WoundView: A Multimodal AI Tool for Real-Time, Remote, and Cost-Effective Wound Risk Assessment

Yash Bhavsar[1], **Ivan Felipe Rodriguez**[2]

Received October 19, 2025
Accepted June 4, 2026
Electronic access July 15, 2026

Wound care management is a major medical challenge, causing a significant clinical, social, and economic burden on global healthcare systems. To address this, we developed *WoundView*, an AI-powered app for real-time wound assessment. *WoundView* utilizes wound images to provide a risk score and generates tailored treatment plans using advanced Large Language Models (LLMs). This system was developed using three distinct, publicly available datasets consisting of more than 4,000 images. For wound classification, a convolutional neural network (CNN) model (ResNet18) was developed, classifying a total of 18 wound types with a macro-averaged F1-score of 0.776 and a weighted F1-score of 0.925. To enable accurate wound segmentation, a SAM2.1-Large model was fine-tuned to precisely identify the region of interest, achieving a prompt-free validation IoU of 0.72 compared to a U-Net baseline of 0.74. This segmentation was used to analyze the wound's color composition, a key factor in the diagnostic process. We developed a novel scoring system, the Wound Risk Assessment Score (WRAS), utilizing wound characteristics, patient demographics, and comorbidities to prompt an LLM to generate a wound risk severity score and tailored treatment plan. To assess performance, six OpenAI models were compared for 10 cases. Five physicians from various disciplines, including primary care, podiatry, vascular medicine, and infectious disease, evaluated the responses and identified models using WRAS as being more predictive of wound severity. *WoundView* demonstrates the potential of AI to support wound care through accessible, cost-effective assessments, though prospective clinical validation on real-world data is required before deployment.

Introduction

Chronic wounds affect approximately 7-10 million individuals in the United States annually[1]. Wounds are typically categorized as either acute or chronic. Acute wounds heal in a predictable and timely manner, whereas chronic wounds fail to progress through the normal healing process[2]. As the population ages, the prevalence of chronic wounds continues to rise, largely due to comorbidities such as diabetes and venous insufficiency. Additionally, chronic wounds make up a large proportion of all wounds, some examples including pressure injuries, venous ulcers and ischemic wounds[3]. On the other hand, injuries such as abrasions, bruises, burns, cuts, and lacerations are categorized as acute wounds.

Multiple physiological and environmental factors affect the healing process of wounds. Risk factors such as age, diabetes, obesity, peripheral arterial disease and smoking play a major role in the development and progression of wounds[4]. Without timely identification and intervention, acute wounds may transition to chronic wounds. Early detection and proper treatment are critical, as untreated wounds can cause serious complications such as infections and amputations leading to increased risks of disability and possibly death[5].

Wound care management presents major challenges due to lack of timely clinical intervention. The nationwide shortage of wound care specialists further exacerbates this problem[6]. In some instances, patients can experience referral delays greater than 3 months before seeing a specialist, meaning only the most severe cases typically reach advanced care[7]. In addition to these delays, patients face logistical barriers, including pain-related immobility, transportation difficulties, long wait times, and high costs of care. These factors discourage many from seeking early medical care. In addition, due to limited access to specialized wound care centers, primary care physicians and nurses often serve as the initial point of assessment and ongoing management, particularly in rural communities and outpatient settings[8]. Although initial wound care is typically handled by general practitioners, the escalating needs of an aging population make it increasingly difficult to provide efficient and thorough wound care management[9]. This gap in care has significant financial consequences. Medicare estimates the annual cost of wound treatment in the U.S. to be $\sim$ \$22.5 billion with outpatient services accounting for the majority of expenditures[1]. Beyond financial strain, chronic wounds inflict social and psychological stress on patients and their family members, especially in underserved populations who have fewer options for high-quality treatment[10]. As a result, there is an urgent need for a robust, scalable assessment

[1] *Moorestown High School*
[2] *Brown University*

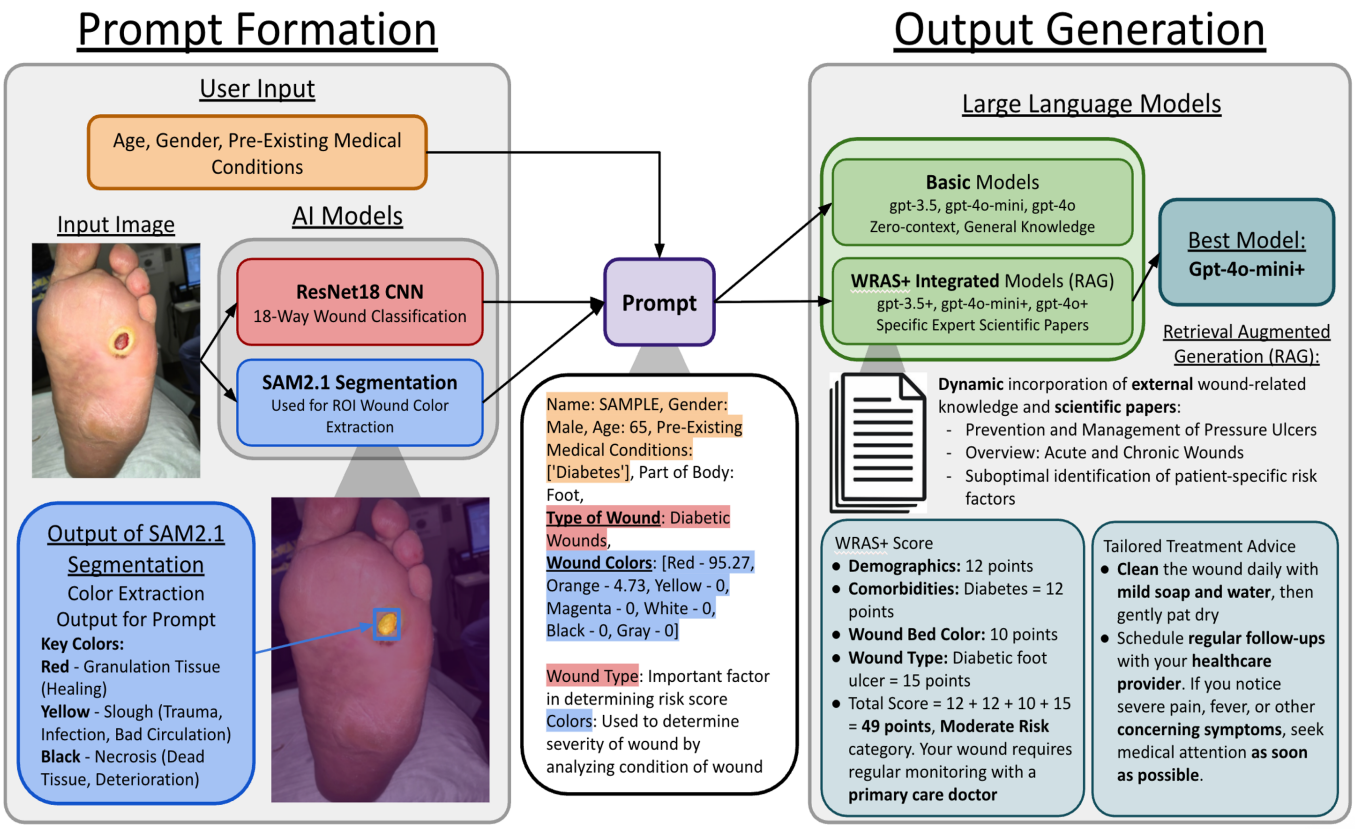

Figure 1: *WoundView* flowchart and pipeline.
This figure shows the pipeline and steps of the *WoundView* app. It starts with user input and the input image; this information is then used to generate a prompt that includes key information like wound type from the ResNet18 CNN and colors of the wound bed from SAM2.1 segmentation model.

tool that can accurately predict wound severity and provide timely treatment recommendations.

The objective of this study was to develop a cost-effective, user-friendly, and medically reliable wound assessment tool that can provide rapid results on a smartphone or web-based platform in minutes. This assessment tool can also be used in outpatient settings by nurses and residents who may not be experienced in treating all types of wounds. The goal is not only to empower patients with clear, personalized advice but also support clinicians with consistent, data-driven recommendations that enhance clinical decision-making.

To address this need, we introduce *WoundView*, an AI-powered system that provides quick, automated and remote wound risk assessment from a single image input. Following deep learning parameters were meticulously tested to improve the accuracy and diagnosis of wound characterization. *WoundView* combines a ResNet18-based CNN for wound classification, a fine-tuned SAM2.1 segmentation model to isolate the region of interest, and a Wound Risk Assessment Score (WRAS) to predict severity of the wound. WRAS is a scoring system (1-100 scale) that integrates patient demographics, comorbidities, wound type, and color-based analysis in the HSV space to generate an interpretable wound severity score. These risk scores were utilized to train LLMs to generate treatment recommendations. Six different LLMs were tested for their performance across each case (gpt3.5, gpt4o-mini, gpt4o, gpt3.5+, gpt4o-mini+, gpt4o+). Output from these LLMs were evaluated by physicians across 120 assessments using a 5-point Likert scale to ensure the model's performance, accuracy, clinical relevance and reliability.

It is important to note that our comprehensive approach to wound diagnosis and treatment recommendation is unique compared to current existing literature. Other groups have used CNN and segmentation in wound detection by comparing different architectures and incorporating wound image and location as well as tracing a wound area to create accurate assessment[11,12]. However, to our knowledge, no existing system combines multi-class wound type classification,

Table 1 Literature comparison

Studies	Method	Dataset	Wound Types	Clinical Validation	Treatment Plans
Howell et al. (2021)[11]	AI validation framework for wound area and granulation tissue	199 photos, 2 centers	11 types	Yes, 4 specialists + 6 reviewers	No
Patel et al. (2024)[12]	Multi-modal CNN + body location (VGG16, ResNet152, EfficientNet-B2)	946 images	4 types	No	No
WoundView	ResNet18 + SAM2.1 + WRAS + RAG-LLMs	4,338 images (3 datasets)	18 types	Yes, 5 physicians, 120 assessments ($\rho = 0.611$, $p < 0.01$)	Personalized via RAG-LLMs

region-of-interest segmentation, color-based tissue analysis, and LLM-generated treatment recommendations into a single patient-facing tool. *WoundView* integrates all of these components alongside a novel, patient-interpretable severity score (WRAS) applicable across 18 wound types.

Recent AI advances have begun addressing these limitations. Howell et al. (2021) developed a validation methodology for AI-based wound assessment tools using 199 photographs from two wound centers[11]. Comparing AI wound area and granulation tissue measurements against expert human annotations, they found AI performed statistically similarly to human specialists. However, their study focused on validation frameworks rather than comprehensive assessment systems, lacking integration of patient demographics, comorbidities, standardized severity scoring, or treatment recommendations.

Patel et al. (2024) introduced a multi-modal network integrating wound images with anatomical location data for classifying four wound types: diabetic, pressure, surgical, and venous ulcers[12]. Their approach achieved 74.79–100% accuracy across various classification tasks, demonstrating that location data significantly improved performance. Despite these advances, the system focused solely on classification without severity assessment, risk stratification, or treatment planning, and evaluation relied on computational metrics without physician validation or real-world deployment.

Both studies reveal several limitations in current AI wound assessment systems. Dataset sizes remain limited (199 and 946 images for Howell and Patel, respectively), potentially under-representing wound diversity in clinical practice[11,12]. While classification has advanced, no standardized severity assessment tool exists that works across wound categories. Existing scales are wound-specific and not generalizable. No prior work has developed unified severity scoring applicable to both acute and chronic wounds incorporating demographics, comorbidities, and wound characteristics to quantify wound severity.

Methods

Datasets

An extensive search was conducted to identify publicly available wound image datasets suitable for model development and evaluation. A total of six datasets were initially considered; however, only three were ultimately selected. The remaining datasets were excluded due to their specialized focuses, which did not align with the broader wound categories targeted in this study. Additionally, several datasets lacked comprehensive annotations or exhibited limited variability in wound types, further limiting their utility. The final three datasets were chosen based on relevance, image quality, and the presence of detailed labels necessary for training deep learning models.

All datasets used in this study were publicly available and previously published with appropriate permissions. Dataset #1 (Fateen, 2023) is available under open license on Kaggle and contains de-identified wound images with no protected health information (PHI). Dataset #2 (Taher, 2021) is available under open license on Kaggle with de-identified images and segmentation masks. Dataset #3 (Kręcichwost et al., 2021) was published in Computerized Medical Imaging and Graphics and contains de-identified clinical images with demographic metadata (age and gender only) and no personally identifiable information.

Dataset #1 contained wound images of the following types: abrasions, bruises, burns, cuts, diabetic foot, lacerations, normal, pressure wounds, surgical wounds, and venous

Table 2 Overview of three selected datasets

	Size	Resolution	Purpose
Dataset #1	2940 images	640×640 (normalized)	**Classification** (Abrasions, Bruises, Burns, Cut, Diabetic Foot, Laceration, Normal, Pressure Wounds, Surgical Wounds, Venous Wounds)
Dataset #2	1210 images	512×512	**Segmentation** (Image + ROI/Label)
Dataset #3	47 cases/patients 188 images	320×240	**Classification** (Amputation, Arteriovenous Fistula Failure, Diabetic Foot, Ischemia, Lymphedema, Mastectomy, Venous Insufficiency, Venous Wounds, Wound Infection, Injury) **Segmentation** (Image + ROI/Label) **Demographics** (Age & Gender)

wounds[13]. Images in this dataset had varying resolutions and were normalized to 640×640 pixels for training. Dataset #2 contained data in the format of a wound image and the corresponding annotation for ROI[14]. All images in this dataset were standardized at 512×512-pixel resolution. In Dataset #1 and #2, the annotation protocols were not specified in the original source, nor were the demographic information about the patients.

Dataset #3 contained data grouped by each patient for classification (amputation, arteriovenous fistula failure, diabetic foot, ischemia, lymphedema, mastectomy, venous insufficiency, venous wounds, wound infection, and injury), segmentation (image + ROI/label), and demographics (age and gender)[15]. All images were captured at 320×240-pixel resolution. Patient age ranged from 33 to 90 years with a mean of 68.5 years, and the cohort included 34 female and 13 male patients. All wound segmentation masks in this dataset were manually delineated by an experienced surgeon who was given unlimited time to carefully outline the wound boundaries on high-quality photographs. These expert annotations were then converted to binary masks for use in model training and evaluation.

Dataset #1 and Dataset #3 shared two overlapping wound categories, Diabetic Foot and Venous Wounds; each was treated as a single class in the combined label set, yielding 18 distinct wound classes from the two 10-label datasets.

Our datasets exhibited significant class imbalance. Dataset #1 showed moderate imbalance with the largest class (Pressure Wounds, n = 602) being 6 times larger than the smallest (Cut, n = 100). Dataset #3 demonstrated severe imbalance, with Venous Wounds (n = 74) being 37 times more represented than the smallest classes (Mastectomy, Lymphedema, and Arteriovenous fistula failure, each n = 2). A detailed visualization of the training and testing data distributions across all three datasets is provided in Appendix B (Fig. 14).

Data was split 80% training / 20% testing using dataset-specific strategies to prevent data leakage. For Dataset #1, a stratified split (StratifiedShuffleSplit, scikit-learn) was used to ensure each of the 18 wound classes was proportionally represented in both train and test sets. For Dataset #3, a patient-level group split (GroupShuffleSplit) was applied using patient ID (case_N) as the group key, guaranteeing that all images from a given patient appear exclusively in either the training or test set, never both. This prevents any patient-level data leakage that would otherwise inflate performance estimates. The same random seed (42) was used across all splits for reproducibility.

Model Architecture

CNN Classifier

In this study, CNNs were employed to classify a diverse set of 18 wound types. CNNs are a class of deep learning models that are highly effective for image analysis. They operate by applying a series of learnable filters, called kernels, to input images in order to detect patterns such as edges, textures, and shapes[16]. These filters move across the image spatially, producing feature maps that highlight important visual characteristics. As the network deepens, each layer builds upon the previous one, capturing increasingly complex features. Pooling layers reduce the spatial dimensions, which helps to generalize patterns and decrease computational load. Activation functions introduce non-linearity, allowing the model to learn complex relationships[17]. Fully connected layers at the end of the network combine the extracted features and output a classification decision.

To select the optimal classification architecture, four CNN models (ResNet18, ResNet34, ResNet50, and EfficientNet-B0) were evaluated under identical conditions: the same data split, AdamW optimizer with OneCycleLR scheduler, 20 training epochs, and the same image preprocessing pipeline. Performance across these architectures is summarized in Ta-

Table 3 Architecture Comparison on Combined (DS1 + DS3) Test Set

Architecture	Parameters	Accuracy	Macro F1	Macro Precision	Macro Recall	Weighted F1
ResNet18	∼11M	0.922	0.776	0.767	0.838	0.925
ResNet34	∼21M	0.916	0.615	0.610	0.633	0.921
ResNet50	∼25M	0.924	0.710	0.695	0.749	0.933
EfficientNet-B0	∼5M	0.937	0.685	0.674	0.742	0.939

ble 3.

Pairwise McNemar's tests (with continuity correction) were performed to assess whether performance differences were statistically significant. Of the six possible pairs, only ResNet34 vs. EfficientNet-B0 was significant (x^2 = 4.65, p = 0.031). No other pair reached significance (all p > 0.17), indicating that the four architectures perform comparably on this dataset. ResNet18 was selected as the final architecture based on its highest macro-averaged F1 score (0.776). In the presence of significant class imbalance, macro F1 is a more informative metric than accuracy, as it gives equal weight to all classes regardless of support. ResNet18 also offered a favorable complexity-performance trade-off (∼11M parameters).

The implemented ResNet18 model begins with an initial 7x7 convolutional layer (112×112×64) and a MaxPool layer (56×56×64), followed by four stages of residual blocks (Stages 1–4). Each stage progressively doubles the feature map depth while halving the spatial resolution: Stage 1 (56×56×64), Stage 2 (28×28×128), Stage 3 (14×14×256), and Stage 4 (7x7×512) (Fig. 2). The residual skip connections facilitate efficient gradient flow and robust feature extraction. To adapt the model to our wound classification task, the standard head was replaced with a Global Average Pooling layer and a fully connected layer modified to output 18 distinct classes [18].

Segmentation

To develop a robust wound segmentation pipeline, two distinct architectural approaches were evaluated: the SAM2.1 foundation model and a specialized U-Net baseline. The objective was to benchmark the automated performance of a state-of-the-art vision transformer against an established medical imaging architecture. While both models were trained on the same data splits, SAM2.1 was ultimately selected as the primary segmentation engine due to its superior generalization capabilities and its support for interactive, clinician-led mask refinement.

SAM2.1 is a state-of-the-art vision transformer-based model developed for image segmentation [19]. Unlike traditional segmentation models that need to be trained specifically for each task, SAM2.1 is designed to generalize across a wide range of image domains with minimal fine-tuning. It takes an image and returns a high-quality object mask that outlines

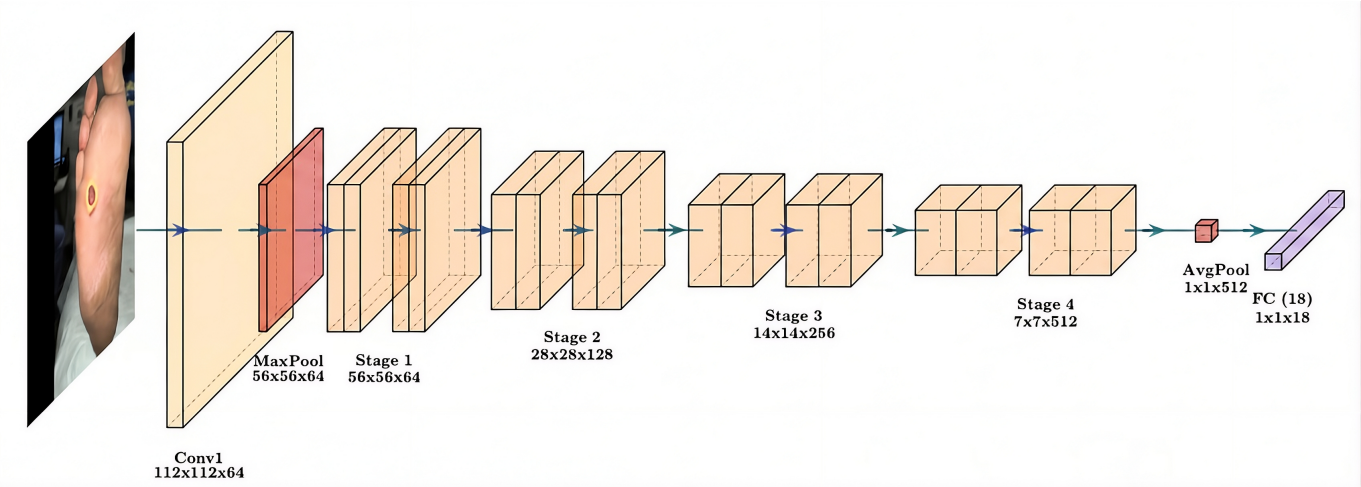

Figure 2: ResNet18 Architecture.
Diagram illustrating the flow of data through the four residual stages, highlighting the spatial dimensions and feature map depths at each transition point.

the region of interest. Internally, SAM2.1 uses a Hiera image encoder, a hierarchical vision transformer, to process the image and capture global context by attending to all parts of the image[20]. These components work together to accurately segment even complex and irregular shapes, such as wounds. For our application, SAM2.1 enabled flexible and precise segmentation, just requiring fine-tuning on our dataset. Its modular design, fast inference time, and effectiveness across varied inputs made it a strong choice for our segmentation pipeline[21].

For fine-tuning, the SAM2.1-Large variant was used. The image encoder was frozen throughout training; only the mask decoder and prompt encoder were updated (∼4M trainable parameters). During training, a single random positive point prompt was sampled uniformly from within the ground-truth mask for each image. At inference, the model was evaluated without point prompts, using only the automatic mask generation pathway. A U-Net baseline with a pretrained ResNet34 encoder was also implemented for comparison, trained for 50 epochs using a combined Dice and binary cross-entropy loss.

Model Training

The ResNet18 CNN was trained with the AdamW optimizer (learning rate 1×10^{-3}, weight decay 1×10^{-2}) and a OneCycleLR scheduler. Images were resized to 224×224 pixels. SAM2.1 was fine-tuned over 2500 steps with a learning rate of 0.0001 and weight decay of 0.01. Learning rate defines how quickly the model adjusts its weights to minimize validation errors, while weight decay is a form of regularization that reduces weight size to prevent overfitting. Intersection over Union (IoU) was calculated every 100 steps to assess precision between the ground truth and predicted masks. As shown in Fig. 3, a greater IoU corresponds with better model performance.

$$IoU = \frac{\text{Area of Overlap}}{\text{Area of Union (Total Area)}}$$

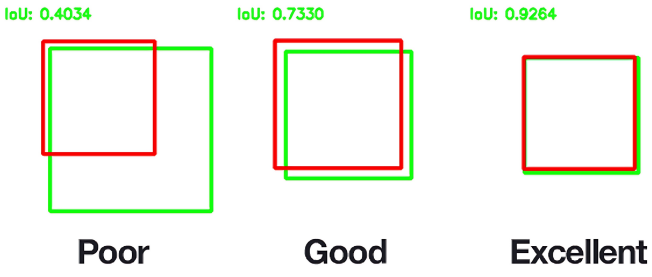

Figure 3: Visual references for poor, good, and excellent IoU metrics.

Color Extraction & Calculation

The color of the wound bed is a critical indicator of healing progression. Predominantly red tissue suggests granulation tissue, indicating that the wound is healing properly. A high percentage of yellow tissue reflects the presence of slough, showing that the wound has a delay in healing and may have an infection. Lastly, black tissue is a sign of necrosis, meaning that the tissue is dying and intervention is crucial[22]. The most effective way to analyze wound bed color is by using the predicted segmentation mask to isolate the region of interest. Once the wound area was identified from a picture, the mask was overlaid onto the original image to extract the relevant pixels. Each pixel within this region was then classified using the HSV (Hue, Saturation, Value) color space. Colors commonly observed in wounds were prioritized, including red, orange, yellow, black, and white. HSV was chosen instead of the RGB color space because its circular hue representation makes it easier to calculate angular distance between colors. It also separates chromaticity from brightness, which is especially important since the lighting in user-uploaded images can vary[23].

WRAS

One of the aims of this study was to develop an accurate WRAS scoring system that can be applied to all types of wounds, both acute and chronic. Several wound assessment instruments are currently in clinical use. Pressure-injury-specific scales such as Norton, Braden, and Waterlow predict risk of pressure ulcer development[24]. Broader wound assessment tools, including the Pressure Ulcer Scale for Healing (PUSH), the Bates-Jensen Wound Assessment Tool (BWAT), and the Pressure Sore Status Tool (PSST), score wound status across multiple tissue characteristics. However, these tools share common limitations: they require trained clinicians for administration, were designed primarily for specific wound subtypes (particularly pressure injuries and venous ulcers), produce scores that are not intuitively interpretable by patients without clinical training, and do not generate automated treatment recommendations. Hence, there was a need for a standardized and easily interpretable wound scoring assessment that can score all different types of wounds. Using existing similar guidelines based on scientific and clinical wound scales, we constructed the novel WRAS scoring system. It includes comorbidities, demographics, and wound characteristics, resulting in a wound severity score ranging from 1 to 100, which is easy for patients to interpret. It is designed to work across a wide variety of wound types, making it a more versatile and accessible tool for both patients and healthcare providers. As a novel wound scoring system, WRAS supports remote evaluations, which makes it particularly valuable in settings with limited access to wound specialists and patients

in underserved communities. A simple score interpretation guides the patient to accurate diagnosis and how they should seek intervention from a medical professional. This app allows for broader implementation in telemedicine and mobile health applications.

$$\text{WRAS} = W_{\text{type}} \times S_{\text{type}} + W_{\text{color}} \times S_{\text{color}} + W_{\text{demo}} \times S_{\text{demo}} + W_{\text{comorbidity}} \times S_{\text{comorbidity}}$$

(Equation 1. WRAS Score Calculation)

The WRAS score is computed as a weighted sum of four components: where S_{type} is the severity contribution of the classified wound type (e.g., necrotic wounds receive a higher base score than abrasions), S_{color} reflects wound bed composition derived from the HSV color analysis (higher proportions of black or yellow tissue increase the score), S_{demo} incorporates patient age and sex, and $S_{\text{comorbidity}}$ applies additive penalties for conditions such as diabetes, peripheral vascular disease, and immunosuppression. Each component is normalized to a 0–100 range and combined with empirically derived weights informed by existing clinical guidelines. The resulting score maps directly to the four risk tiers in Table 4.

Table 4 Interpretation of WRAS scores. The WRAS score ranges from 1-100 in order to provide easy interpretability to the patient. Additionally, the ranges for the score go from low, medium, high, and critical risk corresponding with limited doctor intervention, primary care visits, wound care specialist consultations, and immediate ER visits, respectively.

Score (Range)	Description
Low Risk (1–20)	Likely to heal with minimal intervention.
Moderate Risk (21–50)	Requires regular monitoring and active wound care.
High Risk (51–75)	Advanced intervention required.
Critical Risk (76–100)	ER immediately.

LLM

LLMs are deep learning models trained on vast amounts of text data to understand and generate human-like language. They work by predicting the next word in a sentence based on patterns they've learned[25]. This helps them do tasks like answering questions, writing summaries, or creating organized text. We chose to test six different OpenAI ChatGPT LLMs. The basic models (3.5, 4o-mini, and 4o) were compared with the advanced WRAS-integrated models (3.5+, 4o-mini+, and 4o+) in wound identification, WRAS score, and tailored treatment plan. The basic models represent a baseline understanding of wound care using only the knowledge embedded in their pre-trained parameters. In contrast, the WRAS-integrated models used a Retrieval-Augmented Generation (RAG) framework, which allowed dynamic incorporation of external wound-related knowledge included in the WRAS scoring system to enhance response accuracy and reliability. Each model received a structured text prompt describing the patient's demographics and wound characteristics, such as wound type and wound colors. These characteristics were determined from the ResNet18 CNN and SAM2.1 segmentation models. Based on this input, the model produced two outputs: a WRAS score and a tailored treatment plan, offering guided, interpretable instructions for the patient to follow.

Model Validation

In order to validate the LLM responses, certified physicians were asked to rate the responses of each model on a one-to-five-point Likert scale. Physicians were provided with the image of the wound and a ground truth prompt that included the patient's demographics, type of wound, correct color classification metrics, WRAS score, and associated treatment instructions. For our purposes, a total of 10 cases encompassing a wide variety of patients and wound types were tested. Each of the five physicians was assigned to four cases, making sure that there were two physicians assigned to each case to test reliability. For each case, the physicians rated each ChatGPT LLM response. In other words, 24 ratings were obtained from each physician, culminating in a total of 120 responses. This approach ensured that evaluations were based on clinically relevant information and allowed for consistent assessment across all models. The use of a standardized prompt helped reduce variability in interpretation and focused the evaluation on the accuracy, clarity, and medical appropriateness of each response.

Results

The performance of *WoundView* was evaluated across three core components: wound classification using the ResNet18 CNN, wound segmentation with the SAM2.1 model, and language-based risk assessment and treatment generation via Large Language Models (LLMs). ResNet18 CNN was tested using Datasets #1 and #3, while SAM2.1 was tested using Datasets #2 and #3. Both models were validated through various quantitative metrics. The CNN demonstrated strong discriminative power in classifying 18 wound types, similarly SAM2.1 achieved precise segmentation of wound regions, enabling reliable color-based tissue analysis. These outputs were subsequently integrated into the Wound Risk Assessment Score (WRAS), which informed LLM-driven recommendations. Together, these results highlight the accuracy, reliabil-

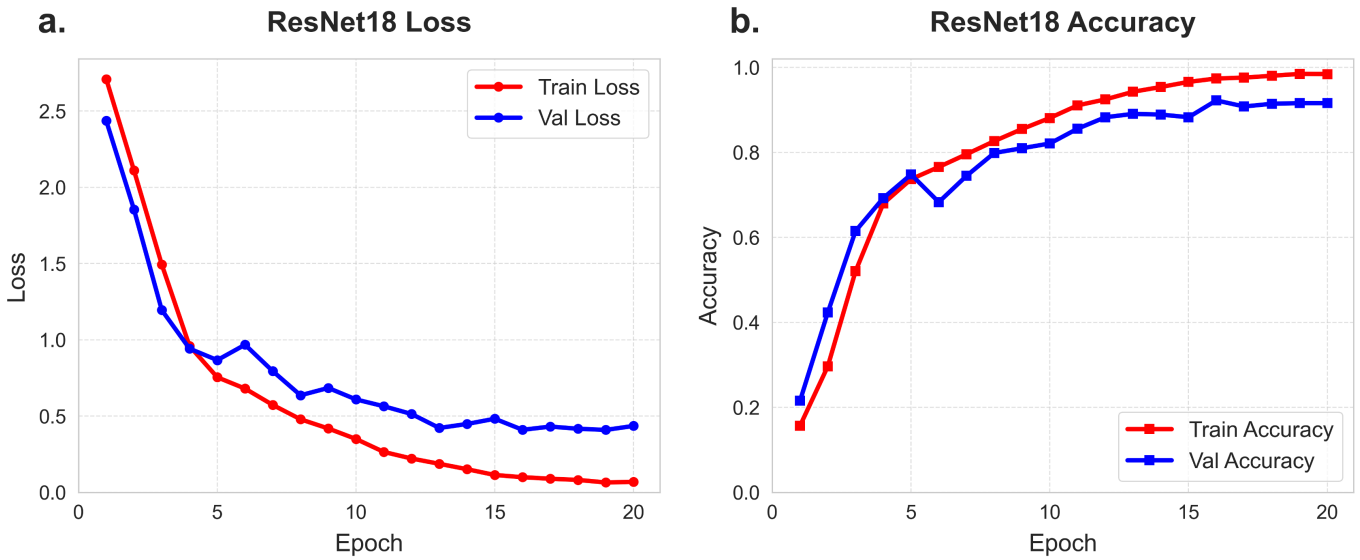

Figure 4: ResNet18 CNN cross-entropy loss (a) and classification accuracy (b) curves.
The model demonstrates strong convergence, with validation accuracy stabilizing at 92.2% accuracy evaluated over 20 epochs.

ity, and clinical interpretability of *WoundView*'s multi-model framework.

ResNet18

The learning and loss curves of the ResNet18 CNN demonstrate strong model performance and effective convergence. As shown in Fig. 4b, validation accuracy increased rapidly during the initial training epochs, reaching a peak of 92.2% and stabilizing thereafter, indicating that the model successfully learned discriminative features from the combined dataset. The cross-entropy loss graph further supports this trend, with both training and validation loss decreasing sharply and approaching zero (Fig. 4a). The minimal gap between the training and validation loss curves suggests that the ResNet18 architecture achieved high performance while avoiding significant overfitting.

Confusion matrices provide key insights into the model's performance across the 18 categories. In our study, the confusion matrix was normalized per wound type to ensure fair comparisons across categories, despite the significant imbalance between Dataset #1 (∼3,000 images) and Dataset #3 (∼200 images). As shown in Fig. 5, the model achieved a near-perfect diagonal for the well-represented classes sourced from Dataset #1, while the Dataset #3-exclusive classes (e.g., lymphedema and ischemia) showed weaker performance, as their extremely small sample sizes limited the model's ability to learn distinct features. However, categories within Dataset #3 that had higher representation still demonstrated high sensitivity, aligning with the strong performance seen across the broader dataset.

The ResNet18 classifier achieved an overall accuracy of 92.2% on the combined held-out test set (DS1 + DS3). To account for the class imbalance, macro-averaged metrics are reported: the model reached a macro F1 of 0.776 (precision: 0.767, recall: 0.838) and a weighted F1 of 0.925.

Detailed per-class performance is presented in Table 5. The ten wound classes sourced from Dataset #1 (Abrasions through Venous Wounds) achieved uniformly high F1 scores (0.90–1.00). In contrast, Dataset #3-exclusive classes showed higher variability. Five classes with extremely limited test support (2–9 samples) received zero F1 scores, while those with marginal representation (e.g., Ischemia and Venous Insufficiency) achieved moderate scores (0.21–0.55). These results highlight that performance gaps are primarily a function of data scarcity in rare chronic categories rather than architectural failure.

To provide a more robust performance estimate for chronic wounds and address potential data leakage, a 5-fold patient-level cross-validation was performed on Dataset #3 using GroupKFold. This ensured that no patient's images appeared in both the training and validation sets within any fold. Results across four architectures are summarized in Table 6.

The cross-validation macro F1 scores (ranging from 0.108 to 0.239) reflect the genuine difficulty of 10-class classification on a dataset of 188 images with extreme imbalance. All architectures showed high fold-to-fold variability, largely driven by the specific patient cohorts assigned to validation in each split. Every model predominantly predicted the majority classes, highlighting the need for substantially more annotated

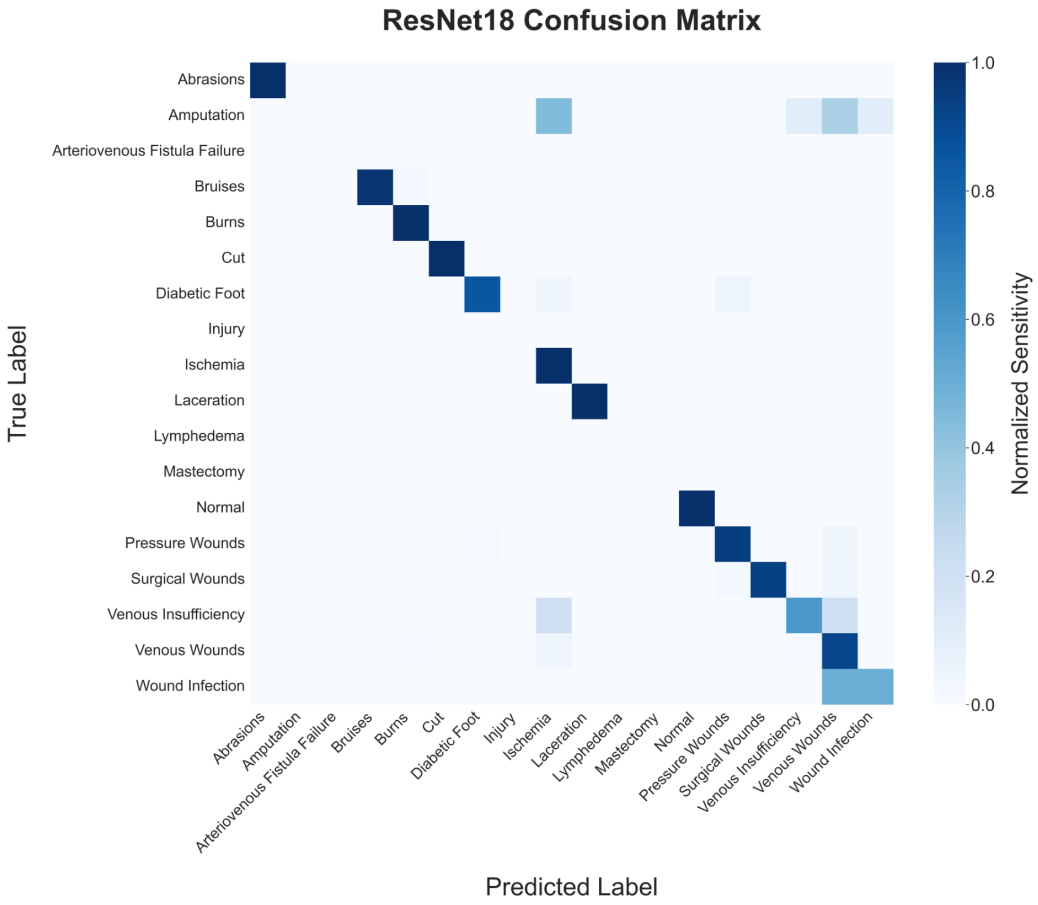

Figure 5: Normalized Confusion Matrix for the ResNet18 CNN.
The matrix represents the 18-way classification performance on the combined testing sets of Datasets #1 and #3. Diagonal elements represent normalized sensitivity (recall) for each wound type.

data for rare chronic wound types-a challenge shared across all evaluated architectures.

Segment Anything Model 2.1 (SAM2.1)

The fine-tuned SAM2.1 model demonstrated a steady improvement in segmentation accuracy throughout the training process. As illustrated in Fig. 6b, the prompt-free validation IoU reached a peak of 0.7151 and remained stable thereafter, suggesting efficient convergence. Training was terminated at 2,500 steps to minimize the risk of overfitting. Both training and validation loss curves (measured via Dice loss) exhibited a consistent downward trend with expected fluctuations (Fig. 6a). Notably, the validation loss closely tracked the training loss, indicating that the frozen Hiera encoder provided robust multiscale features that enabled strong generalization to unseen wound images.

To contextualize performance, we benchmarked SAM2.1 against a supervised U-Net baseline utilizing a pretrained ResNet34 encoder. Quantitative results are summarized in Table 7.

The U-Net baseline marginally outperformed prompt-free SAM2.1-Large by 2.1 percentage points (IoU: 0.7364 vs. 0.7151). However, SAM2.1 achieved this competitive performance with significantly fewer trainable parameters ($\sim$4M vs. $\sim$32M) while keeping the high-capacity image encoder frozen. For clinical deployment, SAM2.1 offers superior utility: it provides a highly accurate automated mask while natively supporting real-time, interactive refinement by a clinician, a “human-in-the-loop" capability that static architectures like U-Net lack.

The SAM2.1 model was then tested on wounds from validation split of Dataset #2 of varying sizes, locations, and wound bed colors (Fig. 7). The three examples validate SAM2.1 performance, showcasing its ability to differentiate the wound bed from the surrounding area. The first two examples demon-

Table 5 Per-Class F1 Score, ResNet18 (Combined Test Set).

Class	F1	Precision	Recall	Test Samples
Abrasions	0.985	0.971	1.000	33
Bruises	0.979	0.979	0.979	48
Burns	0.931	0.871	1.000	27
Cut	1.000	1.000	1.000	20
Diabetic Foot	0.911	0.976	0.854	96
Laceration	1.000	1.000	1.000	24
Normal	1.000	1.000	1.000	40
Pressure Wounds	0.942	0.942	0.942	121
Surgical Wounds	0.963	0.988	0.940	84
Venous Wounds	0.902	0.887	0.917	120
Ischemia	0.211	0.118	1.000	2
Venous Insufficiency	0.545	0.500	0.600	5
Wound Infection	0.500	0.500	0.500	2
Amputation	0.000	0.000	0.000	9
Arteriovenous Fistula Failure	0.000	0.000	0.000	2
Injury	0.000	0.000	0.000	7
Lymphedema	0.000	0.000	0.000	2
Mastectomy	0.000	0.000	0.000	2

Table 6 5-Fold Cross-Validation on Dataset #3 (Patient-Level GroupKFold)

Architecture	Accuracy	Macro F1	Macro Precision	Macro Recall	Weighted F1
ResNet18	0.441 ± 0.089	0.211 ± 0.069	0.235 ± 0.108	0.287 ± 0.100	0.395 ± 0.146
ResNet34	0.510 ± 0.142	0.239 ± 0.095	0.216 ± 0.104	0.307 ± 0.080	0.447 ± 0.193
ResNet50	0.457 ± 0.154	0.234 ± 0.120	0.224 ± 0.118	0.286 ± 0.126	0.405 ± 0.191
EfficientNet-B0	0.346 ± 0.070	0.108 ± 0.028	0.113 ± 0.039	0.142 ± 0.027	0.331 ± 0.119

strate high segmentation accuracy, with IoU scores of 0.9187 and 0.8967, respectively, indicating close alignment between the predicted and ground truth wound areas. In these cases, the SAM2.1 model effectively identifies the wound boundaries despite variability in wound size and location. In contrast, the third example shows a significantly lower IoU score of 0.4263, suggesting poor segmentation performance. The predicted mask reveals over-segmentation and misalignment with the wound region. These results underscore the model's high performance in most cases, while also highlighting the need for improvement in more challenging imaging scenarios.

To validate the utility of the segmentation masks for clinical diagnosis, we analyzed the fidelity of the subsequent HSV color extraction. We calculated the Euclidean distance between the predicted and ground-truth color percentages across seven reference categories (red, orange, yellow, magenta, white, gray, and black). As shown in Fig. 8, 57.4% of all color predictions fell within a Euclidean distance of 0 to 20, indicating high color-matching accuracy. This distribution confirms that the SAM2.1 segmentation engine is precise enough to enable reliable downstream analysis of tissue composition (e.g., detecting necrotic slough), which is a vital component of the automated Wound Risk Assessment Score (WRAS).

WoundView App

A web-based app was created using Gradio and was hosted on the Hugging Face platform. This software allowed the implementation of the multi-model framework that was intended for the final product. By creating *WoundView* as a web-based app, it now became accessible to millions of people all over the world. It is a user-friendly app with a patient entering their demographics, pre-existing health conditions, and a picture of their wound. Within mere seconds, classification and segmentation colors are extracted. Once the user is prompted to enter a conversation with the LLM, which will provide a WRAS score as well as a tailored treatment plan that advises the patient on the next steps they should take. This entire process

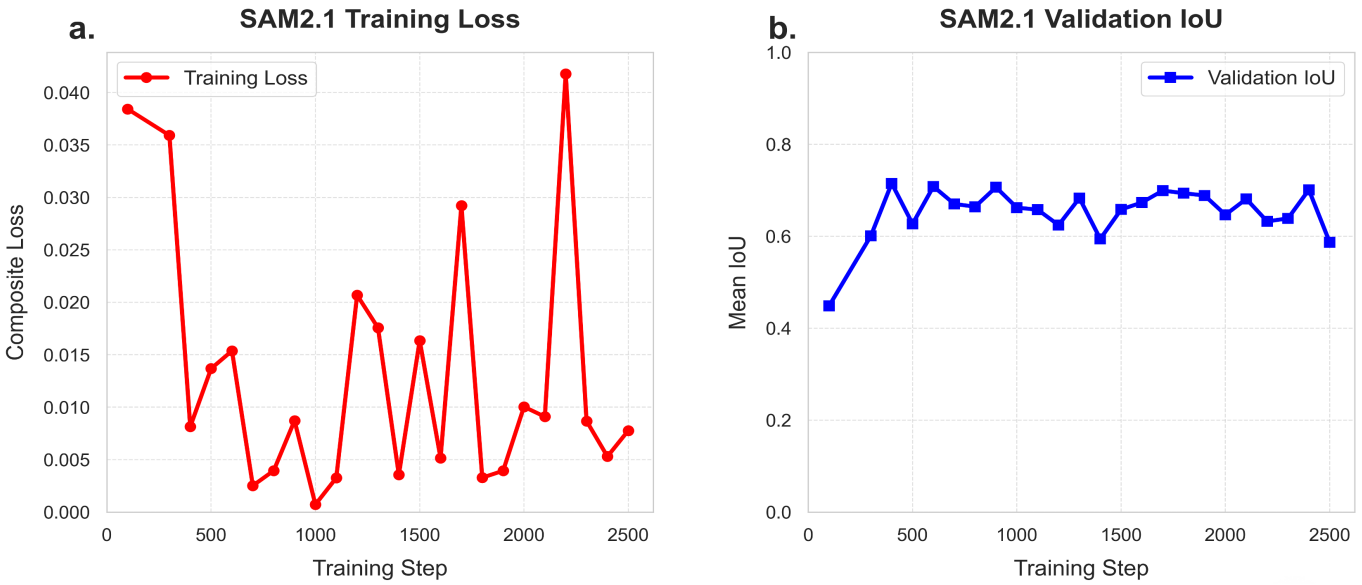

Figure 6: SAM2.1 Dice loss score (a) and segmentation learning curve (b).
The IoU curve illustrates the prompt-free automated performance reaching a peak of 0.7151. The loss function demonstrates stable convergence over the 2,500-step fine-tuning period.

Table 7 Segmentation Model Comparison

Model	Architecture	Validation IoU
U-Net	ResNet34 encoder (∼32M params)	0.7364
SAM2.1 (fine-tuned, prompt-free)	Hiera Transformer Backbone + mask decoder (∼4M trainable)	0.7151

occurs over the span of a few minutes.

Data Analysis

In order to test the reliability of the physicians' responses, Spearman's Rank Correlation Test was performed between the two replicate responses for each case.

$$\rho = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

(Equation 2. Spearman's Rho Correlation Coefficient Equation)

Spearman's Rho Correlation Coefficient was computed on ranks between replicate physician responses for all 10 wound cases (Table 8). Across all 10 cases, the overall correlation was $\rho = 0.611$ ($p < 0.01$), indicating moderate-to-strong agreement. Cases 1, 2, and 8 showed high inter-physician agreement ($\rho \geq 0.85$), while Cases 4 and 7 showed lower agreement ($\rho = 0.237$ and $\rho = 0.200$ respectively). All 10 cases were retained in the final analysis (Table 8).

In order to evaluate the performance of each LLM, physicians' responses were averaged and plotted for comparisons (Fig. 11). They rated LLM's responses on a five-point Likert scale (1 - very poor, 2 - minimally accurate, 3 - partially correct, 4 - mostly correct, 5 - perfect response). Basic models 3.5, 4o-mini, and 4o scored 2.94, 3.19, and 3.75, respectively. WRAS-integrated (+) models 3.5+, 4o-mini+, and 4o+ scored 3.56, 3.81, and 3.69, respectively. The WRAS model 4o-mini+ had the best performance, followed by 4o and 4o+.

Of the 120 physician responses collected, the majority of the responses fell within ratings 4 and 5 which were mostly correct and perfect responses (Fig. 12). The high number of ratings of 4 and 5 suggests that the model's WRAS score and tailored treatment recommendations are closely aligned with clinical evaluations by doctors indicating overall positive acceptance of the models' performance. The scarcity of ratings of 1 and 2 highlights the model's reliability and interpretability, as its WRAS scores and treatment recommendations consistently reflected clinical judgment.

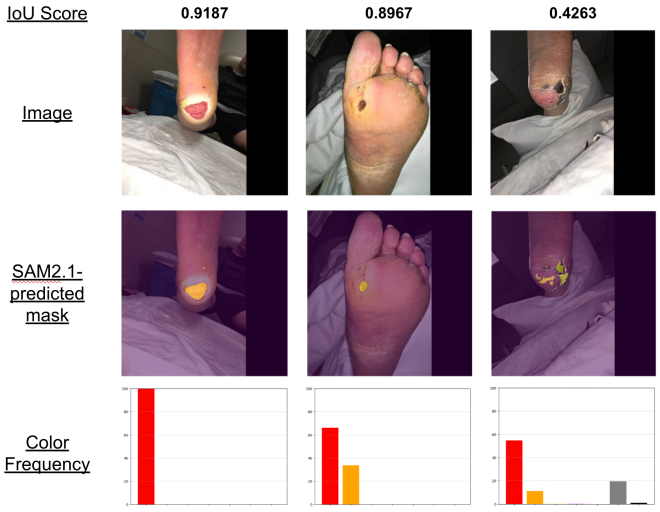

Figure 7: Examples of SAM2.1 performance on selected wound images.
Representative cases from the validation set showing raw images, predicted masks, and corresponding tissue color extraction.

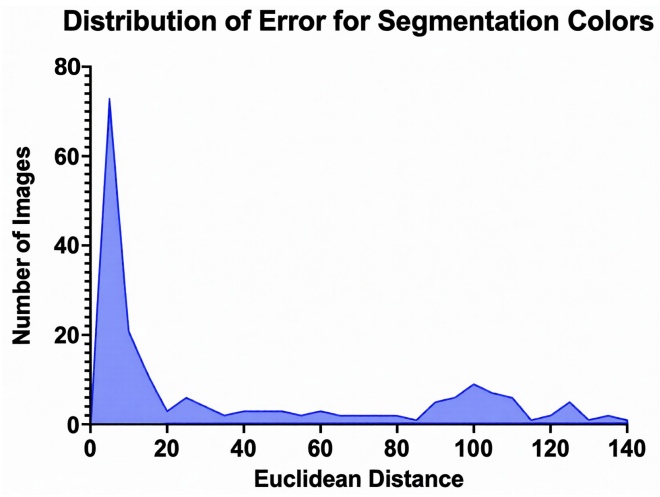

Figure 8: Distribution of error for segmentation colors of Dataset #3.
Histogram showing the Euclidean distance between predicted and ground-truth color classifications in HSV space.

Discussion

WoundView demonstrates strong performance as a proof-of-concept AI-assisted wound assessment tool. It is lightweight, affordable, and accessible via a web interface, making it suitable for low-resource settings. We were successfully able to provide a patient-specific wound assessment and treatment plan in under a minute. There were three critical factors that were involved in this process: the ResNet18 CNN, SAM2.1 segmentation, and OpenAI LLMs. Achieving a macro-averaged F1 of 0.776 in 18-way classification (weighted F1: 0.925) proves the effectiveness of the advanced ResNet18 CNN in accurately categorizing diverse wound types. The SAM2.1 segmentation model exhibited a robust performance, achieving a prompt-free validation IoU of 72%, precisely identifying ROI from the wound images. Additionally, strong correlation among physicians further validates the accuracy, reliability, and decision-making of the AI models. Our novel WRAS score delivers valuable information to patients, as confirmed by high physician evaluation scores for WRAS-integrated LLMs. The multi-faceted approach of integrating three different models (classification, segmentation, LLM), makes this comprehensive app unique.

Table 8 Data Analysis Showing Spearman's Rho Correlation Coefficient for Each Case, Selected Cases, and All Cases

Wound Case	Spearman's Rho Correlation Coefficient
Case #1	0.85257
Case #2	0.88536
Case #3	0.63246
Case #4	0.23735
Case #5	0.71204
Case #6	0.77460
Case #7	0.20000
Case #8	0.86603
Case #9	0.69561
Case #10	0.42906
Total	**0.61067**

WoundView can help reduce the burden of wound care management for patients, caregivers, and the health care system. It demonstrates its potential as an innovative AI-powered diagnostic tool, enabling accurate, real-time, remote and cost-effective wound assessment. In rural areas and developing countries, *WoundView* can help to serve as an initial point of diagnosis in cases where access to doctors and other professionals may be limited or unfeasible. Timely detection, awareness, and appropriate treatment can prevent the possibilities of the wound's development into severe cases such as permanent disability. *WoundView* represents a comprehensive integration of wound classification, segmentation, color analysis, and LLM-generated treatment planning in a single patient-facing application. Although accurate measures of wound care may exist, to this day, no widespread use of a mobile app among doctors and patients with a tailored treatment plan exists. In addition, this can be used as a telemedicine tool that highlights

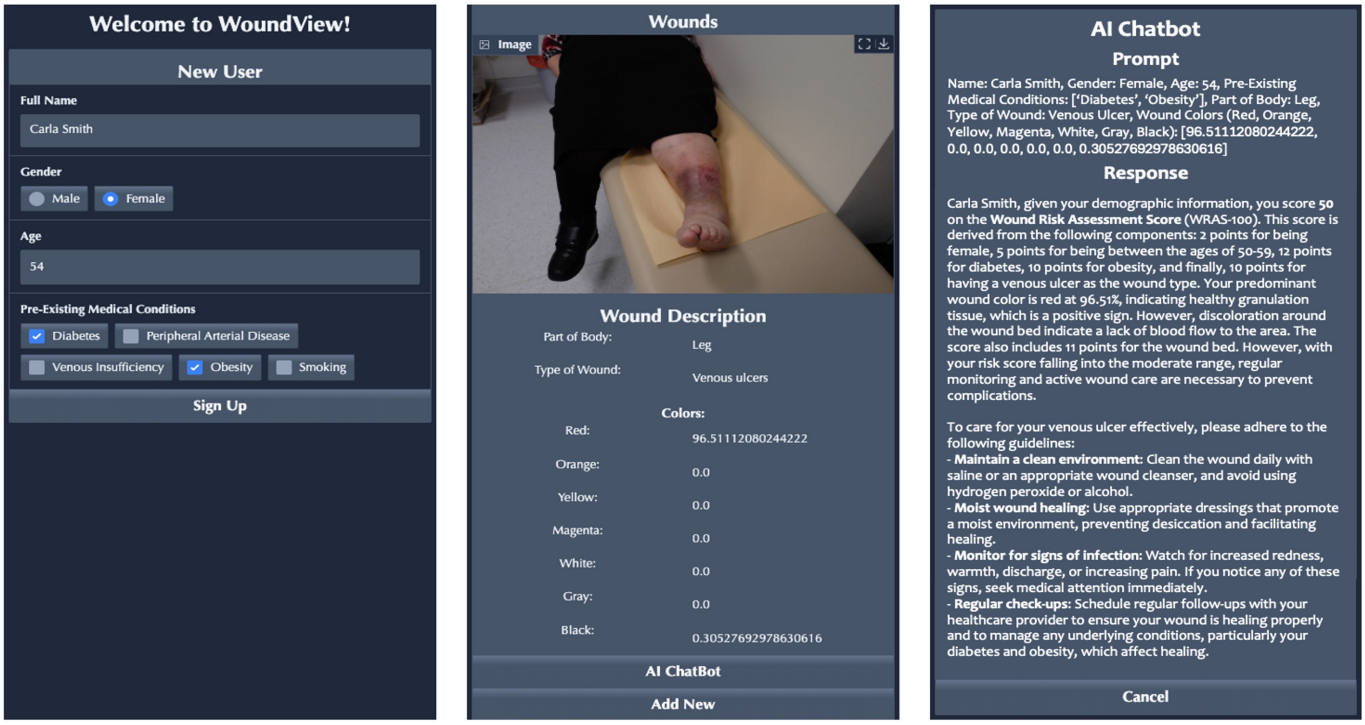

Figure 9: Screenshots from fully functional web-based *WoundView* App.
This is an example of a patient image from our dataset for demonstration purposes. Our app contains a very user-friendly interface and simple instructions for the patient to follow.

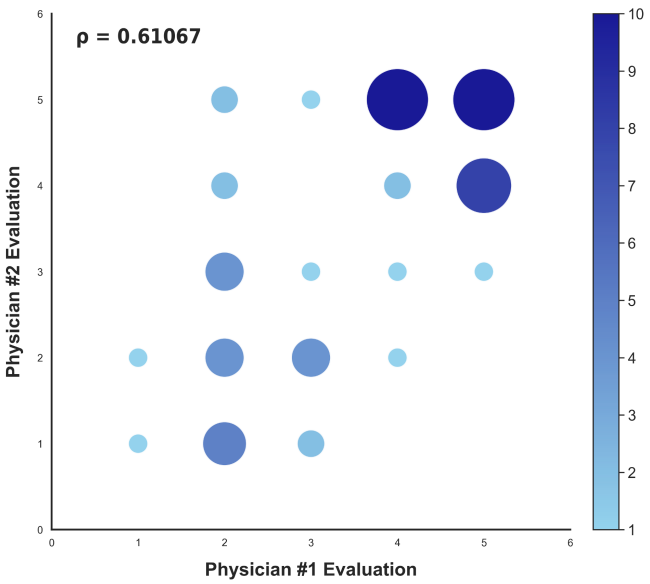

Figure 10: Replicate Physician Response Correlations for all 10 Wound Cases.
Overall $\rho = 0.611$ ($p < 0.01$).

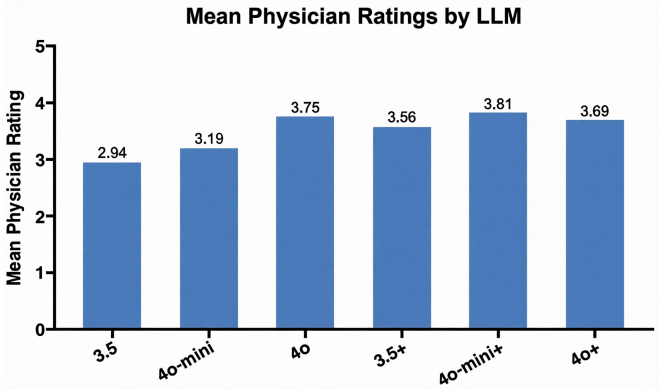

Figure 11: Mean Physician's Rating of LLM Models.

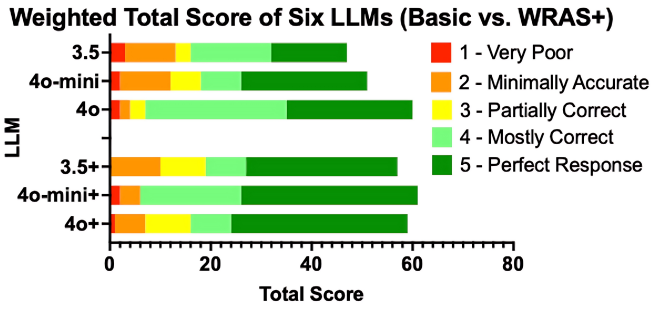

Figure 12: Weighted Total Score of Six LLMs (Basic vs. WRAS+).

wound severity on a simple numeric scale easy for patients to interpret.

This study has several important limitations. First, all models were trained and evaluated on publicly available benchmark datasets; real-world performance will be subject to variability in image quality, lighting, wound presentation, and patient demographics not represented in these datasets. Second, significant class imbalance in Dataset #3 limits classification performance for rare wound types: five of the ten DS3 classes (Amputation, Arteriovenous Fistula Failure, Injury, Lymphedema, Mastectomy) achieved zero F1 across all architectures, reflecting test-set support of 2–9 images each. Five-fold cross-validation on Dataset #3 yielded a macro F1 of at most 0.239 ± 0.095 across all four architectures, confirming that reliable classification of rare chronic wound types will require substantially expanded annotated datasets. Third, the LLM evaluation was limited to 10 wound cases evaluated by physicians; this sample is too small to draw statistically robust conclusions about clinical reliability. Fourth, LLMs carry an inherent risk of hallucination, generating plausible-sounding but medically incorrect treatment recommendations. Fifth, the WRAS scoring system, while grounded in clinical literature, has not been prospectively validated against patient outcomes.

In the future, we would look into incorporating size and depth of wound as well as pain level to further enhance wound analysis. Additionally, access to larger diverse datasets could result in additional improvements, and increased clinical feedback from wound care specialists will reduce variability in the physicians' evaluations of the model. *WoundView* can also benefit from key additive features such as chat history and wound timeline. Our goal would be to convert this app to a mobile application and deploy it to iOS and Android devices.

A formal roadmap for clinical deployment is planned, beginning with an IRB-approved prospective trial to validate *WoundView* against real-world longitudinal patient outcomes. This will be followed by establishing a regulatory pathway for mobile-based medical diagnostic support tools, ensuring full compliance with clinical safety and data privacy standards.

References

1 C. K. Sen. Human wound and its burden: updated 2025 compendium of estimates. Advances in Wound Care. Vol. 14, 2025, https://doi.org/10.1177/21621918251359554.

2 M. Spear. Acute or chronic? Whats the difference? Plastic Surgical Nursing. Vol. 33, pg. 98-100, 2013, https://doi.org/10.1097/psn.0b013e3182965e94.

3 L. Gould, P. Abadir, H. Brem, M. Carter, T. Conner-Kerr, J. Davidson, L. DiPietro, V. Falanga, C. Fife, S. Gardner, E. Grice, J. Harmon, W. R. Hazzard, K. P. High, P. Houghton, N. Jacobson, R. S. Kirsner, E. J. Kovacs, D. Margolis, F. McFarland Horne, M. J. Reed, D. H. Sullivan, S. Thom, M. Tomic-Canic, J. Walston, J. A. Whitney, J. Williams, S. Zieman, K. Schmader. Chronic wound repair and healing in older adults: current status and future research. Journal of the American Geriatrics Society. Vol. 63, pg. 427-438, 2015, https://doi.org/10.1111/jgs.13332.

4 S. Guo, L. A. DiPietro. Factors affecting wound healing. Journal of Dental Research. Vol. 89, pg. 219-229, 2010, https://doi.org/10.1177/0022034509359125.

5 Y. F. Liu, P. W. Ni, Y. Huang, T. Xie. Therapeutic strategies for chronic wound infection. Chinese Journal of Traumatology. Vol. 25, pg. 11-16, 2022, https://doi.org/10.1016/j.cjtee.2021.07.004.

6 L. Gould, I. Herman. Out of the darkness and into the light: confronting the global challenges in wound education. International Wound Journal. Vol. 22, 2025 https://doi.org/10.1111/iwj.70178.

7 J. M. Hwang. Time is tissue. Want to save millions in wound care? Start early: a QI project to expedite referral of high-risk wound care patients to specialised care. BMJ Open Quality. Vol. 12, pg. e002206, 2023, https://doi.org/10.1136/bmjoq-2022-002206.

8 A. Fernández-Araque, M. Martinez-Delgado, J. M. Jiménez, M. López, M. J. Castro, E. C. Gila. Assessment of nurses' level of knowledge of the management of chronic wounds. Nurse Education Today. Vol. 134, pg. 106084, 2024, https://doi.org/10.1016/j.nedt.2023.106084.

9 A. Friman, D. W. Edström, S. Edelbring. General practitioners' perceptions of their role and their collaboration with district nurses in wound care. Primary Health Care Research Development. Vol. 20, 2018,https://doi.org/10.1017/s1463423618000464.

10 T. M. Klein, V. Andrees, N. Kirsten, K. Protz, M. Augustin, C. Blome. Social participation of people with chronic wounds: a systematic review. International Wound Journal. Vol. 18, 2020,https://doi.org/10.1111/iwj.13533.

11 R. S. Howell, H. H. Liu, A. A. Khan, J. S. Woods, L. J. Lin, M. Saxena, H. Saxena, M. Castellano, P. Petrone, E. Slone, E. S. Chiu, B. M. Gillette, S. A. Gorenstein. Development of a method for clinical evaluation of artificial intelligence–based digital wound assessment tools. JAMA Network Open. Vol. 4, pg. e217234, 2021, https://doi.org/10.1001/jamanetworkopen.2021.7234.

12 Y. Patel, T. Shah, M. K. Dhar, T. Zhang, J. Niezgoda, S. Gopalakrishnan, Z. Yu. Integrated image and location analysis for wound classification: a deep learning approach. Scientific Reports. Vol. 14, 2024, https://doi.org/10.1038/s41598-024-56626-w.

13 I. Fateen. Collected and categorized wound images dataset. 2023 https://www.kaggle.com/datasets/ibrahimfateen/wound-classification.

14 M. Taher. Wound data. 2021 https://www.kaggle.com/datasets/mohamadtaher/wound-data.

15 M. Kręcichwost, J. Czajkowska, A. Wijata, J. Juszczyk, B. Pyciński, M. Biesok, M. Rudzki, J. Majewski, J. Kostecki, E. Pietka. Chronic wounds multimodal image database. Computerized Medical Imaging and Graphics. Vol. 88, pg. 101844, 2021,https://doi.org/10.1016/j.compmedimag.2020.101844.

16 G. Kourounis, A. Elmahmudi, B. Thomson, J. Hunter, H. Ugail, C. Wilson. Computer image analysis with artificial intelligence: a practical in-

troduction to convolutional neural networks for medical professionals. Postgraduate Medical Journal. 2023, https://doi.org/10.1093/postmj/qgad095.

17 H. C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, R. M. Summers. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Transactions on Medical Imaging. Vol. 35, pg. 1285-1298, 2016, https://doi.org/10.1109/tmi.2016.2528162.

18 K. He, X. Zhang, S. Ren, J. Sun. Deep residual learning for image recognition. arXiv. 2015, https://doi.org/10.48550/arxiv.1512.03385.

19 N. Ravi, V. Gabeur, Y. T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C. Y. Wu, R. Girshick, P. Dollár, C. Feichtenhofer. SAM 2: segment anything in images and videos. arXiv. 2024, https://doi.org/10.48550/arxiv.2408.00714.

20 X. Li, X. Tian, Z. Wang, F. Zhang, Y. Zhang, N. Yang, C. Tian. SAM2Former: Segment Anything Model 2 Assisting UNet-Like Transformer for Remote Sensing Image Semantic Segmentation. IEEE Access. Vol. 13, pg. 115018-115032, 2025,https://doi.org/10.1109/ACCESS.2025.3583458.

21 Z. Jiaxing, T. Hao. SAM2 for image and video segmentation: a comprehensive survey. arXiv. 2025, https://doi.org/10.48550/arXiv.2503.12781.

22 J. E. Grey, S. Enoch, K. G. Harding. Wound assessment. British Medical Journal. Vol. 332, pg. 285-288, 2006, https://doi.org/10.1136/bmj.332.7536.285.

23 V. Kondekar, S. Bodhe. A comprehensive investigation of color models used in image processing. International Journal of Computer Applications. Vol. 180, pg. 19-24, 2018, https://doi.org/10.5120/ijca2018916507.

24 P. Papanikolaou, P. Lyne, D. Anthony. Risk assessment scales for pressure ulcers: a methodological review. International Journal of Nursing Studies. Vol. 44, pg. 285-296, 2007, https://doi.org/10.1016/j.ijnurstu.2006.01.015.

25 S. Minaee, T. Mikolov, N. Nikzad, M. Chenaghlu, R. Socher, X. Amatriain, J. Gao. Large language models: a survey. arXiv. 2024, https://doi.org/10.48550/arxiv.2402.06196.

Appendix/Supplemental Work

Appendix A: LLM System Prompt Template

The following system prompt was utilized to standardize the OpenAI Large Language Model (LLM) responses. This prompt instructs the model to act as a clinical expert, integrates the multimodal data from the classification and segmentation models, and enforces the use of the Wound Risk Assessment Score (WRAS):

"You are an expert doctor who treats chronic wounds, and you know every single thing about wounds and how to treat them as well as preventing them from getting worse. The user will provide the following inputs: Name, Gender, Age, Pre-existing Medical Conditions, Wound Part of Body, Wound Classification, Colors of the Wounds (as percents out of 100).

Please provide the medical advice in 2 concise paragraphs that must incorporate the following key features every time:

1. Wound Risk Score (1-100): Generate a wound risk score from 1-100, 1 being no risk and 100 being going to see a medical professional immediately! Any color percentages less than 3% should not be taken into consideration when making the score. Specific components of the score must be listed.

2. Medical Advice: Provide directions on monitoring and caring for the wound, explicitly stating if the patient requires immediate medical intervention."

Appendix B: Dataset Visualization

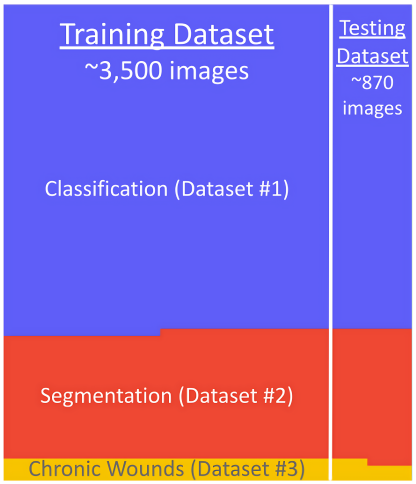

Figure 13: Visual representation of datasets 1-3 and the 80/20 split for training and testing, respectively.

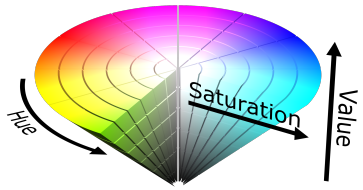

Figure 14: HSV Color Distribution Visualization

Table 9 Color Ranges in HSV Values The HSV ranges utilized for tissue classification were empirically derived based on clinical wound bed color standards (e.g., red for granulation, yellow for slough) and validated against expert-annotated regions from Dataset #2 and #3.

Color Class	Hue Range	Saturation Range	Value Range
Red_1	[0, 10]	[50, 255]	[50, 255]
Red_2	[170, 179]	[50, 255]	[50, 255]
Orange	[11, 25]	[50, 255]	[50, 255]
Yellow	[26, 35]	[50, 255]	[50, 255]
Green	[36, 85]	[50, 255]	[50, 255]
Cyan	[86, 95]	[50, 255]	[50, 255]
Blue	[96, 130]	[50, 255]	[50, 255]
Purple	[131, 160]	[50, 255]	[50, 255]
Magenta	[161, 169]	[50, 255]	[50, 255]
White	[0, 179]	[0, 55]	[200, 255]
Gray	[0, 179]	[0, 50]	[50, 200]
Black	[0, 179]	[0, 50]	[0, 50]