, an AUROC of 0.9852, and an AP of 0.9813. In the PU category classification, the model obtained an F1-score of 75.43%, an AUROC of 0.9384, and an AP of 0.8616. For the IAD category classification, the F1-score was 53.20%, with an AUROC of 0.8391 and an AP of 0.5927.

When examining the optimal combinations per task (refer to Table 3), it is observed that, from an architectural standpoint, transformer models exhibit a superior performance compared to convolution-based models. An exception to this trend is noted in the IAD Category Classification task, where the ConvNeXtV2 model achieves the highest overall performance.

Table 2. Performance of the best models.

Technique	F_1-score	AUROC[a]	AP[b]
Binary	0.9323	0.9852	0.9813
PU[c] category	0.7543	0.9384	0.8616
IAD[d] category	0.5320	0.8391	0.5927

[a] AUROC: area under the receiver operating characteristic curve.
[b] AP: average precision.
[c] PU: pressure ulcer.
[d] IAD: incontinence-associated dermatitis.

Table 3. Best benchmark result overview.

Task	Model	Augmentation	Multimodal technique	Sampling technique	Post processing
Binary	TinyViT	TrivialAugmentWide	Cross-attention	None	Ensemble
PU[a] category	TinyViT	RandAug Strong	None	Synthetic oversampling	Ensemble
IAD[b] category	ConvNeXtV2	Heavy	None	Synthetic oversampling	Ensemble

[a] PU: pressure ulcer.
[b] IAD: incontinence-associated dermatitis.

Regarding augmentations, lighter augmentations enhance performance in the binary classification task. Conversely, the finer category classification tasks benefit from more intensive augmentations, including a heavy augmentation set and significant variations of RandAugment .

Training techniques also show variability across tasks. Multimodality training proves advantageous for the binary classification, whereas it detracts from performance in fine-grained category classification. The cross-attention-based modality fusion approach shows the best performance for the binary classification task. Tailored sampling strategies yield the most substantial performance enhancements, particularly for the PU and IAD category classification tasks, where significant class imbalances are present. Both classic and synthetic oversampling improve performance in these tasks, with the synthetic approach achieving superior results. However, for the binary classification task, neither method provides a noticeable performance increase compared to the standard training regimen.

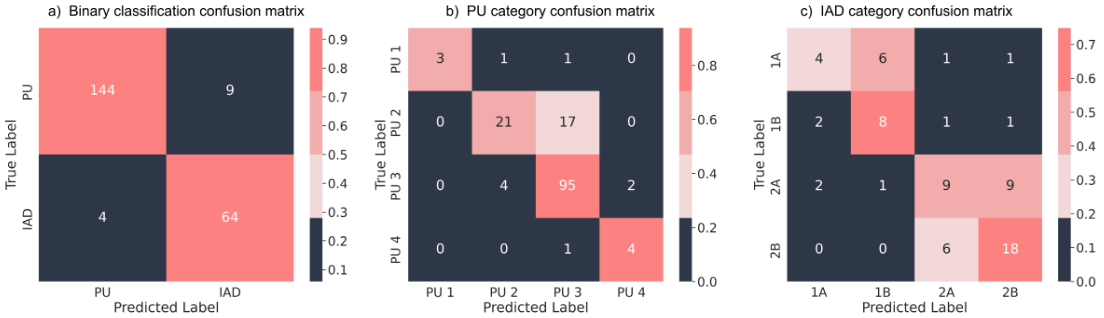
In the realm of postprocessing techniques, there is a discernible preference for ensembling, which enhances performance across all evaluated tasks. While test-time augmentation also positively impacts performance most of the time, its effectiveness is not as pronounced as that achieved through ensembling.

Detailed performance metrics for the tasks are provided in the multimedia appendices (Multimedia Appendices 1–3).

In examining the outcomes of the confusion matrices for the optimal combinations per task, as depicted in Figure 5, a more nuanced understanding of the results and the inherent complexities of the tasks is achieved. The binary classification task demonstrates a high degree of accuracy, achieving low rates of false positives and false negatives, despite the presence of slight class imbalance between the 2 categories. The classification of PU categories presents notable challenges, particularly for categories 1 and 2, which are characterized by their low frequency within the dataset. A mixup between PU-2 and PU-3 is observed to be the most common misclassification, indicating a degree of ambiguity in their differentiation.

A similar pattern is observed in the classification of IAD categories. Categories 1 and 2 prove challenging to classify accurately due to their limited sample sizes. Conversely, categories 3 and 4, while yielding better classification results, also exhibit tendencies for mutual misclassification.

Figure 5. Confusion matrices showing the best benchmarks for different classification tasks: (a) binary classification, (b) pressure ulcer (PU) category, and (c) incontinence-associated dermatitis (IAD) category.



Performance Comparison

To evaluate our model's performance, we conducted a comparative analysis against the initial hospital inputs recorded in the primary hospital systems, as well as the annotation performance of 2 wound experts and a health care provider without extensive wound expertise on the test dataset. Since the primary hospital system does not include detailed wound degree information, we limited this comparison to binary classification. Specifically, because the electronic health records in the hospitals only document the presence of PU, we assumed that all other labels correspond to IAD for the purposes of this comparison. Furthermore, we assessed model performance solely on the subset of data labeled as PUs.

As shown in Table 4, our AI model demonstrates a significant improvement in both accuracy and F1 score compared to the initial hospital inputs and health care provider annotations. Notably, the model also slightly outperforms the wound care experts on the test dataset, indicating its potential to assist in clinical decision-making.

In addition to this binary classification analysis, we evaluated the model's performance on the test datasets with respect to individual wound degree classification, as shown in Table 5. Also, in this more complex classification task, the AI model outperforms the individual wound experts and health care providers.

Table 4. Model performance comparison binary.

Method	All images		PU[a] only
	Accuracy	F_1-score	Accuracy
AI[b] model	0.9412	0.9323	0.9532
Primary system	0.8190	0.7260	0.8366
Wound expert 1	0.8959	0.8774	0.9281
Wound expert 2	0.8914	0.8773	0.8889
Health care provider	0.8190	0.7736	0.9150

[a] PU: pressure ulcer.
[b] AI: artificial intelligence.

Table 5. Model performance comparison for PU[a] and IAD[b] categories.

Method	PU category		IAD category	
	Accuracy	F_1-score	Accuracy	F_1-score
AI[c] model	0.8255	0.7543	0.5655	0.5320
Wound expert 1	0.7047	0.5284	0.4328	0.3445
Wound expert 2	0.7181	0.5229	0.3881	0.2941
Health care provider	0.4698	0.3295	0.1642	0.1450

[a] PU: pressure ulcer.
[b] IAD: incontinence-associated dermatitis.
[c] AI: artificial intelligence.

Discussion

Principal Findings

In this study, we developed a multimodal deep learning framework for the fine-grained classification of PUs and IAD, along with their respective wound severities. By integrating wound images with categorical patient data, we aimed to enhance diagnostic accuracy and support clinical decision-making in wound care management.

Our extensive evaluations demonstrated that transformer-based architectures, particularly TinyViT, achieved superior performance across the classification tasks. The TinyViT model attained an F_1-score of 93.23% in the binary classification of PU and IAD, outperforming both wound care experts and nursing staff on the test dataset. This highlights the model's effectiveness in handling complex visual data and its potential to assist clinicians in accurately distinguishing between these 2 conditions. In the fine-grained classification of PU categories, the TinyViT model again showed the best performance with an F_1-score of 75.43%. However, the performance was notably lower than in the binary classification task, indicating the increased difficulty in distinguishing between the stages of PU due to subtle visual differences and class imbalances—particularly in differentiating the PU categories stages 1 and 2. Similarly, for IAD category classification, the ConvNeXtV2 model performed best with an F_1-score of 53.20%, but the overall performance was modest, reflecting challenges in differentiating between IAD severity levels.

These findings indicate that while our models effectively distinguish between PU and IAD, their performance in classifying the specific categories within each condition can be enhanced, particularly due to challenges posed by subtle visual differences and class imbalances. Misclassifications often occurred between adjacent categories, which may be due to overlapping visual features and insufficient samples in certain classes. This underscores the need for larger and more balanced datasets to enhance model training and improve classification accuracy in fine-grained tasks. To address this, future research could focus on targeted data collection to increase underrepresented classes. Additionally, exploring advanced synthetic data generation techniques could provide valuable insights, as our study demonstrated the effectiveness of stable diffusion–based synthetic oversampling.

The integration of multimodal data, which combines images with patient information, was beneficial in the binary classification task, enhancing the model's ability to differentiate between PU and IAD. This highlights the importance of contextual clinical information in supporting image-based diagnoses. However, the inclusion of multimodal data had less impact on the fine-grained classification tasks. This may be because the categorical patient data do not provide sufficient granularity to assist in distinguishing between wound severities within PU or IAD. Augmentation strategies played a significant role in model performance. Lighter augmentations were more effective for the binary classification task, possibly because they preserved essential image features while providing variability. In contrast, more intensive augmentations benefited the fine-grained classification tasks by helping the models generalize better to subtle variations in wound appearances. This indicates that augmentation techniques should be tailored to the specific requirements of each classification task.

Synthetic data generation and oversampling proved particularly effective in mitigating class imbalances in the PU and IAD category classification tasks, enhancing the model's ability to learn from underrepresented classes. Notably, the synthetic oversampling approach demonstrated superior performance compared to traditional oversampling, highlighting its potential for improving classification in highly imbalanced settings. In terms of postprocessing, ensembling predictions from multiple folds consistently improved model performance across all tasks, providing more robust and reliable results. While test-time augmentation also contributed to performance gains, its impact was less pronounced compared to ensembling.

These findings contribute valuable insights into the development of more effective diagnostic tools and algorithms for wound classification. By addressing the challenges identified, future work can focus on enhancing the precision and utility of clinical assessments, ultimately improving patient care outcomes.

Limitations

This study, while providing significant insights into the classification of PU and IAD using advanced AI techniques, has certain limitations that warrant consideration. The dataset used, although comprehensive, may not adequately represent the vast diversity of clinical environments and patient demographics. This could limit the generalizability of the findings to other settings or populations. Additionally, inherent class imbalances within the dataset, despite efforts to mitigate their effects through techniques like oversampling and synthetic data generation, might have influenced the model's learning, potentially skewing the accuracy toward more frequently represented classes.

Moreover, the integration of multimodal data did not uniformly improve performance, indicating that its effectiveness varies depending on the data's context and characteristics. This suggests a need for further investigation into which data types are most useful and how they should be integrated.

Furthermore, the study did not exhaustively evaluate every conceivable combination of models, augmentations, training techniques, and postprocessing methods. Instead, selections were based on educated predictions, leveraging the highest-performing techniques from prior phases of the research. This approach, while efficient, may have overlooked potentially effective combinations that could offer further insights or enhanced performance. Additionally, fine-grained classification remains a challenging task due to subtle visual differences between wound categories. Future work should explore attention mechanisms to highlight key image regions and improve model focus, as well as few-shot learning techniques to enhance performance on underrepresented classes. Lastly, the comparison of the model's performance with wound care experts and primary systems is constrained by the specific test dataset used in this study, and as such, the findings may not be fully generalizable to broader and more diverse datasets or clinical scenarios.

Conclusions

This study has successfully implemented a framework for classifying PUs and IAD using advanced artificial intelligence methodologies. By systematically evaluating various computational strategies, including different model architectures, augmentation techniques, training methods, and postprocessing approaches, this research provides valuable insights into optimizing AI-driven wound classification models and their potential for real-world clinical application.

The exploration revealed that transformer-based models, notably the TinyViT, generally outperform other architectures, highlighting their suitability for complex visual data processing in fine-grained applications. The effectiveness of different augmentation strategies varied with the complexity of the classification task, emphasizing the need for tailored approaches depending on the specific requirements of the data and the classification objectives.

Furthermore, the study highlights the value of multimodal data integration in enhancing classification accuracy in specific contexts, though its effectiveness varies across tasks. In addition, our findings emphasize the importance of addressing class imbalances, where both classic and synthetic oversampling significantly improved performance, particularly in tasks with severe class disparities. Notably, synthetic oversampling demonstrated superior effectiveness, suggesting that generative models can serve as a powerful tool for augmenting underrepresented classes. Finally, the superior performance of ensembling in postprocessing underscores its potential as a robust strategy for improving prediction reliability, particularly in clinical applications.

In conclusion, our work presents a highly effective classification model capable of accurately distinguishing between PU and IAD images. This model can serve as a valuable tool to assist health care providers in making correct diagnoses, thereby enhancing clinical decision-making and improving patient outcomes in wound care management. The application of our model has the potential to streamline the diagnostic process, reduce the burden on medical staff, and ensure that patients receive appropriate and timely treatment. Furthermore, our extensive benchmarking provides a valuable reference and guidance for future research and development in wound image classification, contributing to the advancement of practical applications within the domain.

Acknowledgments

We would like to express our sincere gratitude to the wound care experts and nursing staff at University Hospital Essen and Ludwig Maximilian University Hospital Munich for their invaluable assistance in data collection and annotation. We also thank the KIADEKU project team for their collaboration and support throughout this study. This work was supported by the German Federal Ministry of Education and Research (V6KIP039)

Data Availability

The code used in this study will be made publicly available on GitHub [40]. Model weights can be shared upon request. Due to the sensitive nature of the medical image data obtained from a private hospital, the datasets used in this study are not publicly available and cannot be shared to protect patient privacy and confidentiality.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Binary Classification Results.

[XLSX File (Microsoft Excel File), 13 KB-Multimedia Appendix 1]

Multimedia Appendix 2

PU Classification Results.

[XLSX File (Microsoft Excel File), 13 KB-Multimedia Appendix 2]

Multimedia Appendix 3

IAD Classification Results.

[XLSX File (Microsoft Excel File), 13 KB-Multimedia Appendix 3]

References

1. Li Z, Lin F, Thalib L, Chaboyer W. Global prevalence and incidence of pressure injuries in hospitalised adult patients: a systematic review and meta-analysis. Int J Nurs Stud. May 2020;105:103546. [doi: 10.1016/j.ijnurstu.2020.103546] [Medline: 32113142]
2. Ousey K, O'Connor L. IAD made easy. 2017. URL: https://eprints.hud.ac.uk/id/eprint/31572/1/content_11936.pdf [Accessed 2025-04-24]

3. Beeckman D. Romanelli M, Clark M, Gefen A, Ciprandi G, editors. Incontinence-Associated Dermatitis (IAD) and Pressure Ulcers: An Overview. Springer; 2018:89-101. [doi: 10.1007/978-1-4471-7413-4_7]
4. Miteinander durch Innovation. KIADEKU — miteinander durch innovation [Website in German]. URL: https://www.interaktive-technologien.de/projekte/kiadeku [Accessed 2024-05-16]
5. Aldughayfiq B, Ashfaq F, Jhanjhi NZ, Humayun M. YOLO-based deep learning model for pressure ulcer detection and classification. Healthcare (Basel). Apr 25, 2023;11(9):37174764. [doi: 10.3390/healthcare11091222] [Medline: 37174764]
6. Chang CW, Christian M, Chang DH, et al. Deep learning approach based on superpixel segmentation assisted labeling for automatic pressure ulcer diagnosis. PLoS ONE. 2022;17(2):e0264139. [doi: 10.1371/journal.pone.0264139] [Medline: 35176101]
7. Seo S, Kang J, Eom IH, et al. Visual classification of pressure injury stages for nurses: a deep learning model applying modern convolutional neural networks. J Adv Nurs. Aug 2023;79(8):3047-3056. [doi: 10.1111/jan.15584] [Medline: 36752192]
8. García-Zapirain B, Elmogy M, El-Baz A, Elmaghraby AS. Classification of pressure ulcer tissues with 3D convolutional neural network. Med Biol Eng Comput. Dec 2018;56(12):2245-2258. [doi: 10.1007/s11517-018-1835-y] [Medline: 29949023]
9. Liu TJ, Christian M, Chu YC, et al. A pressure ulcers assessment system for diagnosis and decision making using convolutional neural networks. J Formos Med Assoc. Nov 2022;121(11):2227-2236. [doi: 10.1016/j.jfma.2022.04.010] [Medline: 35525810]
10. Lau CH, Yu KHO, Yip TF, et al. An artificial intelligence-enabled smartphone app for real-time pressure injury assessment. Front Med Technol. 2022;4(905074):905074. [doi: 10.3389/fmedt.2022.905074] [Medline: 36212608]
11. Kim J, Lee C, Choi S, et al. Augmented decision-making in wound care: evaluating the clinical utility of a deep-learning model for pressure injury staging. Int J Med Inform. Dec 2023;180(105266):105266. [doi: 10.1016/j.ijmedinf.2023.105266] [Medline: 37866277]
12. Swerdlow M, Guler O, Yaakov R, Armstrong DG. Simultaneous segmentation and classification of pressure injury image data using Mask-R-CNN. Comput Math Methods Med. 2023;2023(3858997):3858997. [doi: 10.1155/2023/3858997] [Medline: 36778787]
13. Zahia S, Sierra-Sosa D, Garcia-Zapirain B, Elmaghraby A. Tissue classification and segmentation of pressure injuries using convolutional neural networks. Comput Methods Programs Biomed. Jun 2018;159(51-58):51-58. [doi: 10.1016/j.cmpb.2018.02.018] [Medline: 29650318]
14. Pandey B, Joshi D, Arora AS. A deep learning based experimental framework for automatic staging of pressure ulcers from thermal images. Quant Infrared Thermogr J. 2024:1-21. [doi: 10.1080/17686733.2024.2390719]
15. Huang PH, Pan YH, Luo YS, et al. Development of a deep learning-based tool to assist wound classification. J Plast Reconstr Aesthet Surg. Apr 2023;79(89-97):89-97. [doi: 10.1016/j.bjps.2023.01.030] [Medline: 36893592]
16. Oura P, Junno A, Junno JA. Deep learning in forensic gunshot wound interpretation-a proof-of-concept study. Int J Legal Med. Sep 2021;135(5):2101-2106. [doi: 10.1007/s00414-021-02566-3] [Medline: 33821334]
17. Rostami B, Anisuzzaman DM, Wang C, Gopalakrishnan S, Niezgoda J, Yu Z. Multiclass wound image classification using an ensemble deep CNN-based classifier. Comput Biol Med. Jul 2021;134(104536):104536. [doi: 10.1016/j.compbiomed.2021.104536] [Medline: 34126281]
18. Patel Y, Shah T, Dhar MK, et al. Integrated image and location analysis for wound classification: a deep learning approach. Sci Rep. Mar 25, 2024;14(1):7043. [doi: 10.1038/s41598-024-56626-w] [Medline: 38528003]
19. Liu Z, John J, Agu E. Diabetic foot ulcer ischemia and infection classification using efficientnet deep learning models. IEEE Open J Eng Med Biol. 2022;3(189-201):189-201. [doi: 10.1109/OJEMB.2022.3219725] [Medline: 36660100]
20. Lee S, Lukan J, et al. A deep learning model for burn depth classification using ultrasound imaging. J Mech Behav Biomed Mater. Jan 2022;125(104930):104930. [doi: 10.1016/j.jmbbm.2021.104930]
21. Afza F, Sharif M, Khan MA, Tariq U, Yong HS, Cha J. Multiclass skin lesion classification using hybrid deep features selection and extreme learning machine. Sensors (Basel). Jan 21, 2022;22(3):35161553. [doi: 10.3390/s22030799] [Medline: 35161553]
22. Cheng J, Schmidt C, Wilson A, et al. Artificial intelligence for human gunshot wound classification. J Pathol Inform. Dec 2024;15:100361. [doi: 10.1016/j.jpi.2023.100361] [Medline: 38234590]
23. Odame P, Ahiamadzor MM, Derkyi NKB, et al. Multi-wound classification: exploring image enhancement and deep learning techniques. Engineering Reports. Jan 2025;7(1):e70001. URL: https://onlinelibrary.wiley.com/toc/25778196/7/1 [doi: 10.1002/eng2.70001]
24. Tkachenko M, Malyuk M, Holmanyuk A, Liubimov N. Label Studio: data labeling software. 2020. URL: https://github.com/heartexlabs/label-studio [Accessed 2025-05-17]

25. Büscher A, Blumenberg P, Krebs M, Stehling H, Stomberg D. Expertenstandard dekubitusprophylaxe in der pflege, 2. aktualisierung 2017, stand: mai 2021. schriftenreihe des deutschen netzwerks für qualitätsentwicklung in der pflege. hochschule osnabrück, fakultät für wirtschafts- und sozialwissenschaften. URL: https://www.dnqp.de/fileadmin/HSOS/Homepages/DNQP/Dateien/Expertenstandards/Dekubitusprophylaxe_in_der_Pflege/Dekubitus_2Akt_Auszug.pdf. [Accessed 2024-05-18]. 2021.
26. Beeckman D, Van den Bussche K, Alves P, et al. Towards an international language for incontinence-associated dermatitis (IAD): design and evaluation of psychometric properties of the Ghent Global IAD Categorization Tool (GLOBIAD) in 30 countries. Br J Dermatol. Jun 2018;178(6):1331-1340. [doi: 10.1111/bjd.16327] [Medline: 29315488]
27. Huggingface/pytorch-image-models. Hugging Face. URL: https://github.com/huggingface/pytorch-image-models [Accessed 2024-05-17]
28. Woo S, Debnath S, Hu R, et al. ConvNeXt V2: co-designing and scaling convnets with masked autoencoders. Presented at: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); Jun 17-24, 2023:16133-16142; Vancouver, BC, Canada. [doi: 10.1109/CVPR52729.2023.01548]
29. Timm/convnextv2 tiny.fcmae ft in22k in1k 384. Hugging Face. 2023. URL: https://huggingface.co/timm/convnextv2_tiny.fcmae_ft_in22k_in1k_384 [Accessed 2024-05-17]
30. Tan M, Le Q. EfficientNetV2: smaller models and faster training. arXiv. Preprint posted online on Apr 1, 2021. [doi: 10.48550/arXiv.2104.00298]
31. Timm/tf efficientnetv2 s.in21k ft in1k. Hugging Face. URL: https://huggingface.co/timm/tf_efficientnetv2_s.in21k_ft_in1k [Accessed 2025-05-17]
32. Yu W, Si C, Zhou P, et al. MetaFormer baselines for vision. IEEE Trans Pattern Anal Mach Intell. 2024;46(2):896-912. [doi: 10.1109/TPAMI.2023.3329173]
33. Wu K, Zhang J, Peng H, et al. TinyViT: fast pretraining distillation for small vision transformers. Presented at: Computer Vision – ECCV 2022: 17th European Conference; Oct 23-27, 2022:68-85; [doi: 10.1007/978-3-031-19803-8_5]
34. Timm/tiny_vit_21m_384.dist_in22k_ft_in1k. Hugging Face. URL: https://huggingface.co/timm/tiny_vit_21m_384.dist_in22k_ft_in1k [Accessed 2024-05-17]
35. Zhang H, Cisse M, Dauphin YN, MixUp LPD. Beyond empirical risk minimization. arXiv. Preprint posted online on 2018. [doi: 10.48550/arXiv.1710.09412]
36. Cubuk ED, Zoph B, Shlens J, Le QV. Randaugment: practical automated data augmentation with a reduced search space. Presented at: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); Dec 6, 2020:18613-18624; Seattle, WA, USA. [doi: 10.1109/CVPRW50498.2020.00359]
37. Muller SG, Hutter F. TrivialAugment: tuning-free yet state-of-the-art data augmentation. Presented at: 2021 IEEE/CVF International Conference on Computer Vision (ICCV); Oct 10-17, 2021:774-782; Montreal, QC, Canada. [doi: 10.1109/ICCV48922.2021.00081]
38. TorchMetrics. URL: https://github.com/Lightning-AI/metrics [Accessed 2024-09-23]
39. Maier-Hein L, Reinke A, Godau P, et al. Metrics reloaded: recommendations for image analysis validation. Nat Methods. Feb 2024;21(2):195-212. [doi: 10.1038/s41592-023-02151-z] [Medline: 38347141]
40. Multimodal deep learning for fine-grained classification of pressure ulcers and incontinence associated dermatitis. GitHub. URL: https://github.com/AlexariusIII/Multimodal-Deep-Learning-for-Fine-Grained-Classification-of-Pressure-Ulcers-and-Incontinence-Associa [Accessed 2025-04-24]

Abbreviations

AI: artificial intelligence
AP: average precision
AUROC: area under the receiver operating characteristic curve
DRKS: Deutsches Register Klinischer Studien (German Clinical Trials Register)
GLOBIAD: Ghent Global IAD Categorization Tool
IAD: incontinence-associated dermatitis
PU: pressure ulcer

Edited by Yuankai Huo; peer-reviewed by Gaurav Kumar Gupta, Mehrad Aria; submitted 09.10.2024; final revised version received 19.02.2025; accepted 26.02.2025; published 01.05.2025

Please cite as:

Brehmer A, Seibold C, Egger J, Majjouti K, Tapp-Herrenbrück M, Pinnekamp H, Priester V, Aleithe M, Fischer U, Hosters B, Kleesiek J
Fine-Grained Classification of Pressure Ulcers and Incontinence-Associated Dermatitis Using Multimodal Deep Learning: Algorithm Development and Validation Study
JMIR AI 2025;4:e67356
URL: https://ai.jmir.org/2025/1/e67356
doi: 10.2196/67356