

딥러닝 기반 욕창 분석 모바일 애플리케이션의 단계 분류 성능 및 임상 적용 가능성

장상현¹, 강민수², 임춘화³, 허명행⁴, 배경화⁵

¹대전을지대학교병원 신경과, ²을지대학교 빅데이터의료융합학과, ³대전을지대학교병원 진단검사의학과, ⁴을지대학교 간호학과, ⁵대전을지대학교병원 QI부

Stage Classification Performance and Clinical Applicability of a Deep Learning–Based Mobile Application for Pressure Injury Analysis

Sang-Hyun Jang¹, Min-Su Kang², Chun-Hwa Lim³, Myung-Haeng Heo⁴, Kyung-Hwa Bae⁵

¹Professor, Department of Neurology, Daejeon Eulji University Hospital, Daejeon, ²Professor, Department of Big Data Medical Convergence, Eulji University, Seongnam, ³Professor, Department of Laboratory Medicine, Daejeon Eulji University Hospital, Daejeon, ⁴Professor, College of Nursing, Eulji University, Uijeongbu, ⁵Team Leader, Quality Improvement Department, Daejeon Eulji University Hospital, Daejeon Republic of Korea

Purpose: The purpose of this study was twofold: 1) to reduce inter-rater variability in pressure injury staging and 2) to explore the feasibility of a supportive tool for clinical practice by evaluating the agreement between automated classification results and the clinical applicability of a deep learning-based mobile application for pressure injury analysis.

Methods: In this single-center prospective pilot study, clinical images from 20 patients with pressure injuries were used to calculate Cohen's kappa for inter-expert and expertAI agreements. Stage-specific performance analyses were conducted only for pressure injuries classified as Stages 2-4, as no cases of Stage 1, deep-tissue pressure injury, or unstageable pressure injury were observed. In addition, a 10-item survey was administered to 50 nurses to assess ease of learning, interface intuitiveness, clinical usefulness, and overall satisfaction.

Results: The interexpert kappa was 0.703, indicating substantial agreement, whereas the expertAI kappa ranged from 0.615 to 0.730, reflecting moderate to substantial agreement. Nurse satisfaction (N = 50) showed high positive response rates for ease of learning (94%), intuitive usability (88%), and support for clinical judgment (88%). Opportunities for further functional refinement to support broader clinical implementation were identified.

Conclusion: As this was a single-center pilot study, the findings should be interpreted as preliminary. The AI-based application for determining pressure injury stage classification demonstrated moderate to substantial agreement with expert assessments, and nurse satisfaction was generally favorable, suggesting that the application may serve as an assistive decision-support tool for pressure injury staging. Further multicenter studies are warranted to confirm its clinical applicability and performance.

Keywords: Pressure ulcer, Artificial intelligence, Mobile applications, Deep learning

Received: Jan.06.2026 **Revised:** Feb.09.2026 **Accepted:** Mar.03.2026

Correspondence: Kyung-Hwa Bae

Daejeon Eulji University Hospital, 95 Dunsanse-ro, Seo-gu, Daejeon 35233, Republic of Korea

Tel: +82-42-611-3093 **E-mail:** sori1004@eulji.ac.kr

Funding: None **Conflict of Interest:** None

Quality Improvement in Health Care vol.32 no.1

© The Author 2026. Published by Korean Society for Quality in Health Care; all rights reserved

I. 서론

욕창(pressure injury)은 대표적인 환자안전사고 중 하나로, 특히 장기 입원 환자와 중환자에서 빈번하게 발생하는 예방 가능한 합병증이다[1]. 피부 및 피하조직의 국소적 손상에서 시작된 욕창은 적절한 예방과 관리가 이루어지지 않을 경우 감염, 골수염, 패혈증 등 중대한 합병증으로 진행할 수 있으며[2]. 이는 환자의 삶의 질 저하, 입원 기간 연장, 의료비용 증가로 이어진다 [3, 4]. 이러한 특성으로 인해 욕창은 다수 국가에서 대표적인 예방 가능한 병원 내 위해사건(preventable hospital-acquired condition)으로 분류되고 있다[5].

욕창 관리는 예방 중심의 질 관리 영역으로, 입원 초기의 정확한 사정과 단계 판정은 표준화된 간호중재 제공과 성과 지표 관리의 핵심 요소이다[6]. 그러나 실제 임상에서는 욕창 단계 판정이 평가자의 임상 경험과 숙련도에 크게 의존하며, 미세한 병변 차이를 구분하는 데 어려움이 있어 평가자 간 변이가 발생할 수 있다[7].

특히 미국 욕창 자문위원회(National Pressure Injury Advisory Panel, NPIAP) 기준에 따라 욕창은 Stage 1 - 4 뿐 아니라 심부조직손상(deep tissue pressure injury, DTPI)과 미분류(unstageable)를 포함하는 복합적 분류 체계를 가지므로, 초기 판정의 정확성과 일관성은 치료 방향과 간호중재 선택에 직접적인 영향을 미친다[8, 9]. 최근 AI 기반 영상 분석 기술과 모바일 애플리케이션을 활용한 욕창 사정 보조 시스템이 개발되면서, 욕창 단계 판정의 표준화와 평가자 간 변이 감소를 위한 대안으로 주목받고 있다[10].

Seo 등[11]은 CNN (Convolutional Neural Network) 모델을 활용하여 욕창 단계를 자동 분류한 결과, 임상 간호사 판정과 유사한 수준의 정확도를 보고하였다. Chang 등 [10]은 YOLOv8m 기반 경량 모델을 모바일 환경에 적용하여 Stage 1 - 4, DTPI, unstageable을 포함한 자동 분류 시스템을 개발하고 전체 정확도 84.6%를 보고함으로써, 모바일 환경에서도 임상적으로 활용 가능한 성능을 입증하였다. 그러나 기존 상용 욕창 분석 애플리케이션은 주

로 국외 알고리즘에 기반하거나 국내 임상 환경에서의 검증이 충분하지 않아 실제 간호 실무 적용에는 한계가 있다. 이에 이 연구는 병원과 대학 연구팀이 공동 개발한 AI 기반 욕창 분석 모바일 애플리케이션을 욕창 사정의 표준화와 현장 적용을 위한 질 향상(Quality Improvement, QI) 도구로 보고, 전문가 - 인공지능(Artificial Intelligence, AI)간 일치도와 간호사 수용성을 통합적으로 평가하여 임상 적용 가능성을 검증하고자 하였다.

II. 연구 목적

이 연구의 목적은 AI 기반 욕창 분석 모바일 애플리케이션의 자동 분류 성능과 임상 적용 가능성을 평가하는 것이다. 이를 위해 연구자 판정과 AI 자동 분류 결과 간 일치도를 비교·분석하고, 의료진을 대상으로 사용 편의성과 임상적 유용성에 대한 만족도를 조사함으로써, 이 애플리케이션의 임상적 신뢰성과 실제 간호 실무에서의 적용 가능성을 검증하고자 하였다.

III. 연구방법

1. 연구 설계

이 연구는 딥러닝 기반 욕창 단계 자동 분류 모바일 애플리케이션의 임상적 타당성과 실제 사용 편의성을 평가하기 위해 수행된 전향적 기술·비교 파일럿 연구(descriptive comparative prospective pilot study)이다. 이 연구는 AI 모델 개발을 위한 임상 이미지 제공 단계부터 실제 임상 환경에서의 파일럿 검증 및 간호사 만족도 평가에 이르기까지 일련의 다단계 구조로 설계되었다.

첫째, AI 학습용 데이터 구축을 위해 병원에서는 총 1,200장의 욕창 임상 이미지를 수집하여 대학 연구팀에 제공하였다. 이 이미지는 1-4단계, DTPI, unstageable 등 다양한 욕창 단계와 병변 특성을 포함하고 있었으며, 이미지 품질, 촬영 각도, 피부 톤 등이 임상 현장의 실제 변동성을 반영할 수 있도록 포괄적으로 구성되었다. 제공된 모든

이미지는 철저한 비식별화 절차를 거친 후 AI 학습·검증용 데이터셋으로 활용되었다.

둘째, 대학 연구팀은 제공된 임상 이미지를 기반으로 AI 알고리즘 및 모바일 애플리케이션을 개발하였다. 딥러닝 기반 객체 탐지 모델을 사용하여 상처 영역을 자동으로 인식하고, NPIAP기준에 따라 욕창 단계를 분류하는 모델을 구축하였다. 이후 모델 성능을 최적화하고 모바일 환경에서 실시간 분류가 가능하도록 애플리케이션 형태로 구현하였다.

셋째, 개발된 애플리케이션의 분류 성능과 신뢰성을 확인하기 위해 욕창 환자 20명으로부터 수집된 임상 파일럿 평가용 이미지를 활용하여 전문가 판정과 AI 자동 분류 결과 간의 일치도를 분석하였다. 욕창 사정 경험이 풍부한 간호사 2인이 NPIAP 기준에 따라 독립적으로 이미지를 판독하였으며, 동일 이미지에 대해 AI 자동 분류 결과를 산출하여 연구자 - AI 간의 분류 일치도를 평가하였다.

넷째, 애플리케이션의 사용자 수용성과 실무 적합성을 평가하기 위해 이 연구에서는 애플리케이션 사용 경험이 있는 임상 간호사 50명을 대상으로 학습 용이성, 사용 편의성, 인터페이스 직관성, 임상적 유용성 등 실제 사용 경험에 대한 만족도를 조사하였다. 이를 통해 자동 분류 기능뿐 아니라 실제 간호업무 흐름에서의 활용 가능성을 포괄적으로 검토하였다.

이 연구는 파일럿 연구의 특성에 따라 초기 임상 적용 가

능성을 탐색하는 데 중점을 두었으며, 연구의 구성·수행·보고는 CONSORT 2010 확장 가이드라인(Consolidated Standards of Reporting Trials 2010 extension for pilot and feasibility trials)을 준용하여 이루어졌다[12]. 이러한 구조화된 연구 설계를 통해 모바일 AI 애플리케이션의 기술적 성능과 임상 실무 적용성을 다각도로 평가하고자 하였다.

2. AI개발을 위한 임상 이미지 제공

이 연구에서 사용된 인공지능 모델의 학습 및 검증을 위해, 병원에서는 욕창 환자의 병변 부위를 촬영한 임상 이미지를 수집하여 대학 연구팀에 제공하였다. 총 1,200장의 임상 이미지는 욕창의 다양한 단계(Stage 1 - 4, DTPI, unstageable)를 포함하고 있고, 촬영 각도, 조명 조건, 피부톤 등 실제 임상 환경에서 발생할 수 있는 변동성을 충분히 반영하였다.

이미지 제공에 앞서 모든 자료는 개인 식별이 불가능하도록 철저한 비식별화 절차를 거쳤다. 병원은 임상 데이터 제공 및 요구 기능 정의에 한하여 연구에 참여하였으며, 인공지능 알고리즘의 설계 및 개발 과정에는 직접 관여하지 않았다. Figure 1에는 이 연구에 활용된 임상 이미지의 예시를 제시하였다.



Figure 1. Data collection example.

3. AI 모바일 애플리케이션 개발(대학 연구팀)

AI 알고리즘 개발과 모바일 애플리케이션 구현은 대학 연구팀이 전적으로 수행하였다. 제공된 임상 이미지로 학습 데이터셋을 구축한 뒤, 딥러닝 기반 객체 탐지 모델을 활용해 욕창 상처 영역을 자동 인식하고 CNN 기반 분류 모델을 적용하여 NPIAP 기준에 따라 Stage 1 - 4를 자동 분류하는 알고리즘을 개발하였다 완성된 모델은 모바일 애플리케이션과 연동되도록 구현되었으며, 간호사가 이미지를 업로드하면 실시간으로 단계가 제시되도록 설계하였다(Figure 2).

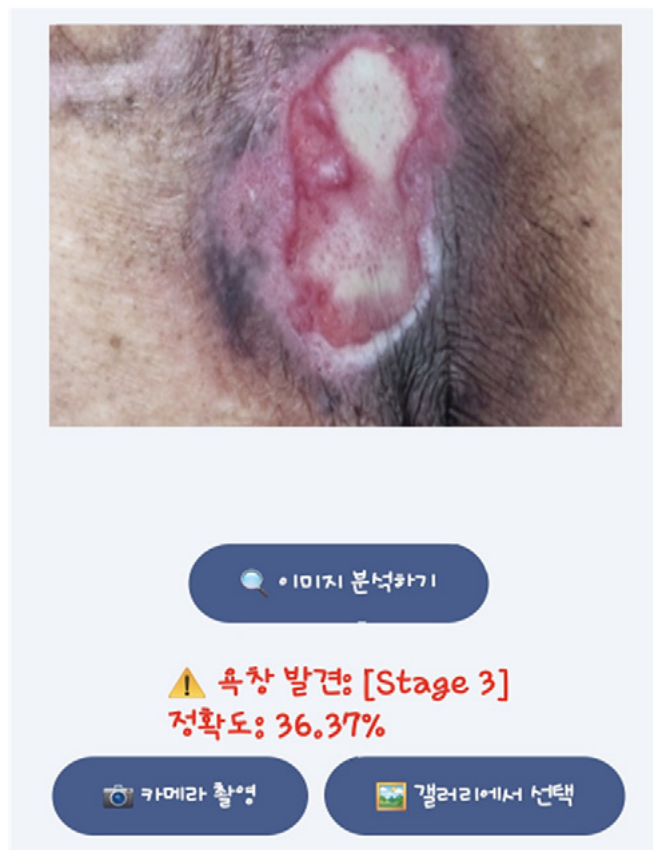


Figure 2. Development and integration of the research mobile application.

4. 임상 파일럿 연구 절차

1) 연구 대상자 산출 근거 및 모집

파일럿 연구에서의 적정 표본 수는 Hertzog [13] 및 Julious [14]의 제안에 따라 10-30명 수준의 소규모 표본이 일반적으로 사용되며, 실제 의료영상 기반 파일럿 연구에서도 유사한 규모가 적용되고 있다. 또한 Chang 등 [10]은 딥러닝 기반 모바일 애플리케이션을 개발하고 파일럿 수준의 테스트를 통해 이미지 기반 모델을 평가하였으며, 초기 검증 단계에서는 제한된 수의 이미지를 대상으로 실제 적용 가능성을 탐색하는 방식을 사용한 바 있다.

이에 이 연구는 임상 적용 가능성과 시스템 작동 정확성을 확인하는 전향적 파일럿 연구로서, CONSORT 2010 확장 가이드라인을 준수[12]하여 설계·보고하였다. 임상 파일럿용 평가를 위해 욕창 환자 20명으로부터 동의를 구한 후 수집한 이미지를 활용하여 연구자 판정과 AI 자동 분류 결과 간 일치도를 분석하였다.

2) 연구 간호사 선정 및 평가

욕창 단계 분류에 대한 임상적 전문성을 확보하기 위해, 욕창 사정 경험이 풍부한 임상 경력 10년 이상의 간호사 2인을 전문가 평가자로 선정하였다. 선정된 간호사들은 모두 NPIAP기준에 기반한 욕창 단계 분류 교육을 사전에 받았으며, 연구 과정에서 서로의 평가에 영향을 주지 않도록 독립적으로 판독을 수행하였다.

3) 욕창 병변 이미지 수집 및 분류

(1) 욕창 평가 단계 일치도 분석

연구자 평가는 욕창 평가 경험이 풍부한 연구 간호사 2인이 수행하였다. 평가 전, NPIAP 기준에 따라 욕창 단계를 일관되게 분류할 수 있도록 평가 지침과 표준화된 예시 이미지를 활용한 사전 교육을 시행하였다. 연구 간호사 1

은 연구 대상자로부터 사전 동의를 받은 후 각 환자의 욕창 사정기록지 및 임상 소견을 참고하여 병변 부위를 욕창으로 독립적으로 판독하였으며, 다른 평가자의 영향을 받지 않도록 상호 간섭이 없는 환경에서 평가를 진행하였다. 판독 결과는 단계별로 기록하였고, 동일 시점에 병변 부위의 임상 이미지를 함께 수집하였다.

(2) 연구 간호사의 욕창 이미지 판독

또 다른 연구 간호사 1인은 비식별화된 욕창 이미지를 제공받아 동일한 NPIAP 기준을 적용하여 단계별 라벨링을 수행하였다. 이 평가는 환자 정보가 완전히 제거된 상태에서 독립적으로 이루어졌으며, 각 이미지를 stage 1-4단계 중 하나로 분류하였다. 두 연구 간호사 간 평가는 서로 간섭 없이 개별적으로 수행되었으며, 동일한 기준과 판정 정의를 사용하여 평가 간 신뢰도를 확보하고자 하였다. 이 연구에서 사용한 NPIAP 기반 욕창 단계 분류 체계는 Figure 3[9]에 제시하였다.

(3) AI 기반 자동 분류

AI 기반 분석은 딥러닝 알고리즘이 탑재된 모바일 애플리케이션을 활용하여 수행되었다. 촬영된 욕창 이미지는 애플리케이션에 입력되었으며, 내장된 CNN 기반 분류 알고리즘에 의해 욕창 상처 영역이 자동 탐지된 후 1-4단계 중

하나로 분류되었다. AI가 산출한 단계 결과는 자동으로 기록되었으며, 연구 간호사 판독 결과와 동일한 형식으로 저장하여 이후 일치도 및 진단 성능 분석에 활용하였다.

4) AI 결과와 연구자 평가 결과 비교

AI 자동 분류 결과와 두 연구자 평가 결과 간 일치도는 Cohen의 카파 계수(Cohen's kappa)를 사용하여 분석하였다. Fleiss의 카파 계수는 두 연구자와 AI를 포함한 다평가자 환경에서의 일치도를 평가하기 위해 사용하였다. 단계별 판정의 정확도(accuracy), 민감도(sensitivity), 특이도(specificity)를 산출하여 AI 모델의 분류 성능을 평가하였다. 이 연구에서의 정확도는 독립적인 기준(reference standard)에 근거한 진단 정확도가 아니라, 연구자 판정을 기준으로 산출한 AI분류 성능(Classification performance)지표를 의미한다.

Stage 1, DTPI, unstageable 사례는 없어 단계별 성능 분석에서 제외하였다. 이러한 분석은 연구자 간 평가 일관성을 확인하고, AI 분류 결과가 전문가 판독과 어느 수준까지 일치하는지를 평가하기 위한 목적으로 수행되었다. 두 연구자는 각 이미지에 대해 독립적으로 욕창 단계를 판정하였으며, 두 연구자의 판정이 서로 다른 경우에도 해당 사례를 제외하거나 조정하지 않고 불일치(disagreement)로 처리하여 Cohen의 카파 계수 산출에 그대로 반영하였다.

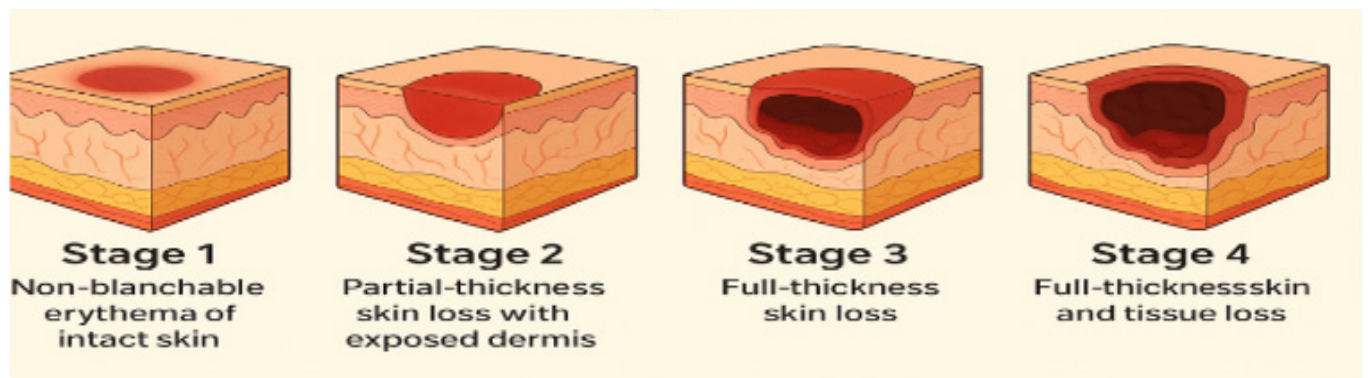


Figure 3. NPIAP-based pressure injury staging system.

Adapted from national pressure injury advisory panel (2016). Pressure injury stages and definitions.

5) 간호사 만족도 및 사용성 조사

이 연구의 간호사 만족도 설문지는 의료진 대상 애플리케이션 만족도 평가 선행 연구[15]와 전문직 실무용 애플리케이션 인지된 만족도 및 사용성 도구(Perceived Satisfaction and Usability of Applications for Professional Practice, pSUAPP)의 신뢰도·타당도 구조[16]를 참고하여 구성하였으며, 학습 용이성·인터페이스·임상 유용성·전반적 만족도 등 10개 문항을 5점 Likert 척도로 평가하도록 하였다. 설문은 애플리케이션을 일정 기간 사용한 간호사 50명을 대상으로 시행하였고, 도구의 신뢰도는 Cronbach's $\alpha = 0.90$ 으로 확인되었다.

5. 자료 분석

통계 분석은 SPSS 29.0 프로그램을 사용하였으며 각 분석 지표의 정의와 산출 방식, 해석 기준은 Table 1에 제시하였다.

- 1) 대상자의 일반적 특성(연령, 성별, 병동 유형, 욕창 발생 부위, 기저질환 등)은 평균, 표준편차, 빈도 및 백분율을 이용한 기술통계로 분석하였다.
- 2) 욕창 단계 분류의 일치도는 전문가 1 - 전문가 2(연구 간호사 간), 전문가 - AI 간 Cohen의 카파 계수를 산출하여 평가하였다. 또한 각 평가자 간 분류 결과의 전체 일

치율(percent agreement), 평균, 표준편차, 최소 - 최대값(min - max)을 산출하여 전체 분류 일치율을 산출하였다. 세 평가자(전문가 1·전문가 2·AI) 간의 다자 간 일치도는 Fleiss의 카파 계수(Fleiss' kappa)를 이용하여 분석하였다.

- 3) AI 자동 분류 성능을 보다 구체적으로 평가하기 위해 전문가 평가를 기준(reference standard)으로 하여, 단계별 민감도(sensitivity)와 특이도(specificity)를 산출하였다.
- 4) 간호사 만족도는 Likert 5점 척도로 측정된 각 문항의 빈도와 백분율을 산출하여 분석하였다.
- 5) 모든 분석에서 통계적 유의수준은 양측 검정 기준으로 $p < .05$ 로 설정하였다.

6. 윤리적 고려

이 연구는 E대학교병원 기관생명윤리위원회(Institutional Review Board, IRB)의 승인(IRB No. 2025-06-003)을 받은 후 수행되었다. 모든 대상자에게 연구 목적과 절차, 개인정보 보호 방침을 충분히 설명한 뒤 서면 동의를 얻어 이미지를 수집하였다. 수집된 욕창 이미지는 모두 비식별화하여 개인 식별 정보가 포함되지 않도록 처리하였으며, 연구 목적 외 다른 용도로 사용되지 않았다. 모든 자료는 기관 규정에 따라 보안이 유지되는 환경에서 안전하게 관리하였다.

Table 1. Definitions of agreement and performance metrics used in the quality improvement evaluation.

Category	Analysis metric	Definition and calculation	Interpretation and purpose
Agreement analysis	Cohen's Kappa	Measures the agreement between the research nurses' assessments and AI classification results	A coefficient that adjusts for agreement occurring by chance beyond simple percent agreement. Values between 0.61 and 0.80 are generally interpreted as substantial agreement (Landis and Koch)
Performance analysis	Accuracy	$(TP^1 + TN^2) / (TP^1 + TN^2 + FP^3 + FN^4)$	Proportion of correctly classified cases among all evaluated cases
	Sensitivity (Recall)	$TP^1 / (TP^1 + FN^4)$	Proportion of true cases of a given stage correctly identified by the AI
	Specificity	$TN^2 / (TN^2 + FP^3)$	Proportion of non-target cases correctly excluded by the AI
Statistical test	p-value	$p < .05$	Statistical significance threshold

1) TP: True Positive 2) TN: True Negative 3) FP: False Positive 4) FN: False Negative

IV. 연구 결과

1. 대상자의 일반적 특성

이 연구는 욕창 환자 20명을 대상으로 하였으며, 대상자의 일반적 특성은 Table 2와 같다. 성별 분포는 남성이 11명(55.0%)으로 여성 9명(45.0%)보다 다소 많았다. 평균 연령은 68.9세(± 15.9)로, 연구 대상자의 연령대가 비교적 고령층에 집중되어 있었다. 세부 연령군을 살펴보면 60-69세가 25.0%로 가장 많았고, 80세 이상 초고령군이 35.0%를 차지하여 전체의 3분의 1 이상이 초고령 환자였다. 이는 욕창이 고령 환자군에서 흔히 발생하는 특성과 일치하는 결과이다.

입원 병동 유형은 내과병동이 35.0%, 외과병동이

20.0%, 중환자실이 45.0%였다. 특히 중환자실 입원 환자가 전체의 절반에 가까운 비율을 차지하고 있어, 욕창이 중증 환자나 장기 입원 환자에서 더욱 빈번하게 발생함을 시사한다.

욕창 발생 부위는 천골(sacrum)이 60.0%로 가장 높은 빈도를 보였으며, 둔부(buttocks)가 30.0%, 척추(spinal)가 5.0%, 천골과 후방이 동시에 관찰된 경우가 5.0%였다. 이는 천골 부위가 체중 부하와 압박에 취약하여 욕창이 가장 흔히 발생하는 부위라는 기존 문헌과 일관된 결과이다[2].

욕창 발생 시점은 입원 전에 이미 발생한 경우가 35.0%, 입원 후 발생한 경우가 65.0%로, 병원 입원 중 발생한 욕창이 더 많았다. 이는 장기적인 체위 유지, 기동성 저하, 중환자 치료 과정 등이 욕창 발생의 주요 요인으로 작용했음을 시사한다.

Table 2. General characteristics

(N=20)

Characteristics	Categories	n (%)	Mean $\pm$ SD
Gender	Male	11 (55.0)	68.9 $\pm$ 15.9
	Female	9 (45.0)	
Age (yr)	40-49	2 (10.0)	
	50-59	4 (20.0)	
	60-69	5 (25.0)	
	70-79	2 (10.0)	
	≥ 80	7 (35.0)	
Ward type	Internal medicine ward	7 (35.0)	
	Surgical ward	4 (20.0)	
	Intensive care unit	9 (45.0)	
Pressure Injury Location	Buttocks	6 (30.0)	
	Sacrum	12 (60.0)	
	Buttocks, Sacrum	1 (5.0)	
	Spinal	1 (5.0)	
Onset of Pressure Injury	Before hospitalization	7 (35.0)	
	Hospitalized	13 (65.0)	
Underlying disease ¹⁾	Hypertension	11 (55.0)	
	Cancer	6 (30.0)	
	Diabetes	4 (20.0)	
	Chronic Kidney Disease	2 (10.0)	
	Others	3 (15.0)	
	doesn't exist	2 (10.0)	

1) Participants could have multiple underlying diseases; therefore, the sum of percentages exceeds 100%.

기저질환은 고혈압(Hypertension)이 55.0%로 가장 많았으며, 암(Cancer) 30.0%, 당뇨병(Diabetes) 20.0%, 만성 신부전(Chronic Kidney Disease) 10.0% 순으로 나타났다. 대상자의 상당수는 고혈압을 포함하여 당뇨병, 암, 만성 신부전 등의 기저질환을 단독 또는 복합적으로 보유하고 있었으며, 기저질환은 중복 응답이 가능하여 각 항목의 비율 합은 100%를 초과하였다. 이는 대부분의 환자가 복합적인 만성질환을 가지고 있었음을 의미하며, 이러한 기저질환이 욕창 발생 위험을 높이는 중요한 요인임을 보여준다.

2. 연구자 및 AI 간 욕창 단계 분류

전체 20건의 욕창 사례 중 Stage 2는 13건, Stage 3는 5건, Stage 4는 2건이었으며, Stage 1, DTPI, unstageable 사례는 없어 단계별 성능 분석에서 제외되었다.

Table 3은 연구자 간, 연구자와 AI 간, 그리고 세 집단(연구자 1 · 연구자 2 · AI) 간의 전체 일치율과 카파 계수를 제시한 것이다. 연구자 1과 연구자 2의 전체 일치율은

85.0%였으며, Cohen의 카파 계수는 0.703으로 나타나 상당한 수준의 합의(substantial agreement)를 보였다. 이는 두 연구자 간 욕창 단계 판정이 비교적 안정적이고 신뢰할 수 있는 수준임을 의미한다.

연구자 1과 AI 간 전체 일치율은 80.0%, 카파 계수는 0.615로 확인되었다. 이는 중등도에서 상당한 수준(moderate to substantial)의 합의를 보여주며, AI의 판정이 연구자 1의 판정과 대부분 일치했음을 시사한다.

연구자 2와 AI 간 전체 일치율은 85.0%, 카파 계수는 0.730으로, 연구자 1 - AI 간 결과보다 더 높은 합의 수준을 보였다. 이는 AI 기반 분류 결과가 연구자 2의 평가와 비교적 높은 수준의 일치도를 보였음을 의미한다.

마지막으로 연구자 1, 연구자 2, AI 세 집단 간 3자 비교에서는 전체 일치율이 75.0%, Fleiss의 카파 계수는 0.682로 나타났다. 이는 세 집단 모두가 동일한 판정을 내린 경우가 전체의 약 4분의 3을 차지함을 의미하며, AI가 연구자 간 합의(concordance)와도 상당한 수준의 일치를 보였다는 점에서 임상적 의미가 크다.

Table 3. Agreement between researchers and AI in pressure injury staging.

(N=20)

Comparison	Stage	Sensitivity (%)	Specificity (%)	Accuracy (%)	Cohen's Kappa	Fisher's Exact (p)
Researcher 1 vs Researcher 2	Overall	-	-	85.0	0.703	<.001
Researcher 1 vs AI	Stage 2	84.6	100.0	90.0	0.615	<.001
	Stage 3	71.4	84.6	80.0	0.615	<.001
Researcher 2 vs AI	Stage 2	83.3	87.5	85.0	0.730	.004
	Stage 3	83.3	85.7	85.0	0.730	.004
	Stage 4	100.0	100.0	100.0	0.730	.004
Researcher 1 & 2 vs AI (3-way)	Overall	-	-	75.0	0.682 (Fleiss)	

Note. Values are presented as percentages unless otherwise indicated. Cohen's kappa indicates the level of agreement. Fisher's exact test was used for small cell counts. Statistical significance was set at $p < .05$.

3. 간호사 만족도

설문지는 학습 용이성, 인터페이스의 명확성, 평가 소요 시간의 효율성, 바쁜 임상 환경에서의 사용 용이성, 임상적 의사결정 지원, 단계 판정의 일관성과 안정성, 전반적 만족도 및 향후 사용 의향을 평가하였다. 총 50명의 간호사가 참여한 만족도 조사 결과, 애플리케이션의 학습 용이성은 94%(매우 그렇다 48%, 그렇다 46%)로 매우 높은 긍정 평가를 받았다. 화면 구성 및 버튼 배치의 직관성 역시 88%가 긍정적으로 응답하여 사용자 편의성에 대한 만족도가 높게 나타났다.

반면, 애플리케이션 사용으로 욕창 평가 시간이 줄었다고 응답한 비율은 72%(매우 그렇다 38%, 그렇다 34%)였으며, 바쁜 임상 상황에서도 무리 없이 사용할 수 있다는 응답은 60%(매우 그렇다 32%, 그렇다 28%)로 상대적으로

낮게 나타났다. 이는 실제 임상 환경에서의 시간·업무 부담이 애플리케이션 활용에 일정 부분 제한 요인으로 작용할 수 있음을 시사한다.

그럼에도 불구하고, 애플리케이션이 욕창 단계 판정의 일관성을 높여준다는 항목에서 82%(매우 그렇다 54%, 그렇다 28%), 임상 판단에 도움이 되었다는 항목에서 88%(매우 그렇다 48%, 그렇다 40%)가 긍정적으로 응답하였다. 또한 반복 평가에서도 일관적이었다는 항목에서도 82%(매우 그렇다 38%, 그렇다 44%)가 긍정하여 AI 기반 판단의 신뢰성과 안정성이 높게 평가되었다.

전반적 만족도는 86%(매우 그렇다 46%, 그렇다 40%)였으며, 향후에도 애플리케이션을 계속 사용하고 싶다는 응답은 84%(매우 그렇다 48%, 그렇다 36%)로 확인되었다. 이는 애플리케이션이 임상적으로 유용한 도구로 인식되고 있음을 의미한다(Figure 4).

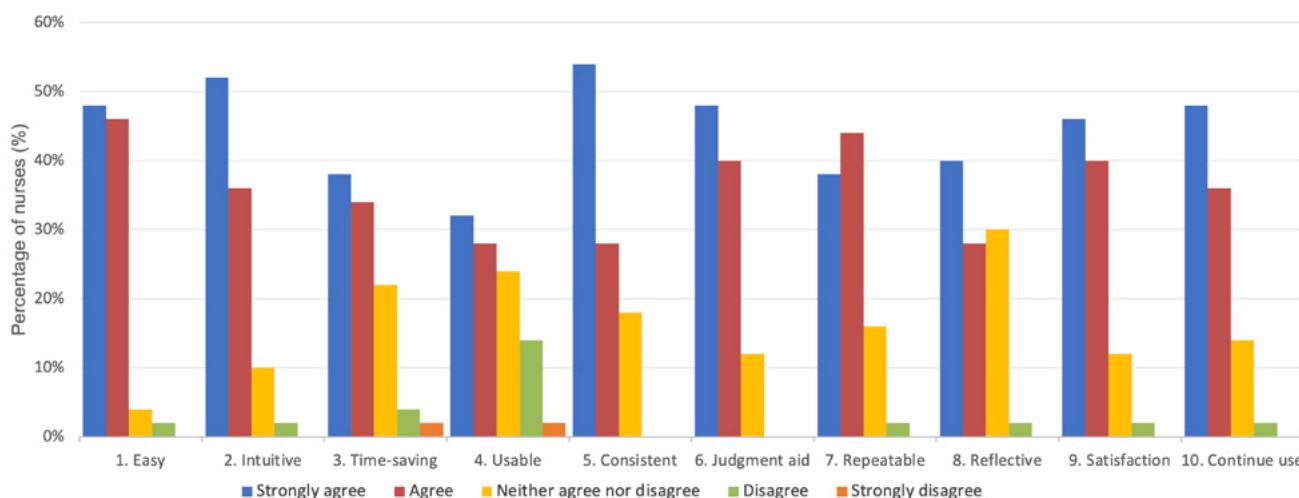


Figure 4. Distribution of nurse satisfaction responses regarding the AI-based pressure injury assessment mobile application (N=50)

note: Nurse usability and perceived clinical utility of the AI based pressure injury application

V. 고찰

이 연구는 병원과 대학 연구팀에서 공동 개발한 AI 기반 욕창 분석 모바일 애플리케이션에 대해 전문가 판정과의 일치도 및 간호사 만족도 및 사용성조사를 통해 임상 적용의 가능성을 탐색한 파일럿 연구이다. 그 결과, AI 기반 분류 시스템은 연구자 평가와 상당한 수준의 일치도를 보였으며, 간호사들은 애플리케이션의 사용성 측면에서 전반적으로 긍정적인 반응을 보였다.

먼저, 연구자 간 카파 계수(0.703)는 기존 문헌에서 보고된 욕창 분류 연구자 간 일치도(중등도-상당 수준)와 유사하여, 이 연구의 연구자 평가가 신뢰도 높은 기준으로 기능했음을 의미한다. 연구자 - AI 간 일치도(Cohen's kappa 0.615-0.730) 역시 중등도에서 상당 수준으로 나타났으며, 이는 딥러닝 기반 모델이 임상 간호사의 판단을 보조할 수 있는 잠재적 가능성을 보여준다. 이러한 결과는 CNN 및 YOLO 기반 욕창 분류 연구에서 중등도 이상의 일치도가 보고된 이전 연구[17]들과 맥락을 같이 한다.

이 연구에서 간호사 50명을 대상으로 수행한 만족도 조사에서는 애플리케이션의 학습 용이성(94%), 화면 구성 직관성(88%)이 매우 높게 나타나, 임상 현장에서 빠르게 적용할 수 있는 도구임을 보여주었다. 이는 간호사가 복잡한 정보시스템보다 직관적 구조의 모바일 기반 평가 도구를 선호한다는 선행 연구[18]와도 일치한다. 또한 임상 판단 보조(88%), 평가 일관성 향상(82%), 반복 평가의 안정성(82%) 등 AI 판단의 신뢰성과 관련된 항목 역시 높은 긍정률을 보여줌으로써, 경험 수준에 따른 평가 편차를 완화하고 표준화된 판단을 보조할 수 있는 도구로 활용될 가능성을 시사한다.

반면, 바쁜 임상 상황에서의 사용 용이성(60%)과 평가 시간 단축(72%) 항목은 상대적으로 낮게 나타났다. 이는 실제 임상 환경에서 시간·업무 부담이 여전히 AI 기반 애플리케이션 활용의 주요 제약 요인임을 의미한다. 이러한 점은 향후 개발 과정에서 자동 촬영 가이드, 신속 분석 모드, UI(User Interface) 단순화 등의 업무 개선이 필요하다. 또한 이 연구에서는 실제 평가 소요 시간 감소나 오류 감소와

같은 질 향상 효과는 직접 측정하지 않았다.

한편, 전체 50명 중 상당수가 '보통'으로 응답하였으나, 이를 긍정적 경향으로 포함하여 수행한 보조 분석에서는 대부분 항목에서 긍정 응답률이 크게 증가하는 양상이 확인되었다. 이는 간호사들이 이 애플리케이션에 대해 전반적으로 우호적인 인식을 가지고 있으며, 초기 적응 기간을 거친 이후에는 실제 활용도가 더욱 증가할 가능성이 있음을 시사한다.

이 연구의 강점은 실제 임상 현장에서 촬영된 욕창 이미지와 전문가 판정, AI 자동 분류 결과, 그리고 간호사 만족도를 통합적으로 분석하였다는 점이다. 이는 단순한 분류 성능 지표를 넘어, 임상 적용 가능성(clinical applicability)과 기술 수용성(technology acceptance)을 함께 탐색했다는 점에서 의미가 있다. 또한 본원 임상 환경에 특화된 딥러닝 모델을 실제 간호 업무 흐름에서 적용하여 평가함으로써, AI 기반 욕창 사정 시스템의 실제 활용 가능성을 제시하였다.

그러나 이 연구는 단일 기관에서 수행된 파일럿 연구로 표본수가 제한적이며, 이미지 촬영 환경(빛 반사, 각도, 피부 톤 등) 및 병변 특성이 완전히 통제되지 못했다는 한계가 있다. 또한 AI 모델이 특정 병원 환경의 데이터를 기반으로 학습된 점은 외부 기관에서의 일반화 가능성을 제한할 수 있다. 또한 이 연구에서는 두 연구자의 판정을 기준으로 활용하였기 때문에, 전문가 판정 자체의 주관성과 변동 가능성이 존재한다는 한계가 있다.

향후 연구에서는 다기관 기반의 대규모 데이터 확보, 다양한 촬영 조건을 반영한 모델 재학습, 중증 환자나 다양한 피부색 환자를 포함한 알고리즘 외부 검증이 필요하다. 더불어 EMR과의 시스템적 연계를 통해 상처 경과 추적과 간호중재 기록을 자동화하는 기능 확장이 이루어진다면 임상 적용성은 더욱 강화될 것으로 기대된다.

한편, 이러한 한계에도 불구하고 이 연구에서 개발된 AI 기반 욕창 단계 분석 애플리케이션은 독립적인 진단 도구가 아닌, 임상 현장에서 간호사의 판단을 보조하는 의사결정 지원 도구로 활용되는 것을 목적으로 한다. 이 애플리케이션은 특히 정기 간호사정 과정에서 신규로 발견된 욕창

의 초기 단계 평가 상황, 욕창 단계 분류에 대한 이해와 숙련도를 향상시키기 위한 교육 목적, 욕창 사정 경험이 상대적으로 부족한 간호사의 임상 판단을 지원하는 상황에서 가장 효과적으로 활용될 수 있다.

이러한 임상 활용 시나리오에서 이 애플리케이션은 실시간 참고 정보를 제공함으로써 간호사의 임상적 판단을 보조하고, 욕창 단계 분류의 일관성을 향상시키는 데 기여할 가능성이 있다. 다만 이 애플리케이션은 전문가의 임상 판단을 대체하기 위한 목적이 아니며, 기존 임상 업무 흐름 내에서 보조적 도구로 제한하여 활용되어야 한다.

VI. 결론

이 연구는 단일 기관에서 수행된 파일럿 연구로, 연구 결과는 예비적(preliminary) 수준으로 해석되어야 한다. AI 기반 욕창 분석 모바일 애플리케이션은 전문가 판정과 중등도에서 상당한 수준의 일치도를 보여 욕창 단계 분류를 위한 임상 의사결정 보조 도구로서의 잠재적 활용 가능성을 시사한다. 또한 간호사를 대상으로 한 사용성 평가에서 전반적으로 긍정적인 반응을 보였다. 이 연구에서는 평가 시간 단축이나 오류 감소와 같은 질 향상 효과는 직접적으로 평가하지 않았다. 향후 다기관 연구를 통해 이 시스템의 임상 적용 가능성을 추가로 검증할 필요가 있다.

VII. 참고문헌

1. Haesler E. Prevention and Treatment of Pressure Ulcers/Injuries: Clinical Practice Guideline: The International Guideline. 3rd ed. Cambridge, Australia: Cambridge Media; 2019.
2. Pan Pacific Pressure Injury Alliance. Prevention and Treatment of Pressure Ulcers/Injuries: Clinical Practice Guideline. 2nd ed. Osborne Park, Australia: Cambridge Media; 2014.
3. Padula WV, Delarmente BA. The national cost of hospital-acquired pressure injuries in the United States. *International Wound Journal*. 2019;16(3):634-40.
4. Hughes RG. Patient Safety and Quality: An Evidence-Based Handbook for Nurses. Rockville, USA: Agency for Healthcare Research and Quality; 2008.
5. Padula WV, Black JM, Davidson PM, Kang SY, Pronovost PJ, Makic MBF, et al. Are evidence-based practices associated with effective prevention of hospital-acquired pressure ulcers in US academic medical centers?. *Medical Care*. 2016;54(5):512-8.
6. Gould LJ, Horn SD, Fife CE, Lesselroth H, Murphy CL, Simman R, et al. WHS guidelines for the treatment of pressure ulcers—2023 update. *Wound Repair and Regeneration*. 2024;32(1):6-33.
7. Chao Q, Zhang X, Liu Y, Wang J, Li H, Chen S, et al. Evaluation methods of pressure injury stages: A systematic review and meta-analysis. *Journal of Tissue Viability*. 2025;34:100894.
8. Haesler E. Prevention and Treatment of Pressure Ulcers: Quick Reference Guide. 2nd ed. Osborne Park, Australia: Cambridge Media; 2014.
9. Edsberg LE, Black JM, Goldberg M, McNichol L, Moore L, Sieggreen M. Revised National Pressure Ulcer Advisory Panel pressure injury staging system. *Journal of Wound, Ostomy and Continence Nursing*. 2016;43(6):585-97.
10. Chang Y, Kim JH, Shin HW, Ha C, Lee SY, Go T. Diagnosis of pressure ulcer stage using on-device AI. *Applied Sciences*. 2024;14(16):7124.
11. Seo S, Kang J, Eom IH, Song H, Park JH, Lee Y, et al. Visual classification of pressure injury stages for nurses: A deep learning model applying modern convolutional neural networks.

- Journal of Advanced Nursing. 2023;79(8):3047-56.
12. Eldridge SM, Chan CL, Campbell MJ, Bond CM, Hopewell S, Thabane L, et al. CONSORT 2010 statement: extension to randomised pilot and feasibility trials. *BMJ*. 2016;355:i5239.
 13. Hertzog MA. Considerations in determining sample size for pilot studies. *Research in Nursing & Health*. 2008;31(2):180-91.
 14. Julious SA. Sample size of 12 per group rule of thumb for a pilot study. *Pharmaceutical Statistics*. 2005;4(4):287-91.
 15. Ke Z, Liu Y, Wang H, Huang Q, Wu Y, Yang J, et al. Improve the satisfaction of medical staff on the use of home nursing mobile app by using a hybrid multi-standard decision model. *BMC Nursing*. 2024;23(1):302.
 16. de Dios López A, Martínez-Pérez B, de la Torre-Díez I, López-Coronado M, de la Torre-Díez J, Martínez-Pérez A, et al. Validation of a questionnaire to assess the usability of and user experience with mobile health applications. *Healthcare*. 2024;12(23):2328.
 17. Kim J, Lee H, Park S, Kim S, Min J, Lee J, et al. Augmented decision-making in wound care: evaluating the clinical utility of a deep-learning model for pressure injury staging. *International Journal of Medical Informatics*. 2023;180:105266.
 18. Nowicki GJ, Ślusarska B, Bartoszek A, Kocki B, Deluga A, Kanczuga M, et al. Development and pre-evaluation of a “DiagNurse” mobile app to support nurses in clinical diagnosis using the ADDIE model. *Scientific Reports*. 2024;14(1):29765.

Supplementary 1. [욕창 애플리케이션 임상 적용 후 간호사 만족도] 설문 조사

[욕창 애플리케이션 임상 적용 후 간호사 만족도]

설문 조사

대전을지대학교병원 QI부

본 설문은 욕창 단계 평가 애플리케이션을 실제 임상 현장에서 사용한 간호사들의 경험을 바탕으로, 사용 편의성 · 효과성 · 임상적 활용 가능성에 대한 만족도를 조사하기 위해 마련되었습니다.

응답해주신 내용은 연구 목적으로만 활용되며, 모든 자료는 익명으로 처리되어 개인이 식별되지 않습니다. 설문 응답에는 약 5분 정도 소요되며, 귀하의 소중한 의견은 앱의 개선 및 임상 적용 확대에 큰 도움이 될 것입니다.

바쁘신 와중에도 연구에 협조해 주셔서 감사드립니다.

1. 응답자 기본정보

- 1) 근무부서 ○ 병동 ○ 중환자실 ○ 응급실
- 2) 직종/직위 ○ 간호사 ○ 주임 ○ 파트장
- 3) 임상경력 ○ 1년 이내 ○ 1년~3년 ○ 4~6년 ○ 7~10년 ○ 10년 이상
- 4) 욕창 평가 경험 ○ 적음(주1회 미만) ○ 보통(주 5회 미만) ○ 많음(주 5회 이상)
- 5) 앱 사용 빈도(1주) ○ 1~2회 ○ 3~5회 ○ 6~10회 ○ 10회 이상

2. 앱 사용 만족도(5점 척도)

(1=전혀 그렇지 않다, 2=그렇지 않다, 3=보통이다, 4=그렇다, 5=매우 그렇다)로 응답해 주십시오.

번호	내용	전혀 그렇지 않다	그렇지 않다	보통 이다	그렇다	매우 그렇다
1	앱의 전반적 사용법을 배우기 쉬웠다.	①	②	③	④	⑤
2	화면 구성과 버튼 배치가 직관적이었다.	①	②	③	④	⑤
3	앱 사용으로 욕창 평가 시간이 줄었다	①	②	③	④	⑤
4	바쁜 상황에서도 앱 사용에 무리가 없었다.	①	②	③	④	⑤
5	앱이 욕창단계 판정의 일관성을 높여주었다.	①	②	③	④	⑤
6	앱 결과가 임상 판단에 도움이 되었다.	①	②	③	④	⑤
7	앱의 판정 결과가 반복 평가에서도 일관적이었다.	①	②	③	④	⑤
8	앱의 단계 판정이 실제 상처 상태를 잘 반영했다.	①	②	③	④	⑤
9	전반적으로 이 앱에 만족한다	①	②	③	④	⑤
10	향후에도 이 앱을 계속 사용하고 싶다	①	②	③	④	⑤

3. 개방형 질문

- 1) 앱 사용 시 가장 만족스러웠던 점은 무엇이었나요?
- 2) 개선이 가장 시급하다고 생각하는 부분은 무엇인가요?