Automated Unified Reasoning with Vision-Language Models for Multi-modal Burn Assessment

Md Masudur Rahman[1], **Mohamed El Masry**[2,3], **Gayle Gordillo**[2, 4], **Juan P Wachs** [1]

[1]Edwardson School of Industrial Engineering, Purdue University, West Lafayette, IN 47907, USA
[2]McGowan Institute for Regenerative Medicine (MIRM), Pittsburgh, PA 15219, USA
[3] Department of Surgery, University of Pittsburgh School of Medicine, Pittsburgh, PA 15213, USA
[4] Department of Plastic Surgery, University of Pittsburgh School of Medicine, Pittsburgh, PA 15261, USA
rahman64@purdue.edu, moelmasry@pitt.edu, gordillogm@upmc.edu, jpwachs@purdue.edu

Abstract

In emerging clinical applications such as ultrasound-based burn assessment, the lack of domain-specific data presents a significant challenge for developing robust AI systems. Vision-language models (VLMs) have shown strong performance in general computer vision tasks, yet their application to medical imaging remains limited, particularly due to insufficient reasoning capabilities and the scarcity of high-quality training data. We introduce **AURA** (*Automated Unified Reasoning for Burn Assessment*), a multi-modal approach that integrates pre-trained VLMs with symbolic first-order logic (FOL) reasoning to improve diagnostic accuracy and interpretability in this data-limited setting. For this study, we collected real-patient data over a one-year period at a U.S. burn center, performing all experiments in a real clinical setting to ensure practical relevance. The dataset includes both conventional B-Mode ultrasound and Tissue Doppler Imaging (TDI), with TDI introduced here for the first time in burn assessment, underscoring the emerging nature of this work. Beyond burn severity classification, we assess the system's ability to produce expert-level surgical insight directly from imaging data. On the retrospective dataset, it achieves up to 93% accuracy in surgical classification and 87% in fine-grained burn depth prediction, comparable to expert-informed predictions and substantially exceeding the 70% accuracy of traditional visual inspection by human experts. These results, obtained from a novel multi-modal dataset collected in a real clinical burn center setting, highlight the potential of this approach to improve decision-making in burn care. To further support future deployment, we demonstrate a prototype integration with an Electronic Medical Record (EMR) system that aligns with clinical workflows and supports scalable, real-world implementation.

Introduction

Burn injuries are a major global health concern, often requiring rapid and accurate assessment to guide surgical intervention and treatment planning. The severity and depth of burns directly influence clinical decisions, including the need for surgery, the extent of debridement, and long-term recovery strategies. Current diagnostic practice in many burn centers relies primarily on expert visual inspection, which is subjective and can lead to inconsistent outcomes, particularly in borderline cases. Objective assessment methods, such as imaging-based analysis, have the potential to improve diagnostic consistency and support decision-making. Ultrasound imaging offers a portable, non-invasive, and relatively low-cost modality for evaluating burn injuries. Conventional B-Mode ultrasound has been investigated in burn care, but its diagnostic use remains limited. Tissue Doppler Imaging (TDI) (Ho and Solomon 2006), commonly applied in cardiology and musculoskeletal assessment, has not previously been explored for burn severity evaluation. Introducing TDI into burn assessment expands the range of measurable tissue properties, potentially enabling more nuanced severity classification.

Despite these opportunities, the application of artificial intelligence (AI) to burn imaging faces two key challenges. First, the availability of domain-specific datasets is limited, particularly for emerging modalities such as TDI, which restricts the use of fully supervised approaches. Second, existing vision-language models (VLMs) (Achiam et al. 2023; Touvron et al. 2023; Liu et al. 2023, 2024; Radford et al. 2019, 2021; Li et al. 2023; Zhang et al. 2024b; Li et al.

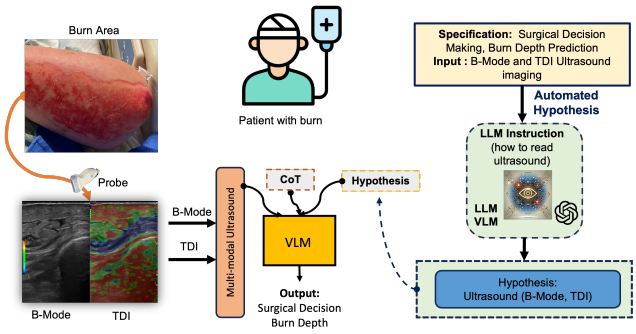

Figure 1: Overview of the proposed framework for burn depth assessment. The system takes multi-modal ultrasound inputs, including B-mode and Tissue Doppler Imaging (TDI), from the burn site. Structured diagnostic hypotheses are automatically generated by a large language model (LLM). These hypotheses guide the vision–language model (VLM) through chain-of-thought (CoT) reasoning to produce interpretable outputs for surgical decision-making and fine-grained burn depth prediction.

2024; Guo et al. 2024; Zhang et al. 2024a; Shakeri et al. 2024), while successful in general computer vision and natural language tasks, often lack the structured reasoning capability required to integrate complex visual and clinical information into interpretable diagnostic outputs (Wei et al. 2022; Wang et al. 2023; Chen et al. 2024; Gu et al. 2024). These limitations are further amplified in emerging domains like ultrasound-based burn care (Tuncer et al. 2024), where imaging protocols are evolving and modality-specific datasets are rare.

To address these challenges, we propose **AURA** (***A**tomated **U**nified **Rea**soning for Burn Assessment*), a multi-modal approach that adapts pre-trained, general-purpose VLMs to medical imaging tasks by incorporating symbolic first-order logic (FOL) reasoning (see Figure 1). The method processes detailed textual descriptions of imaging conditions, modalities, and patient context alongside visual data to generate diagnostic hypotheses and corresponding FOL premises that encode clinical rationale. These premises are validated using an SMT solver (de Moura and Bjørner 2008) to detect and resolve inconsistencies, enabling iterative refinement until a consistent, clinically relevant conclusion is reached.

For this study, we collected a one-year dataset of real-patient ultrasound scans from a U.S. burn center, performing all experiments in a clinical burn care setting. The dataset includes both conventional B-Mode and TDI modalities, marking the first introduction of TDI in burn severity assessment. We evaluate the proposed approach on two clinically relevant tasks: (i) binary classification to determine surgical intervention requirements, and (ii) fine-grained, three-class burn severity classification. On the retrospective dataset, the method achieves up to 93% accuracy in surgical decision-making and 87% in burn depth prediction, comparable to expert-informed predictions and substantially exceeding the $\sim$70% accuracy of traditional visual inspection by human experts. To demonstrate deployment readiness, we also integrate the framework into a prototype Electronic Medical Record (EMR) system, showing its compatibility with real-world clinical workflows.

Methodology

We present **AURA** (*Automated Unified Reasoning for Burn Assessment*), a multi-modal diagnostic reasoning framework that integrates pre-trained Vision-Language Models (VLMs) with symbolic first-order logic (FOL). AURA performs clinical reasoning over ultrasound-based burn imaging, generating interpretable hypotheses and refining diagnostic conclusions through logical consistency checks. The system processes textual descriptions of imaging conditions and clinical protocols alongside visual data to perform both hypothesis generation and downstream classification.

Automated Hypothesis Generation

To enable structured reasoning without relying exclusively on human expertise, we introduce an automated hypothesis generation module (Figure 2). This component transforms clinical knowledge and experimental details into machine-readable hypotheses that guide the VLM in interpreting ultrasound data for burn depth classification.

The process begins by constructing a prompt that fuses two textual contexts: the experimental setup and the clinical interpretation of imaging cues. Let $\mathcal{D}_{\text{exp}}$ represent descriptive details of imaging modalities (e.g., "TDI provides color-coded velocity maps; B-mode offers structural tissue layers"), and $\mathcal{D}_{\text{clin}}$ capture clinical heuristics (e.g., "dominant blue regions in TDI and disrupted layers in B-mode correlate with full-thickness burns"). These are combined as:

$$p = \text{PromptBuilder}(\mathcal{D}_{\text{exp}}, \mathcal{D}_{\text{clin}}),$$

which yields a structured query supplying the language model with sufficient background knowledge.

Given this prompt p, a large language model M_θ generates both a natural-language hypothesis h and a set of first-order logic (FOL) premises $\Phi = \{\phi_1, \phi_2, \ldots, \phi_K\}$ that encode specific diagnostic rules. Sampling parameters such as temperature and top-p sampling are varied to promote diverse rule generation.

To validate the FOL premises, we employ the Z3 SMT solver (de Moura and Bjørner 2008), which ensures logical consistency. The solver iteratively removes contradictions and guides refinement. Once a consistent set is reached or a maximum iteration threshold is met, the remaining premises are summarized into a final natural-language hypothesis. A typical output might be: *"Based on the presence of dominant blue regions in TDI and discontinuous layers in B-mode, the burn is indicative of full-thickness injury and may require surgical intervention."*

This automated pipeline allows the system to dynamically generate domain-specific reasoning without relying on manual annotations or handcrafted logic, enhancing interpretability and diagnostic performance.

Hypothesis and Logical Premise Generation

Given the unified prompt p, the model M_θ produces both a hypothesis h and a corresponding FOL rule set Φ:

$$(h, \Phi) = M_\theta(p \mid \tau, p_{\text{top}}),$$

where τ and p_{top} control generation diversity.

Consistency Verification via SMT Solver

To verify logical consistency, the set Φ is evaluated using an SMT solver. If inconsistencies are detected, a refinement loop is triggered:

$$\Phi^{(\ell+1)} = \Gamma\left(M_\theta\left(\text{RefinePrompt}(p, \Phi^{(\ell)})\right)\right),$$

repeating until a consistent set is obtained or iteration limits are reached.

Final Hypothesis Generation

Validated FOL rules are summarized into the final natural-language hypothesis: *"Based on the dominant blue pattern in TDI and discontinuous layers in B-mode imaging, the burn is likely full-thickness, suggesting surgical intervention may be necessary."*

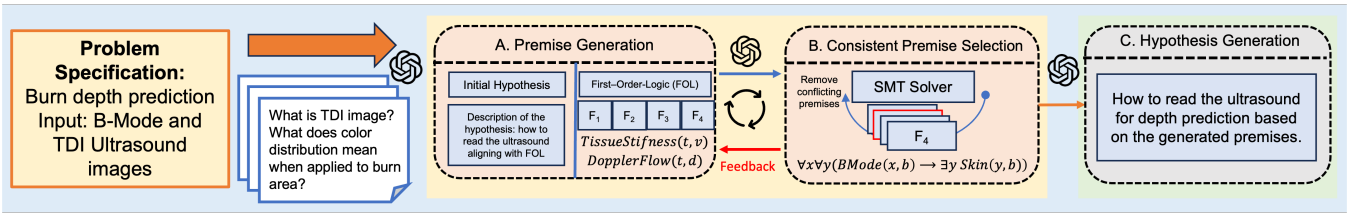

Figure 2: **Automated hypothesis generation pipeline.** The AURA framework initiates diagnostic reasoning by constructing an input prompt that combines experimental descriptions with clinical context. (A) A pre-trained language model generates initial diagnostic hypotheses along with first-order logic (FOL) premises describing associations between ultrasound imaging features and burn severity. (B) An SMT solver checks the logical consistency of these premises, filtering out contradictions and iteratively refining the rule set. (C) The validated premises are summarized into a final natural-language hypothesis, which serves as structured guidance for downstream vision-language burn classification tasks.

Downstream Classification Tasks

Each ultrasound sample $x_i = (x_i^{\mathrm{TDI}}, x_i^{\mathrm{B}})$ is converted into a composite RGB image $z_i \in \mathbb{R}^{H\times W\times 3}$. Classification uses both visual and symbolic signals:

Binary classification:

$$\hat{y}_i = \arg\max_{y\in\{0,1\}} \{P(y \mid z_i) + \alpha\, \mathcal{S}(h, \Phi, y)\}\,,$$

Multi-class classification:

$$\hat{c}_i = \arg\max_{c\in\{1,2,3\}} \{P(c \mid z_i) + \beta\, \mathcal{S}(h, \Phi, c)\}\,.$$

Classification Variants with Hypothesis Integration

We experiment with three variants that integrate h:

1. Hypothesis+VLM:

$$\hat{y}_i = f_{\mathrm{VLM}}(z_i, h) = \arg\max_{y\in\mathcal{Y}} P(y \mid z_i, h).$$

2. Chain-of-Thought (CoT):

$$r^{(t)} = M_\theta(z_i, h, r^{(1)}, \ldots, r^{(t-1)}), \quad \hat{y}_i = f_{\mathrm{VLM}}^{\mathrm{CoT}}(z_i, h, r).$$

3. CoT with Self-Consistency:

$$\hat{y}_i = \mathrm{Aggregate}\left(\{\hat{y}_i^{(k)}\}_{k=1}^{K}\right),$$

using majority vote or averaging over multiple completions.

Logical Support Function

The alignment score $\mathcal{S}(h, \Phi, y)$ quantifies how well the hypothesis and rules support a given label y, using VLM-based responses to queries such as: "Given the diagnostic hypothesis: [h] and premises: [Φ], to what extent does this support the diagnosis of [y]?" The VLM's answer is mapped to a numeric value used in classification.

Experiments

Dataset and Experimental Setup

We evaluate AURA on a retrospective dataset collected over a one-year period at Eskenazi Burn Center, Indianapolis, **Trial Registration: NCT05167461**[1]. To the best of our knowledge, this is the **first dataset** to combine **Tissue Doppler Imaging (TDI)** and **B-Mode ultrasound** for **burn depth assessment**, enabling multi-modal reasoning beyond traditional RGB-based approaches. The dataset includes ultrasound data from 29 patients with histologically or clinically confirmed burn injuries, covering the full spectrum of severity: superficial, superficial partial-thickness, deep partial-thickness, and full-thickness (third-degree) burns. Ground-truth labels were assigned via histopathology when available (5 cases) or determined through consensus by board-certified burn surgeons.

Each ultrasound sample contains both B-Mode frames, capturing structural echogenicity, and TDI frames, encoding perfusion-sensitive velocity information using pseudo-color. To ensure quality, we retained only TDI frames flagged as diagnostically optimal by the acquisition system—identified by green markers indicating proper probe alignment and coupling. From the raw sequences, 950 high-quality frames were extracted and uniformly subsampled to reduce redundancy and preserve scene diversity, resulting in 324 unique frames for downstream analysis. Of these, 130 frames from 15 subjects are held out for evaluation, while the remaining are used for few-shot prompts, chain-of-thought demonstrations, and calibration examples.

While bedside digital photographs of the burn sites are included, they are not temporally aligned with ultrasound frames. These single still images are processed independently by the VLM, and their outputs are fused with ultrasound-based reasoning at the decision level. Imaging parameters such as probe frequency and TDI velocity ranges follow clinical best practices and are documented in the supplementary material. Representative examples are shown in Figure 3, illustrating the complementary information provided by the photographic and ultrasound modalities.

Hypothesis Generation and Vision–Language Models. For automatic diagnostic hypothesis generation, we use OpenAI's `o3-mini-high`, a compact LLM tuned for symbolic reasoning and logical chaining. Both expert-curated and automatically generated hypotheses are evaluated within our framework to assess reasoning quality. For vision–language inference, we utilize multiple foundation models, including `gpt-4o`, `gpt-4o-mini`, `gpt-4-turbo`, `gemini-2.0-flash`,

[1] https://clinicaltrials.gov/study/NCT05167461

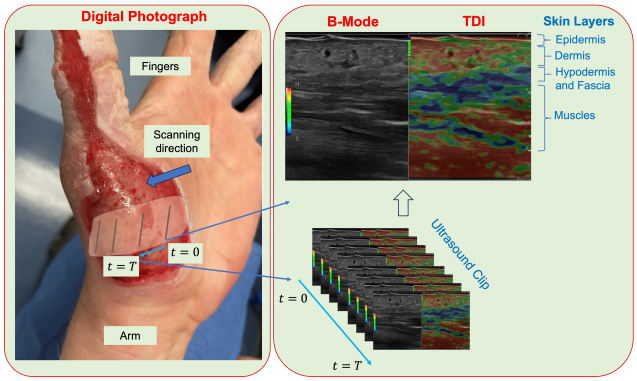

Figure 3: A data sample from the burn dataset. Ultrasound B-Mode and TDI are generated from wound.

and `gemini-1.5-flash`, selected for their strong multimodal reasoning capabilities, low latency, and compatibility with structured prompting involving both visual and textual inputs.

Unless otherwise specified, all experiments are conducted under zero-shot or few-shot settings. To isolate the effect of structured reasoning, we benchmark model performance under identical input conditions using both expert-authored and automatically generated hypotheses as scaffolds for downstream prediction.

Evaluation Design

We evaluate AURA in three structured diagnostic tasks, each designed to assess the framework's reasoning capabilities and its alignment with clinical expertise. Two of these settings rely on automatically generated hypotheses, one for binary surgical decision-making and another for fine-grained classification, while the third offers a comparative evaluation against expert-authored reasoning. This design isolates the impact of symbolic hypothesis generation across varied levels of diagnostic complexity.

Surgical Decision-Making with Automated Hypotheses. The first task evaluates binary classification: whether surgical intervention is needed, based solely on B-Mode and Tissue Doppler Imaging (TDI) inputs. Diagnostic hypotheses are generated automatically using modality-specific prompts. This setting assesses the framework's ability to support critical clinical decisions with no expert involvement.

Fine-Grained Burn Depth Classification with Automated Hypotheses. The second task addresses three-way classification, distinguishing superficial second-degree, deep second-degree, and third-degree burns, using ultrasound inputs alone. Hypotheses are generated entirely by the automated pipeline. This experiment probes AURA's capacity to resolve subtle structural and perfusion differences across similar burn types.

Comparative Evaluation: Expert vs. Automated Hypotheses. To evaluate the quality of automated diagnostic reasoning, we conduct a direct comparison with structured hypotheses written by board-certified burn surgeons. Using the same ultrasound inputs as in the surgical decision task, we compare VLM predictions driven by expert- versus LLM-generated reasoning. This isolates the contribution of hypothesis source to final prediction quality.

Implementation Details. All experiments employ chain-of-thought prompting and self-consistency decoding (Wei et al. 2022; Wang et al. 2023). The VLM is queried multiple times per input with varied temperature (0.5–1.0) and top-p (0.5–1.0) sampling. Chain-of-thought exemplars are randomly ordered and sampled to encourage diverse reasoning paths. Final predictions are determined by majority vote over the model outputs.

Results

We report findings across three evaluation settings aligned with our proposed experimental design. Results assess (1) surgical decision classification using automated hypotheses, (2) fine-grained burn depth classification using automated hypotheses, and (3) a comparative evaluation between automated and expert-generated hypotheses in surgical settings.

Surgical Decision with Automated Hypotheses Table 1 presents binary surgical decision performance when guided solely by automatically generated hypotheses. The highest-performing model, `GPT-4o`, achieves 93% accuracy and 0.93 F1-score, significantly outperforming its unguided base version (33% accuracy). Similarly, `GPT-4 Turbo` achieves 93% accuracy, confirming that structured diagnostic guidance provides consistent benefits.

Smaller and faster models also see major gains. For instance, `GPT-4o-mini` improves from 67% to 80% accuracy, and `Gemini 2.0` improves from 47% to 87%.

Table 1: Surgical decision performance with automatically generated hypotheses.

Model	Accuracy	F1	Precision	Recall
GPT-4o + Auto Hypothesis	**93%**	**0.93**	**0.94**	**0.93**
GPT-4o (Base)	33%	0.17	0.11	0.33
GPT-4o-mini + Auto Hypothesis	80%	0.77	0.85	0.80
GPT-4o-mini (Base)	67%	0.67	0.69	0.67
GPT-4 Turbo + Auto Hypothesis	**93%**	**0.93**	**0.94**	**0.93**
GPT-4 Turbo (Base)	87%	0.87	0.87	0.87
Gemini 2.0 + Auto Hypothesis	87%	0.86	0.89	0.83
Gemini 2.0 (Base)	47%	0.41	0.79	0.47
Gemini 1.5 + Auto Hypothesis	80%	0.79	0.85	0.80
Gemini 1.5 (Base)	60%	0.50	0.42	0.60

Fine-Grained Burn Classification with Automated Hypotheses Table 2 shows model performance on three-class burn depth classification (second-degree superficial, second-degree deep, and third-degree). Once again, hypothesis-guided models substantially outperform their base variants.

`GPT-4o` leads with 87% accuracy, while `Gemini 1.5` improves from 47% to 67%. However, some base models (e.g., `GPT-4o-mini`) perform competitively without hypothesis support, suggesting inherent capabilities for moderate-granularity distinctions.

Expert-Guided vs. Automated Hypotheses (Surgical Decision) As part of our comparative evaluation, we assess

Table 2: Fine-grained burn classification using automated hypotheses.

Model	Accuracy	F1	Precision	Recall
GPT-4o + Auto Hypothesis	**87%**	**0.87**	**0.87**	**0.87**
GPT-4o (Base)	27%	0.27	0.34	0.27
GPT-4o-mini + Auto Hypothesis	53%	0.42	0.53	0.53
GPT-4o-mini (Base)	73%	0.71	0.73	0.73
GPT-4 Turbo + Auto Hypothesis	53%	0.52	0.56	0.53
GPT-4 Turbo (Base)	60%	0.59	0.62	0.60
Gemini 2.0 + Auto Hypothesis	60%	0.50	0.64	0.60
Gemini 2.0 (Base)	47%	0.46	0.60	0.47
Gemini 1.5 + Auto Hypothesis	67%	0.62	0.79	0.67
Gemini 1.5 (Base)	47%	0.43	0.46	0.47

how the best-performing automated setup (GPT-4o with auto-generated hypotheses) fares against expert-written hypotheses. Results are shown in Table 3.

While the expert-guided model achieves slightly higher accuracy (95% vs. 93%) and perfect recall, the automated hypothesis system demonstrates nearly equivalent performance across all metrics. Specifically, GPT-4o with auto hypotheses achieves a 0.93 F1-score and 0.94 precision—matching or closely approaching expert-level diagnostics.

This minimal performance gap highlights the clinical potential of automated reasoning to deliver expert-aligned outputs, especially when expert curation is not feasible in real-time or resource-limited environments.

Table 3: Comparison between expert-guided and automated hypotheses (GPT-4o) on surgical decision classification.

Method	Accuracy	F1	Precision	Recall
GPT-4o + Expert Hypothesis	**95%**	**0.95**	**0.94**	**1.00**
GPT-4o + Auto Hypothesis	93%	0.93	0.94	0.93

Qualitative Analysis. Figure 4 showcases a representative example where multimodal cross-reasoning plays a critical role. The input consists of co-registered B-mode and TDI ultrasound frames. The hypothesis-based framework leverages a structured natural language explanation that explicitly links B-mode indicators of tissue integrity with perfusion patterns visible in TDI—particularly the presence of blue hues in subcutaneous regions, which signify preserved blood flow and tissue viability. The vision-language model, guided by these hypotheses, performs coherent cross-modal alignment, correctly inferring a deep partial-thickness burn that does not warrant surgery. This contrasts sharply with the base GPT-4o output, which lacks structured reasoning and instead defaults to a coarse interpretation of visual cues. As a result, it erroneously predicts a third-degree (surgical) burn by misinterpreting surface texture and depth cues.

This example illustrates how integrating symbolic reasoning enables the VLM to synthesize insights across structural and perfusion modalities—an essential capability for nuanced clinical assessments.

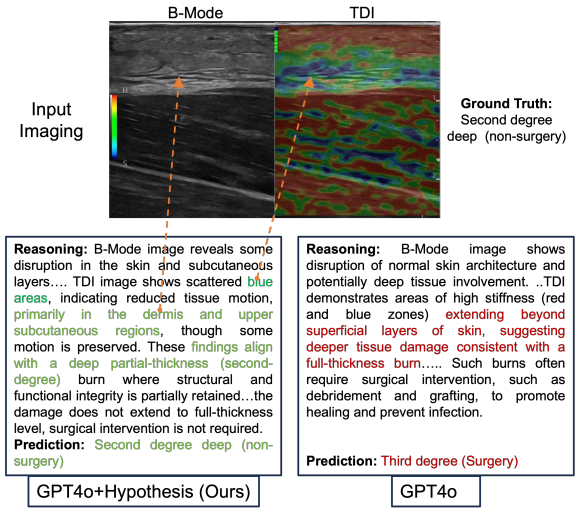

Figure 4: Qualitative comparison between the hypothesis-based framework (left) and base GPT-4o (right) on multimodal ultrasound input. The hypothesis-guided model integrates structural (B-mode) and perfusion (TDI) cues to reason that blue-coded regions localize within the dermis and subcutaneous tissue, supporting a deep partial-thickness (non-surgical) diagnosis. In contrast, GPT-4o without structured reasoning overestimates severity based solely on surface features, yielding a false third-degree burn prediction.

Deployment

We designed the proposed system with practical clinical deployment in mind, focusing on real-world feasibility, interpretability, and integration into decision-support workflows. While not yet deployed in live hospital settings, the system was tested retrospectively using data collected in authentic clinical environments and integrated into a simulated Electronic Medical Record (EMR) interface to demonstrate deployment readiness. Below, we describe key aspects relevant to the deployment.

Integration with Clinical Workflow

We integrated the AI framework with *DrChrono* (EverHealth 2025), a commercial EMR platform. Using a training account provided for medical professionals, we developed a Python-based application that interacts with the DrChrono Developer API to upload clinical imaging data, such as digital photographs and ultrasound videos, and retrieve corresponding AI-driven diagnostic predictions. The system operates through a Web-based API, which ensures compatibility with mobile and tablet platforms. This demonstrates technical feasibility and supports future deployment in real-time clinical or telemedicine environments. An illustration is given in Figure 5.

Hardware and Infrastructure

Although ultrasound is not yet a standard diagnostic tool in burn care, we demonstrated feasibility by collecting B-mode and Tissue Doppler Imaging (TDI) data using the LOGIQ E9 ultrasound system (GE Healthcare Technologies Inc.)

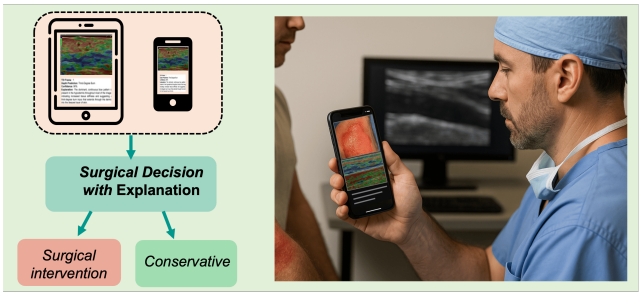

Figure 5: Illustration of the AI-enabled decision support system for burn diagnosis integrated with an EMR platform. The system retrieves imaging data (e.g., ultrasound and digital photographs) from the EMR, performs real-time reasoning using proposed AI model, and returns surgical triage recommendations alongside natural language explanations with confidence scores. The mobile-ready interface enables clinicians to review diagnostic reasoning at the point of care, supporting expert decision-making in both hospital and remote settings.

with a 16 MHz probe. All scans followed standard clinical procedures, and data collection was conducted in real clinical environments, although not during real-time hospital operations. This study is the first to propose and evaluate the use of TDI for burn diagnosis, highlighting its utility in assessing deep tissue injuries in severe burn cases.

Human-in-the-Loop and Decision Support

The system is explicitly intended to support clinical experts by serving as a decision support tool. It provides interpretable justifications for every diagnostic output, which can help guide human experts in high-risk or ambiguous cases. The design allows for expert override at all stages of the workflow, reinforcing safe and collaborative AI-assisted medical decision-making.

Findings and Practical Insights

- The system provides near real-time diagnosis, with end-to-end inference typically completed in under one minute. This includes both EMR data processing and AI inference.
- Digital photographs proved effective for identifying superficial burns, whereas ultrasound imaging, especially TDI, was crucial for accurate assessment of deep burn injuries.
- The framework integrates diverse imaging modalities to produce coherent and interpretable diagnostic outputs, reducing reliance on expert presence at the point of care.
- Diagnostic outputs are expressed in natural language with associated confidence scores in percentage format, supporting clinical interpretability and trust.
- The modular inference pipeline enables flexible cost-aware diagnostics. For example, ultrasound imaging may be skipped when superficial burns are confidently detected from photographs.
- The architecture is suitable for deployment in low-resource environments and can be extended easily to mobile or tablet-based platforms via its API design.

Remaining Challenges and Deployment Path

The current system relies on API-based access to vision-language models, but it is compatible with open-source or in-house alternatives. Since the framework only requires model inference rather than training, it can be deployed using standard GPUs (e.g., with 16GB memory), enabling cost-effective implementation in clinical or remote environments without specialized hardware. Although Tissue Doppler Imaging (TDI) is not widely used in burn care, this study demonstrates its feasibility using conventional ultrasound systems, potentially encouraging its adoption.

Conclusion

This work introduces a framework that combines automated diagnostic hypothesis generation with vision-language models (VLMs) to enable robust multimodal reasoning over medical imaging for burn severity assessment. By incorporating structured first-order logic to guide visual interpretation, the system performs expert-level surgical triage and fine-grained classification directly from raw clinical imaging without requiring large-scale annotated datasets. The framework was evaluated on a real-patient dataset collected at a U.S. burn center, achieving up to 93% accuracy in surgical classification and 87% in burn depth prediction. These results significantly surpass the accuracy of traditional clinical practice based on human visual inspection. The findings demonstrate the system's ability to integrate heterogeneous modalities, including B-Mode ultrasound and Tissue Doppler Imaging, into a coherent diagnostic reasoning process that enhances both reliability and interpretability in clinical decision-making. The paper also presents a deployment-oriented system design, including EMR integration, interface prototyping, and performance considerations for real-world implementation. Ultimately, this work highlights the transformative potential of automated multimodal reasoning to bridge the gap between general-purpose foundation models and domain-specific medical applications. It offers a scalable and clinically meaningful approach for deploying AI in real-world burn care, particularly in data-limited clinical settings.

Acknowledgments

This work was supported by the Office of the Assistant Secretary of Defense for Health Affairs under Award No. W81XWH-21-2-0030 and by NIH under Grant No. 5R21LM013711-02. We thank burn surgeons Brett C. Hartman, MD, and Leigh Spera, MD, from the Eskenazi Burn Center and the Division of Plastic Surgery at the Indiana University School of Medicine, Indianapolis, IN, for their contributions and assistance in facilitating data collection from patients with burn.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; and OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Chen, X.; Chi, R. A.; Wang, X.; and Zhou, D. 2024. Premise Order Matters in Reasoning with Large Language Models. In *Forty-first International Conference on Machine Learning*.

de Moura, L. M.; and Bjørner, N. S. 2008. Z3: An Efficient SMT Solver. In *International Conference on Tools and Algorithms for Construction and Analysis of Systems*.

EverHealth. 2025. DrChrono. Electronic health record software.

Gu, B.; Desai, R. J.; Lin, K. J.; and Yang, J. 2024. Probabilistic medical predictions of large language models. *npj Digital Medicine*, 7(1): 367.

Guo, Y.; Zeng, X.; Zeng, P.; Fei, Y.; Wen, L.; Zhou, J.; and Wang, Y. 2024. Common Vision-Language Attention for Text-Guided Medical Image Segmentation of Pneumonia. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 192–201. Springer.

Ho, C. Y.; and Solomon, S. D. 2006. A clinician's guide to tissue Doppler imaging. *Circulation*, 113: e396–e398.

Li, C.; Wong, C.; Zhang, S.; Usuyama, N.; Liu, H.; Yang, J.; Naumann, T.; Poon, H.; and Gao, J. 2023. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *arXiv preprint arXiv:2306.00890*.

Li, Q.; Yan, X.; Xu, J.; Yuan, R.; Zhang, Y.; Feng, R.; Shen, Q.; Zhang, X.; and Wang, S. 2024. Anatomical structure-guided medical vision-language pre-training. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 80–90. Springer.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26296–26306.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual Instruction Tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1: 9.

Shakeri, F.; Huang, Y.; Silva-Rodríguez, J.; Bahig, H.; Tang, A.; Dolz, J.; and Ben Ayed, I. 2024. Few-shot adaptation of medical vision-language models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 553–563. Springer.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; and Lample, G. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Tuncer, H. B.; Akın, M.; Çakırca, M.; Erkılıç, E.; Yıldız, H. F.; and Yastı, A. Ç. 2024. Do pre-burn center management algorithms work? Evaluation of pre-admission diagnosis and treatment adequacy of burn patients referred to a burn center. *Journal of Burn Care & Research*, 45(1): 180–189.

Wang, X.; Wei, J.; Schuurmans, D.; Le, Q. V.; Chi, E. H.; Narang, S.; Chowdhery, A.; and Zhou, D. 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *The Eleventh International Conference on Learning Representations*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Zhang, J.; Wang, G.; Kalra, M. K.; and Yan, P. 2024a. Disease-informed adaptation of vision-language models. *IEEE Transactions on Medical Imaging*.

Zhang, S.; Xu, Y.; Usuyama, N.; Xu, H.; Bagga, J.; Tinn, R.; Preston, S.; Rao, R.; Wei, M.; Valluri, N.; Wong, C.; Tupini, A.; Wang, Y.; Mazzola, M.; Shukla, S.; Liden, L.; Gao, J.; Crabtree, A.; Piening, B.; Bifulco, C.; Lungren, M. P.; Naumann, T.; Wang, S.; and Poon, H. 2024b. A Multimodal Biomedical Foundation Model Trained from Fifteen Million Image–Text Pairs. *NEJM AI*, 2(1).