WHITE PAPER

Accelerated Computing for Medical AI

Wound Analysis for Diabetic Patients

OUTLINE

FIGURES

TABLES

Executive Summary

Diabetic wounds, exacerbated by elevated blood glucose levels (i.e., hyperglycemia), present unique complexities and risks of severe complications, necessitating rigorous monitoring and intervention. Effective wound care is critical for managing these wounds and improving patient outcomes. To address this, we are developing an AI-driven solution designed to monitor wound status continuously and facilitate timely intervention on QCT DevCloud, which is developed by Quanta Cloud Technology in collaboration with Intel. Our approach leverages a balanced compute infrastructure to ensure optimal performance and cost-efficiency: inference is accelerated by the Intel Flex 170, while the NVIDIA H100 powers the training process. This configuration supports a scalable, real-time monitoring system tailored for diabetic wound care, ensuring both high accuracy and economic viability in clinical settings.

Introduction

Diabetic patients often face heightened risks of severe complications, including amputation and mortality, due to complex wound healing processes exacerbated by high blood glucose levels. Effective wound care is essential, particularly for foot wounds that can quickly worsen. However, wound management for diabetic patients is often inconsistent and limited, especially since many patients only seek hospital care when complications are advanced, an issue more pronounced in rural areas with limited healthcare access.

A key barrier in diabetic wound management is the subjective nature of in-person assessments. Patients must typically visit healthcare facilities to receive evaluations, which depend heavily on clinicians' personal impressions, leading to inconsistent judgments and misdiagnoses at times. To overcome these limitations, we are developing an AI-driven tool to empower diabetic patients to assess and monitor their wound conditions independently, using smartphone-based technology. This solution allows continuous tracking and recording of wound status, reducing dependence on direct physician intervention. By integrating this tool into diabetic wound care, we aim to enhance accuracy and timeliness in wound assessment, ultimately reducing the risk of critical complications. Furthermore, this system allows healthcare providers to monitor wound progression remotely, supporting active intervention when necessary. The data-driven design ensures that wound conditions are recorded objectively, leading to more accurate diagnoses and resource-efficient management. Ultimately, this approach aims to improve patient outcomes and alleviate strain on healthcare resources.

To develop an effective AI-based wound assessment system, it is essential to align our approach with the diagnostic criteria used by clinicians. Physicians assess wound healing based on indicators such as tissue color, distribution of granulation tissue, and patient history. In contrast, our solution will rely on images to extract measurable features, including color, texture, and wound structure. We use a neural network approach to analyze wound structure and tissue distribution, focusing initially on identifying granulation tissue — a key marker of healing. By employing a composite model strategy, where each model specializes in detecting a particular tissue type, we reduce model complexity while enhancing performance. Our chosen architecture is a modified ResUNet[1] model, optimized to capture the nuances of granulation tissue. The model is implemented in PyTorch[2] to support cross-platform deployment, ensuring compatibility across various hardware configurations.

Challenges in Diabetic Wound Management

Diabetic wound management presents unique challenges due to the complex nature of wound healing in patients with high blood glucose levels. A key aspect of effective wound care is the formation of granulation

tissue, an indicator that the wound is progressing through the healing process. Shown as Figure 1, granulation tissue, which appears bright red, bumpy, and moist, consists largely of microvascular white blood cells and serves as a foundation for new tissue and blood vessel growth. The extent and distribution of this tissue are critical markers of wound condition, indicating whether a wound is healing or deteriorating. However, in diabetic patients, wounds are often more severe and may penetrate deeply, exposing underlying structures such as bone, ligaments, and muscle. This complexity, combined with a high risk of infection and the formation of microbial biofilms, complicates assessments based on granulation tissue alone. Biofilms create a barrier that hinders healing and may mask infection severity, further challenging wound evaluation and management.

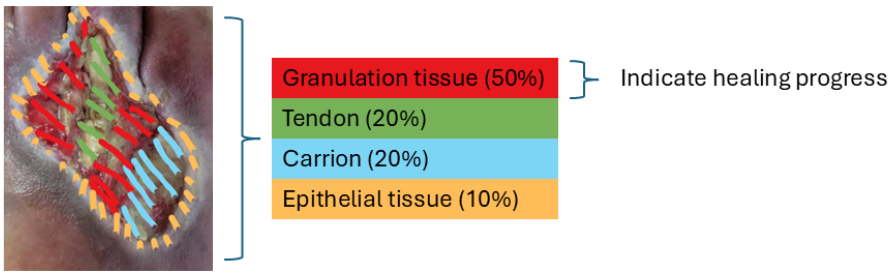

Figure 1: The different types of wound issues

The COVID-19 pandemic has amplified these issues, as hospitals now prioritize limiting non-emergency visits, which can significantly impact diabetic patients' access to timely wound care. While younger patients may manage with remote guidance, middle-aged and elderly diabetic individuals are at greater risk. These patients, who are often more vulnerable to injury and are less equipped to assess wound condition accurately, face increased mortality risk from infections and complications arising from delayed intervention.

To address these challenges, healthcare providers must adopt innovative, remote assessment tools that allow for accurate, real-time monitoring of diabetic wounds. Such solutions could greatly enhance the continuity of care, mitigate infection risks, and improve outcomes for this high-risk population.

Materials and Methods

Dataset, Preprocessing and Training:

The dataset is provided by Chan Gung Memorial Hospital. We collected 1200 wound pictures and augment them up to 640 times (i.e., the number of augmented datasets is 1200*640 in data preprocessing). The dataset is divided into two groups, 90% images for training, and 10% images for inference. In our case collection process, we observed significant variability in wound assessments among clinicians, resulting in inconsistent labeling of identical wound cases. To mitigate this, we implemented a voting system using a crowdsourced labeling approach. By having multiple clinicians label each wound and applying a consensus-based voting mechanism, we minimized individual bias and established a more reliable ground truth for model training.

To enable patients to operate the system independently from home using their smartphones, we must account for variability in image quality due to differences in device capabilities and environmental conditions. Consequently, it is essential to simulate a range of real-world scenarios during model training, including variations in lighting, resolution, and angle. This approach helps the model become more robust, allowing it to perform accurately across diverse conditions and device types typically encountered in home settings.

Our training process is configured to ensure high precision, essential for accurate wound assessment at the pixel level. With a learning rate of 0.00001, we address the risk of gradient explosion inherent in our model and dataset. Data is processed in FP32 to capture subtle wound tissue variations accurately, as any pixel-level errors could undermine diagnostic accuracy. This precision level supports the model's ability to distinguish fine tissue details for reliable wound monitoring. The training consists of 30,000 steps, optimized to achieve the required depth and accuracy in wound feature extraction.

Detection Model:

The task of wound assessment through computational methods primarily relies on analyzing texture and color information in images. Convolutional Neural Networks[3] (CNNs) are particularly suited for this, as they excel in feature extraction and semantic segmentation by combining extracted features across layers. Therefore, CNN-based architecture is the optimal choice for wound tissue feature analysis and segmentation.

In medical imaging research, it is often necessary to use image segmentation to locate specific target areas within an image and clearly identify their locations. This typically involves using CNNs + Fully Convolutional Networks[4] (FCNs) + Class Activation Mapping[5] (CAMs) or construct the decoder and encoder by CNN.

For clinicians, a wound represents an image composed of different tissues, while in the computational realm, it is viewed as an image made up of sub-images characterized by various colors and textures. U-Net facilitates identifying these different colors and textures. It is a model that incorporates both encoder and decoder structures based on CNNs. This principle holds true for most medical imaging applications, where it is sufficient to identify the specific features which exist in designated locations. However, pure U-Net could lose raw information after a few layers, and we believe that raw texture information is essential at every layer[6]. To address this, we adopted ResUNet[1] with attention (transformer) as Figure 2 shown, which can pass raw information to subsequent layers. The attention layer is crucial because wounds are complex, and granulation tissue is not a single object, such as a dog or cat. Instead, it is dispersed throughout the wound. This means the model requires global information to accurately determine whether a tissue is granulation tissue or not. Due to the complexity of our model structure, which incorporates transformers and residual blocks and the very small learning rate caused by our training data and model characteristics, substantial computational power is necessary.

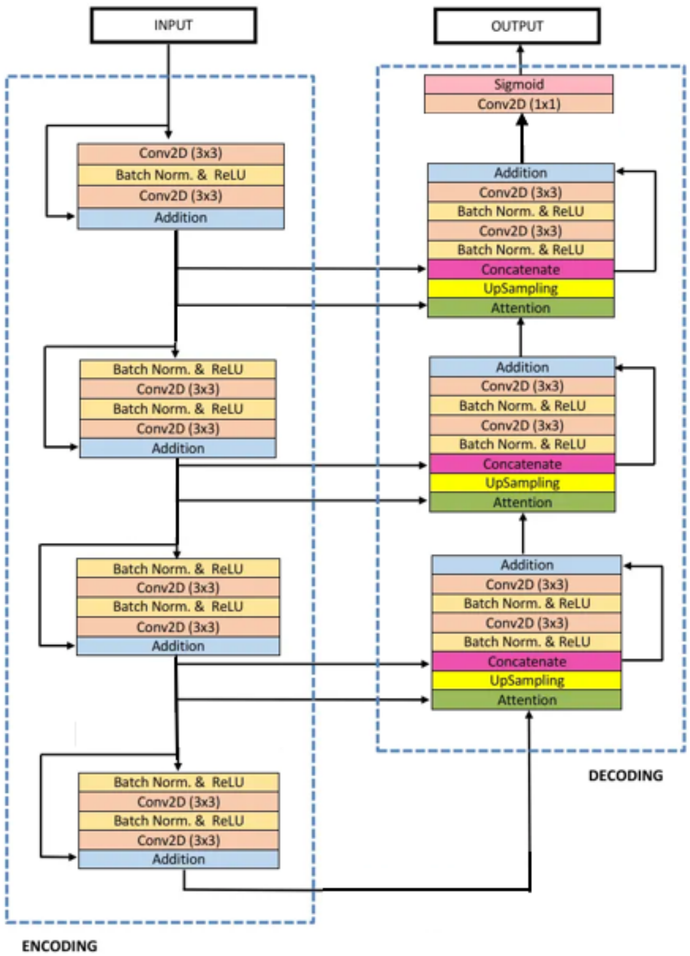

Figure 2: The architecture of attention layer in our model.

Resource Evaluation

Evaluating computational resources is a critical first step in optimizing system performance. To achieve this, we utilized the QCT DevCloud platform[7], developed by Quanta Cloud Technology (QCT) in collaboration with Intel. This platform is specifically designed to address challenges in the development and testing of high-performance computing (HPC) and AI applications by providing a comprehensive remote development environment.

The QCT DevCloud offers access to cutting-edge hardware, storage, and networking resources, along with advanced features of Intel Xeon processors, such as enhanced security and accelerated HPC/AI capabilities. It also includes precompiled workloads, a robust software stack, and tools like Intel oneAPI, streamlining development efforts by eliminating the need for complex installations. Moreover, the platform enables testing across various CPU and GPU environments, allowing developers to evaluate system performance efficiently. This capability reduces development time and resource constraints, ultimately leading to significant cost savings. Figure 3 outlines the key benefits of the QCT DevCloud for developers, emphasizing its role in expediting resource evaluation and deployment processes while ensuring optimal hardware utilization.

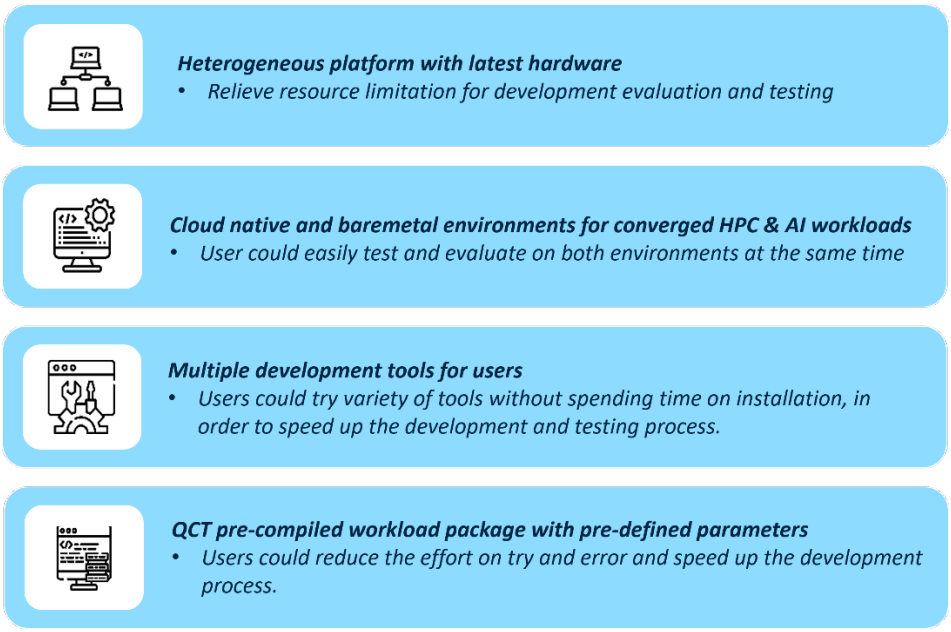

Figure 3: The QCT DevCloud benefits for developer.

The benchmark results are shown in Table 1. We present several examples and provide benchmark comparisons of training and inference performance across different batch sizes and hardware configurations. And the information of software stack and specification of computational nodes are shown in Table 2. The training process involved four primary stages: data copying (HDD/SSD to main memory), message passing (main memory to accelerator's memory), mathematical calculation on accelerator, and other minor processes. Among these, data copying and miscellaneous processes were minimal, so our analysis emphasized message passing and mathematical calculations.

Accelerator	Training (seconds/step)				Inference (seconds)	
	Batch Size	Math. Cal. Proc[1]	MSG. Pass.[2]	Full Step[3]	Batch Size	Full Step
Nvidia H100	4	0.03	0.36	0.39	1	0.04
Nvidia H100 x 8	24 x 8	1.57	0.96	2.53	-	-
Intel Flex 170	4	4.89	0.06	4.95	1	0.76

[1] Mathematical calculation process.
[2] Message passing process.
[3] In this table, for comparison easily, the full step = mathematical calculation process + message passing process
[4] Less is better.

Table 1: The benchmark with different batch sizes of training and inference.

	NVIDIA H100 SXM5	Intel Flex 170
Software Stack	Python 3.9.18	
	CUDA 12.4 NCCL 2.20.5	Intel OneAPI 2024.2
	Pytorch 2.0.1	Pytorch 2.0.1 Intel extension for Python (IPEX)
Model	QuantaGrid D74H-7U	QuantaGrid D54U-3U
Accelerator	NVIDIA H100 SXM5 80GB x 8	Intel Flex 170 PCIE 16GB x 4
CPU	Xeon Platinum 8480+ x 2	Intel Xeon Platinum 8458P x 2
RAM	DDR5-4800 64GB x 32, Total Size: 2 TB	DDR5-4800 64GB x 16, Total Size: 1 TB
OS	Rocky Linux release 9.2 (Blue Onyx)	

Table 2: The software stack and specifications of computation nodes.

For message passing, which transfers data from main memory to accelerator's memory, the H100 over NVLink exhibited longer transfer times, with a single H100 requiring 0.36 seconds (batch size 4). In contrast, the Flex 170, operating over PCIe, completed message passing in 0.06 seconds for the same batch size. For larger configurations, eight H100s' processing batch sizes of 24 each needed 0.96 seconds for message passing. In terms of mathematical calculation performance, the H100 demonstrated significant efficiency, with a single card needing just 0.03 seconds for computations at batch size 4. The Flex 170, by comparison, required 4.89 seconds, indicating the H100's processing advantage at roughly 163 times faster. During inference, which generally demands less computational power, the H100 continued to demonstrate superior performance, completing a single inference in 0.04 seconds—19 times faster than the Flex 170, which required 0.76 seconds.

Feature	NVIDIA H100 SXM5	Intel Flex 170
FP32 Performance	67 TFLOPS	16 TFLOPS
Power Consumption	700W (TDP)	150W (TDP)
Power Efficiency	~ 95.71 GFLOPS/W (FP32)	~ 106.66 GFLOPS/W (FP32)

Table 3: The comparison of power efficiency between H100 and Flex 170.

Our testing revealed that the inference times of 0.04 seconds (NVIDIA H100) and 0.76 seconds (Intel Flex 170) result in negligible differences in user experience. However, when factoring in power efficiency, as outlined in Table 3, a strategic division of tasks becomes evident. The NVIDIA H100, with its exceptional performance in mathematical computations (67 TFLOPS FP32), is highly suited for computationally intensive training tasks, leveraging its superior processing power despite higher power consumption (700W TDP).

In contrast, the Intel Flex 170, with a power consumption of only 150W and a higher power efficiency of 106.66 GFLOPS/W (FP32), presents a cost-effective solution for inference workloads, especially for operations with lower computational requirements. This optimization balances performance and operational costs while ensuring sustainability.

Figure 4 illustrates the overall data flow diagram, where the green arrow represents the training process flow utilizing the H100's capabilities, and the blue arrow signifies the inference process managed by the Flex 170. This division ensures both performance optimization for training and cost-effectiveness during inference, making the solution highly adaptable to real-world use cases.

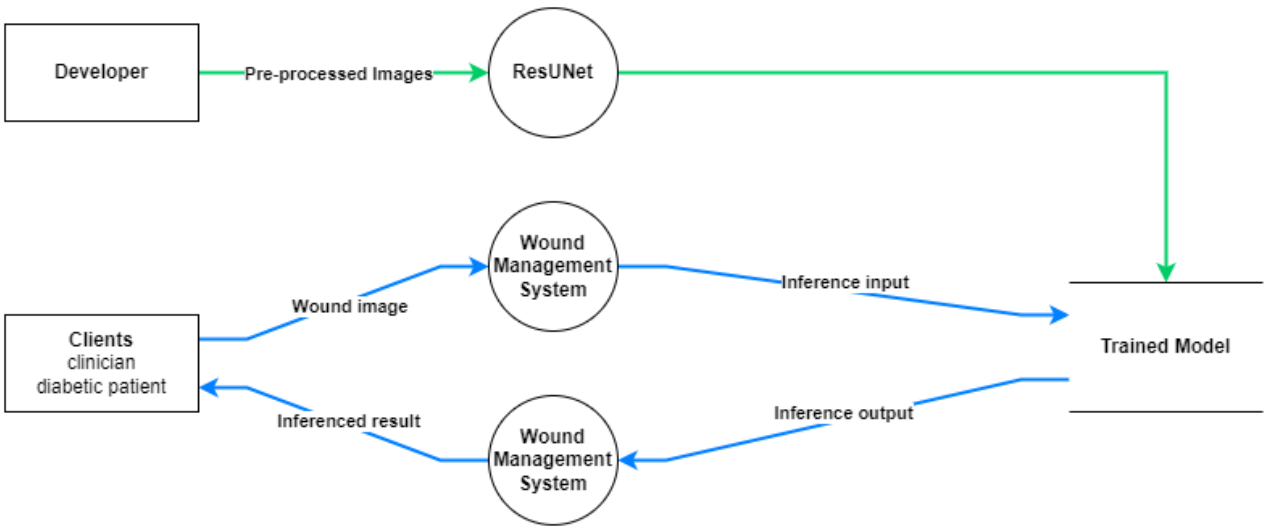

Figure 4: Overview of data flow diagram.

This approach meets real-time response requirements without incurring the high expenses associated with training hardware. By aligning computational power with task-specific needs, we achieve efficient resource allocation across the training and inference phases. This accelerated solution provides diabetic patients with timely, reliable insights into their wound status, enhancing their ability to manage care independently and efficiently.

Results

Our model demonstrated robust performance on both simple and complex wound cases. In this example of a complex foot wound, the "test plus raw" output shows the model's accurate coverage of granulation tissue across the wound area. As Figure 5 shown, compared to the ground truth, only a small section of the microbial biofilm is misclassified. Additionally, certain dark regions in the image, which are challenging for experts to assess, were effectively identified by the model as potential granulation tissue. This capability underscores the model's precision in highlighting wound characteristics, even under suboptimal lighting conditions.

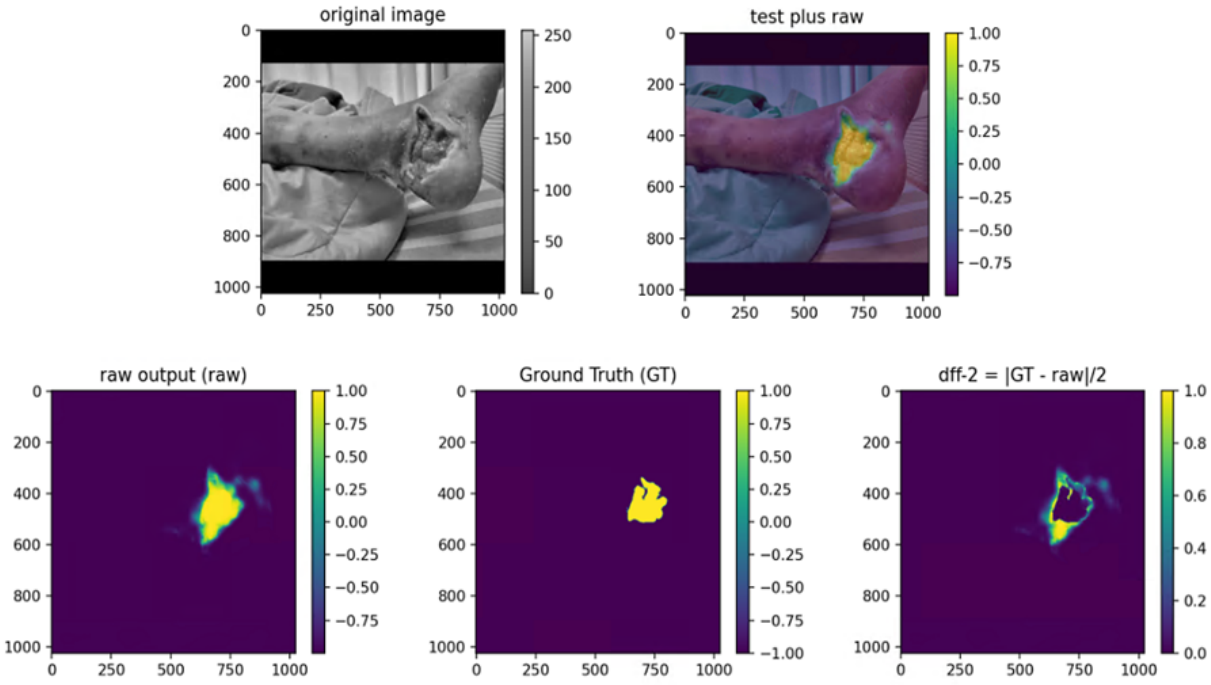

Figure 5: Comparing inference result and ground truth of case 1.

In this example of a simple foot wound, which primarily consists of a large area of granulation tissue with minimal presence of other tissue types, the model demonstrated high accuracy. The "test plus raw" output shows that all granulation tissue is correctly identified, even outperforming the ground truth in some respects. For instance, the model successfully marked a small region of granulation tissue (highlighted by the red circle in Figure 6) that the expert annotator overlooked. This highlights the model's sensitivity to fine details, ensuring comprehensive coverage that may surpass human labeling consistency.

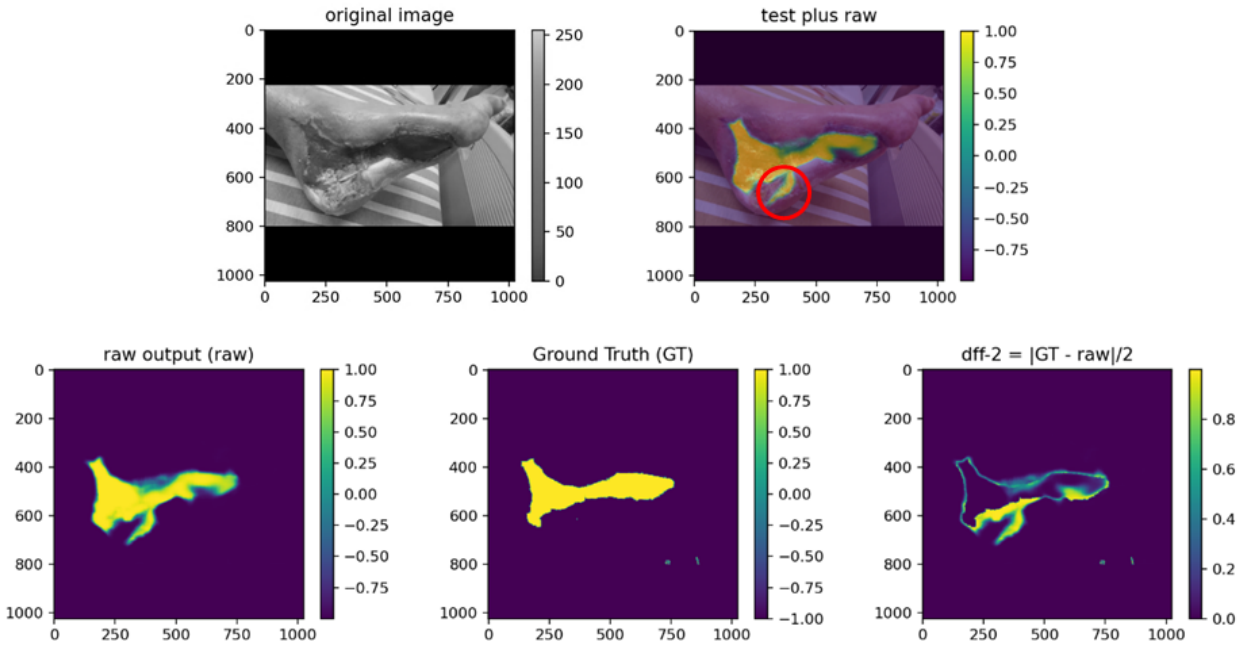

Figure 6: Comparing inference result and ground truth of case 2.

In our previously published paper[6], the model's performance on complex wounds was limited, resulting in suboptimal outcomes. However, with our new ResUNet model enhanced by transformer-based attention mechanisms, we have achieved significant improvements in handling complex wounds. This updated model consistently delivers strong performance, with an average Area Under the Curve (AUC) exceeding 0.9. The closer the AUC is to 1, the better the model's performance, indicating that the model can more accurately distinguish between positive and negative class samples. This demonstrates its ability to accurately assess challenging wound types and reliably differentiate intricate tissue features.

Conclusion

Implementing our system is expected to notably reduce the risks associated with diabetic wound self-care by providing a more reliable assessment tool. This AI-driven solution enhances diagnostic accuracy by equipping clinicians with objective, data-backed insights, thereby supporting improved patient outcomes through timely and precise interventions. Leveraging accelerated computing, the system also expedites the diagnostic process, easing the burden on medical resources and improving access to fast, reliable care. Furthermore, the QCT DevCloud platform has streamlined our development pipeline, reducing setup and preparation time for the system and development environment. The platform also enables efficient evaluation of resource requirements, helping us identify optimal configurations to support high-performance computing needs. These advancements make our system an efficient and scalable solution in diabetic wound management.

Future Work and Enhancements

Enhanced Model Accuracy through Advanced Accelerated Computing: Leveraging additional accelerated computing resources could improve model accuracy, allowing for more refined analysis and reliable wound assessment in real-world scenarios.

Expansion to Broader Wound Types and Conditions: Future developments could extend the system to monitor a variety of wound types and other medical conditions, making it a versatile tool in wound care management.

Integration with Wearable Devices for Continuous Monitoring: Integrating wearable devices would enable continuous, real-time monitoring of wounds, providing a more comprehensive view of wound healing progression and alerting patients and healthcare providers to any significant changes.

Real-Time Feedback and Alerts Powered by Enhanced Computing: Enhanced computing capabilities would support real-time feedback and alert systems, offering immediate insights and recommendations, and further improving patient engagement and care quality.

Acknowledgement

We would like to express our sincere gratitude to Angela Shin-Yu Lien, Chen-Yao Lai, Jyh-Da Wei, and their team for their invaluable support and contributions to this project.

Angela Shin-Yu Lien, from the School of Nursing, College of Medicine, Chang Gung University, and the Department of Endocrinology and Metabolism, Chang Gung Memorial Hospital Linkou Branch, played a key role in the research. Her work was supported by the Ministry of Science and Technology of Taiwan under grant MOST-109-2314-B-182-048, and by Chang Gung University and Chang Gung Memorial Hospital under grant BMRPJ25.

Jyh-Da Wei and Chen-Yao Lai, from the Department of Computer Science and Information Engineering, College of Engineering, and the Artificial Intelligence Research Center at Chang Gung University. The research of Jyh-Da Wei was supported by the Ministry of Science and Technology of Taiwan under grant MOST-110-2221-E-182-047, and by Chang Gung University and Chang Gung Memorial Hospital under grant BMRPB21.

References

1. Diakogiannis FI, Waldner F, Caccetta P, Wu C. ResUNet-a: a deep learning framework for semantic segmentation of remotely sensed data. Published online January 15, 2020. doi:10.48550/arXiv.1904.00592

2. Paszke A, Gross S, Massa F, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. Published online December 3, 2019. doi:10.48550/arXiv.1912.01703

3. O'Shea K, Nash R. An Introduction to Convolutional Neural Networks. Published online December 2, 2015. doi:10.48550/arXiv.1511.08458

4. Long J, Shelhamer E, Darrell T. Fully Convolutional Networks for Semantic Segmentation. Published online March 8, 2015. doi:10.48550/arXiv.1411.4038

5. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning Deep Features for Discriminative Localization. Published online December 14, 2015. doi:10.48550/arXiv.1512.04150

6. Lien ASY, Lai CY, Wei JD, Yang HM, Yeh JT, Tai HC. A Granulation Tissue Detection Model to Track Chronic Wound Healing in DM Foot Ulcers. *Electronics*. 2022;11(16):2617. doi:10.3390/electronics11162617

7. Ware A. QCT and Intel Codeveloped QCT DevCloud Program to Jumpstart HPC and AI Development. HPCwire. August 7, 2023. Accessed December 4, 2024. https://www.hpcwire.com/2023/08/07/qct-devcloud-program-provides-a-platform-to-jumpstart-hpc-and-ai-development/