UNIVERSITY OF TURKU

Multimodal Artificial Intelligence Model for Wound Care

Medical Imaging
Master's Degree Programme in Biomedical Engineering and Health Technology
Department of Computing, Faculty of Technology
Master of Science in Technology Thesis

Author:
Benjamin Pörhö

Supervisors:
Assoc. Prof. Antti Airola
Asst. Prof. Jan Akmal
Adj. Prof. Tatu Rojalin
Prof. Marjo Yliperttula

26.8.2025
Turku

Master's thesis
Subject: Biomedical Engineering and Health Technology
Author(s): Benjamin Pörhö
Title: Multimodal Artificial Intelligence Model for Wound Care
Supervisor(s): Assoc. Prof. Antti Airola, Asst. Prof. Jan Akmal, Adj. Prof. Tatu Rojalin, Prof. Marjo Yliperttula
Number of pages: 56
Date: 26.8.2025

Chronic wounds, particularly diabetic wounds represent a significant clinical and economic burden due to prolonged healing times and high rates of complications. Artificial intelligence (AI) is a promising approach to support wound assessment and guide treatment decisions. However, most existing models lean on single-modality inputs. This thesis presents a proof-of-concept (POC) multimodal AI model that combines RGB images, thermal images, and wound area measurements to predict treatment responses on wounds treated with nanofibrillated cellulose (NFC) hydrogel.

Biological mechanisms of wound healing, conventional and advanced wound care, and prior AI-based methods in wound care were reviewed. The hypothesis was defined: *A multimodal AI model integrating RGB images, thermal images, and wound area measurements would succeed in classifying the post operative day (POD), of an input wound. Contribution of different modalities would vary, RGB images providing the most predictive value for the model, and the thermal images adding confusion due to their lower quality and homogeneous nature in the later PODs. Still the multimodal model would achieve higher accuracies in predicting treatment outcomes compared to the unimodal approach*. The specific aims were to design a three-branch neural network combining convolutional feature extraction for RGB and thermal images with a fully connected branch for wound area data. Another aim was to evaluate the contribution of each modality through ablation experiment and fusion strategies, such as concatenation, attention based fusion and weighted static fusion. Finally, Grad-CAM interpretability techniques were applied to visualize the model's decision-making.

Dataset details, preprocessing techniques, and cross-validation strategy were covered along with performance metrics including accuracy, precision, recall, and F1-score. Model was implemented using TensorFlow, and hyperparameter optimization was performed with KerasTuner. Classes were defined by POD, although this label introduced impracticalities due to varying individual healing trajectories.

Results demonstrated that models trained without strict subject-level splitting showed inflated performance, underscoring the necessity of partitioning by subject (wound). In an ablation study, RGB-only model achieved a classification accuracy of 0.70, combining wound area improved accuracy to 0.75, indicating that explicit size context improves predictions. Adding thermal images to form a three-input model provided a slight decrease in overall accuracy (~0.73), suggesting that thermal data is most informative during early days, but may introduce noise later in the timeframe. Grad-CAM visualizations confirmed that the model was focusing on healing relevant features, rather than noise or image specific features. Exploring different fusion strategies revealed that the weighted static fusion model provided the highest accuracies, reaching testing accuracy of ~0.85 with the same subject that was held out in the ablation study. KerasTuner hyperparameter tuning process involving nested Leave-One-Group-Out Cross-Validation (LOGO CV) reached an average accuracy of 0.71.

Key limitations included the small, homogeneous dataset, and the use of POD as the main label for healing stage, which did not align uniformly with biological wound progression. Future work should utilize clinically relevant labels, focus on high-quality RGB images, precise sizing, and careful integration of thermal cues, avoiding confusing the model. This multimodal POC model introduced a scalable framework for AI-driven wound assessment, highlighting best practices in data handling, model architecture, and interpretability. By addressing these considerations, AI-based wound care tools may achieve greater reliability and clinical relevance when extended to human datasets.

Key words: multimodal, artificial intelligence, machine learning, deep learning, wound care, nanofibrillated cellulose hydrogel, thermal imaging, medical imaging

Table of contents

Abbreviations

Preface

1 Introduction 1

1.1 Background 1

1.2 Research Problem 2

1.3 Scope of the Thesis 3

1.4 Structure of the Thesis 5

1.5 Generative AI-based Tools 5

2 Artificial Intelligence in Wound Care 6

2.1 Wounds, Wound Classification and Wound Care 6

2.2 Artificial Intelligence 10

2.3 Application to Wound Care 16

2.4 About the Data 19

2.4.1 RGB Images 19

2.4.2 Thermal Images 20

2.4.3 Wound Area 21

2.4.4 Amount of Data 21

3 Hypothesis and Aims 24

4 Materials and Methods 25

4.1 Data 25

4.2 Preprocessing 27

4.3 Evaluation Metrics and Methods 29

4.3.1 Cross-Validation Strategy 29

4.3.2 Performance Metrics 30

4.3.3 Model Interpretability 32

4.4 Technologies 33

4.4.1 TensorFlow 33

4.4.2 KerasTuner 33

4.4.3 Environment and Other Libraries 34

5 Results 36

5.1 Healing Scale Model Architecture 36

5.2 Base Model 37

5.3 Weighted Fusion Strategies 41

6 Discussion 50

6.1 Classification Challenges 50

6.2 Model Performance 51

6.3 Future and Recommendations 53
7 Conclusions 56
References 57

Abbreviations

.CR2	Canon RAW image format extension
Adam	Adaptive Moment Estimation
AI	Artificial Intelligence
API	Application Programming Interface
AUC	Area Under the Curve
AUROC	Area Under the Receiver Operating Characteristic Curve
CCD	Charge-Coupled Device
CMOS	Complementary Metal-Oxide-Semiconductor
CNN	Convolutional Neural Network
CT	Computerized Tomography
CV	Cross-Validation
DFU	Diabetic Foot Ulcer
EHR	Electronic Health Records
FDA	Food and Drug Administration
FOV	Field of View
GLP	Good Laboratory Practice
GPU	Graphics Processing Unit
HSM	Healing Scale Model
IRT	Infrared Thermography
JPEG	Joint Photographic Experts Group image format
KT	KerasTuner
LOGO CV	Leave-One-Group-Out Cross-Validation
LOOCV	Leave-One-Out Cross-Validation
MAE	Mean Absolute Error
ML	Machine Learning
MRI	Magnetic Resonance Imaging
NaCl	Sodium Chloride
NFC	Nanofibrillated Cellulose
NLP	Natural Language Processing
PCA	Principal Component Analysis
PDF	Portable Document Format
POC	Proof-of-concept
POD	Post Operative Day
POSAS	Patient and Observer Scar Assessment Scale
PRP	Platelet-Rich Plasma
ReLU	Rectified Linear Unit
ResNet	Residual Network
RGB	Red Green Blue
SVM	Support Vector Machine
TEWL	Transepidermal Water Loss
TPU	Tensor Processing Unit
VGG	Visual Geometry Group
ViT	Vision Transformers

Preface

I am grateful to all those who have guided and supported me throughout this study. First, I would like to thank Tatu Rojalin for our weekly meetings and long brainstorming sessions, which helped me to clarify my ideas and keep my work on track. Tatu was a real game changer for this project.

Special thanks to Antti Airola for guidance that strengthened the methodological rigor and credibility of this study. I also want to thank Jan Akmal for each discussion, which brought valuable insights and fresh perspectives to this study.

Thanks to Elle Koivunotko for providing the data that made this study possible. Finally, I am thankful to Marjo Yliperttula for providing this whole opportunity and guiding this project to success.

Helsinki, August 26th, 2025

Benjamin Pörhö

1 Introduction

1.1 Background

Chronic wounds associated with diabetes represent a significant challenge in healthcare. Wound healing is a complex process that can be easily disrupted by local or systematic factors, resulting in the formation of chronic wounds. In the United States alone, chronic wounds affect an estimated 8.2–10.5 million people (approximately 2.5% of the population) and are expected to increase with aging and diabetes prevalence [1]. The global prevalence of chronic wounds in general has been estimated at 1.5–2.2/1000 population and the numbers are constantly increasing [2]. Diabetic foot ulcers (DFUs) are a common chronic wound in patients with diabetes, with a global prevalence of approx. 6.3% [3]. This means that there are approximately 9–26 million new DFUs worldwide each year [4]. Alarmingly, individuals with diabetes have a 15–34% lifetime risk of developing foot ulcers, and these ulcers precede approximately 80% of diabetes related lower limb amputations [4], [5]. The financial burden is real; chronic wounds are estimated to cost tens of billions of dollars annually. For example, Medicare spending on chronic wound care in the U.S. has been estimated to be between 28 and 97 billion dollars per year, with diabetic foot ulcers alone accounting for up to approximately 18.7 billion dollars [6]. Such wounds significantly degrade quality of life and can lead to severe complications if not managed effectively.

These statistics highlight the urgent need for scalable and innovative approaches to overcome this challenge, especially in areas with limited access to advanced wound care and diagnostic tools. There is often no clear consensus on the best treatment for a given wound, and clinicians must choose from a wide variety of dressings and interventions [7]. Selecting the appropriate treatment is critical, as an incorrect initial approach can delay healing. If a wound does not show sufficient healing after approximately four weeks of standard care, guidelines suggest re-evaluating the strategy and considering advanced therapies [8]. However, the selection of advanced therapy is often not based on evidence, highlighting the uncertainty of current wound management [8]. This underscores the importance of better decision support tools in wound care. DFUs demonstrates the challenge of chronic wounds. They often heal slowly and are prone to infection; over 50% of DFUs become infected, and approximately 20% of those with moderate to severe infection result in some level of amputation [4]. Even with appropriate therapy, many diabetic ulcers take months to heal, although studies have shown that aggressive, specialized care can achieve approximately 77% healing within one

year [4]. The prolonged healing time contributes to extended suffering, high healthcare utilization, and elevated mortality risk, and a patient with DFU has a 5-year mortality rate twice as high as a diabetic patient without an ulcer [4]. These outcomes highlight the need to improve wound care through better initial treatment selection and ongoing treatment monitoring.

Recently, artificial intelligence (AI) has been studied to facilitate wound treatment and automate wound diagnosis [9]. Computer vision techniques, particularly deep learning with convolutional neural networks (CNNs), have been applied to wound images to automate assessments that traditionally rely on expert eyes [10]. The most recent studies have introduced advanced AI technologies where multimodal inputs are used in model training instead of a single input (e.g. images and tabular biomarker data versus images only). CNN-based models have already been successful in classifying wound types, for example, distinguishing diabetic foot ulcers from pressure ulcers, and even assessing wound severity or stage from wound images [10]. Such image-based classifiers can help wound specialists make faster, more consistent diagnoses and potentially suggest optimal treatments [11]. Researchers have also developed deep learning models for wound segmentation, segmenting the wound area in an image, which enables objective measurement of wound size and healing progress over time [12]. These AI-driven image analyses can reduce the burden on clinicians to provide personalized treatment which is also more repeatable than visual inspection alone. In combination with state-of-the-art wound care methods, these solutions can significantly improve modern chronic wound care.

1.2 Research Problem

Chronic wounds have prolonged healing times, sometimes taking months or years to fully heal if complications arise. During this period, patients may undergo various treatments sequentially, starting with standard dressings and switching to advanced therapies if there is no significant improvement after weeks. Various treatment options are available, including different dressing materials, topical agents, and devices; however, predicting which therapy will work best for a particular wound is challenging [8]. The initial choice of treatment often lacks strong comparative evidence [8]. This trial-and-error approach means that some wounds do not receive optimal therapy until late, thus delaying the healing process. Each additional week a wound remains open increases the risk of infection, hospitalization, or even amputation in patients with diabetes [4]. This study focuses on shortening the trial-and-error

period by better informing initial treatment decisions and ongoing monitoring. It is crucial to address this problem in clinical practice. Choosing the right treatment early could significantly improve patient outcomes, and applying effective therapy from the start may close the wound faster, reducing vulnerability to infection and complications. It could also lower healthcare costs by avoiding ineffective treatments and lengthy or frequent hospital stays.

AI offers a potential solution to this decision-making challenge. AI algorithms can be trained on large amounts of wound data to recognize patterns that may reveal potential healing outcomes. By identifying wounds likely to heal slowly or not at all under certain treatments, an AI tool could alert clinicians to modify the plan early on. A recent machine learning study used electronic health record data from over a million wounds to predict which wounds would fail to heal within 12 weeks, achieving high accuracy (AUC ~0.85) [13]. Berezo et al noted that earlier identification of non-healing wounds via such models "may improve treatment decisions and patient outcomes" by prompting timely interventions [13]. An AI-based decision support system could act as a second pair of eyes, suggesting that patients need more aggressive therapy or specialized care from the beginning. AI-powered diagnostic tools can standardize wound assessment by automatically measuring wound size and analyzing tissue characteristics, providing objective feedback on wound improvement or deterioration. This continuous monitoring can help clinicians decide sooner if a strategy is ineffective. For example, if an AI algorithm analyzing weekly wound images detects stagnation in healing, it could recommend re-evaluating the treatment plan. By detecting non-responding wounds earlier than a human might, such tools would support more proactive care adjustments. In cases of diabetic ulcers, even a few weeks of delay in switching to effective therapy can be the difference between limb salvage and amputation. Therefore, developing AI-based diagnostic and monitoring systems addresses a real clinical need to guide clinicians in choosing the right treatment from the start and in making informed changes along the way, ultimately shortening healing time and improving patient outcomes.

1.3 Scope of the Thesis

This thesis focuses on the development of a proof-of-concept (POC) architecture for an AI model for wound care. The aim was to create a POC model to demonstrate how a multimodal AI approach could assist in treatment decisions for wounds and to propose an architecture that can be scaled to future development.

The scope of this study was narrowed in several ways. There are a wide variety of treatment methods that could be included when training an AI model for wound care; however, suitable available data is a major challenge. This study considered only three types of treatment: nanofibrillated cellulose (NFC) hydrogel with glucose, without glucose, and saline solution as a control. The NFC hydrogel is a medical-grade, plant-derived hydrogel manufactured in compliance with ISO13485 and IS10993 standards, aligned with current sustainability values and directives (EU 2022/2464, EU 2023/2772, EU 2024/1760). The NFC hydrogel is non-toxic, xenon-free, and wood-based, representing a groundbreaking development in this field. Its dressing form has already demonstrated clinical success in treating skin graft donor sites (FibDex®) [7]. While focusing on the two variants of NFC hydrogel treatment, one of the tasks of this model is to predict the outcomes of these specific treatments.

The most significant limitation of this study was the nature of the dataset. The data were from a mouse study; therefore, the wound images were from mice that were treated with one of the two types of NFC hydrogel and the control wounds with saline. Obviously, mouse wounds are not comparable to human wounds in this context; therefore, without re-training the model with a human wound dataset, it may not work for human data. Another limitation of the dataset was its small size and limited diversity. The limited sample size causes challenges in training a reliable deep learning model, as deep networks typically require at least thousands of examples rather than just a few hundred, which was the situation with this dataset. Overfitting is a concern, and careful techniques, such as data augmentation, regularization, and cross-validation, are required to make the most of the data.

This study aimed to evaluate whether a multimodal input (RGB image, thermal image, wound area) AI approach can predict or assess wound healing stage under given constraints. Due to the limited dataset, this thesis was treated as a pilot study. Success is indicated by the model capturing meaningful signals from the images and data that correlate with healing. However, the main aim is to achieve a scalable architecture and promising results to lay the foundation for future development when real human data are available. This thesis also provides a brief comparison between single modality and multimodality, while the model's branches are evaluated both individually and in every possible pair combination. This aligns with the POC nature of this study, demonstrating the predictive value of multimodal inputs for AI in wound care.

1.4 Structure of the Thesis

This thesis begins by introducing the challenges of chronic wounds and the potential role of AI in wound care, defining the research problem of trial-and-error treatment approaches, and stating the scope, structure, and generative AI tools used in this study. Chapter 2 reviews the existing literature on wounds and wound care, including DFU classification, and summarizes AI methods relevant to wound analysis. Prior work is presented on image-only and multimodal AI models for wound care, and three data modalities: RGB images, thermal images, and wound areas, are described.

Chapter 3 includes the main hypothesis that a multimodal AI model integrating RGB images, thermal images, and wound area measurements would succeed in classifying the post operative day (POD), of an input wound. Contribution of different modalities would vary, RGB images providing the most predictive value for the model, and the thermal images adding confusion due to their lower quality and homogeneous nature in the later PODs. Still the multimodal model would achieve higher accuracies in predicting treatment outcomes compared to the unimodal approach. and lists the study aims: to build a POC architecture combining all three modalities, evaluate each modality's contribution, and use Grad-CAM for interpretability. Chapter 4 introduces the materials and methods, including data collection, image labeling, preprocessing steps, cross-validation strategies, and performance metrics. The software frameworks and hardware configurations used for model training are also noted.

Chapter 5 presents the results, including the performance metrics and visualizations. The evaluation of each modality combination is also reported, followed by a focus on the HSM approach with three fusion strategies: simple concatenation, attention based fusion, and weighted static fusion. Chapter 6 discusses the results and addresses challenges, such as data leakage, labeling, and dataset quality. Chapter 7 summarizes the main findings and confirms that the multimodal AI approach is promising as a POC. References are listed at the end.

1.5 Generative AI-based Tools

In this thesis, generative AI tools were utilized to enhance writing quality and code development. Paperpal Prime was used for proofreading, trimming and rephrasing [14]. ChatGPT (o4-mini-high) assisted in code and model architecture design [15], while Google Gemini integrated with Google Colab helped with code debugging [16]. All AI-generated content was critically evaluated and modified to maintain accuracy.

2 Artificial Intelligence in Wound Care

2.1 Wounds, Wound Classification and Wound Care

A wound can be broadly defined as damage to or disruption of the cellular, anatomical, and functional integrity of living tissue [17]. Wounds can be acute or chronic based on their healing behavior [17]. Acute wounds often heal within 4-6 weeks, while chronic wounds fail to progress through normal repair or restore integrity after three months [18], [19]. Some experts define chronicity as non-healing over six weeks [19]. In general, wounds can be classified as traumatic (injury-related), surgical (incisions), thermal (burns), or pathological (due to an underlying disease) [20]. The main chronic wound types are diabetic, arterial, venous, and pressure ulcers, each with their common locations [18]. Chronic wounds share features such as pro-inflammatory cytokines, persistent infection, and senescent cells [18]. They affect 1-2% of developed countries' population and cause substantial treatment costs of tens of billions of dollars annually in the U.S. alone [18].

Wound healing is a complex biological process in which the body repairs damaged tissue. **Figure 1** demonstrates the four overlapping phases: haemostasis, inflammation, proliferation, and remodelling [18]. In acute wounds, these phases occur sequentially to restore skin integrity. Haemostasis (days 0–3) begins after injury (bleeding wound). Haemostasis has two main stages. In the primary stage, blood vessels constrict, platelets adhere to the injured site, and platelets aggregate to form a platelet plug [18], [21]. In the secondary phase of haemostasis, thrombin generated through both extrinsic and intrinsic coagulation pathways transforms fibrinogen into fibrin. This fibrin then combines with platelet clots and blood cells to form a thrombus [21]. During inflammation (days 1–25), immune cells infiltrate the wound, phagocytose bacteria and debris, and release cytokines. This phase shows signs of inflammation, including redness, warmth, swelling, and pain [18]. During proliferation (days 1–25), fibroblasts and endothelial cells generate new tissues through angiogenesis, collagen synthesis, granulation tissue formation, and re-epithelialization [18]. Maturation/remodelling begins around day 20 and continues for weeks to months. Collagen is remodelled from type III to type I, and scar tissue gains strength up to 80% of original skin strength [18].

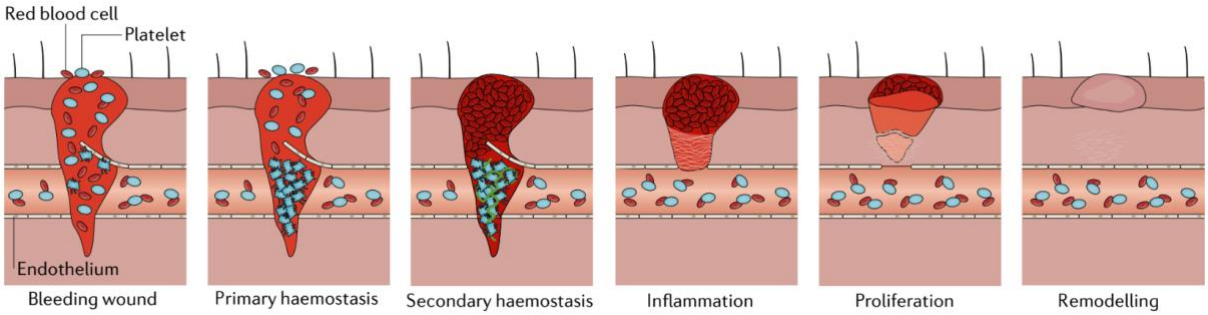

Figure 1 The four consecutive but overlapping phases of wound healing: haemostasis, inflammation, proliferation and remodelling. Haemostasis is divided to two main stages: primary haemostasis and secondary haemostasis. Figure from a review by Guo et al. [21].

These healing phases overlap, and chronic wounds commonly stall in the inflammatory phase, failing to transition to proliferation [18]. They show prolonged inflammation with high protease levels, insufficient growth factors, and impaired angiogenesis [22]. Factors such as hypoxia, trauma, bacterial biofilms, and diseases can disrupt their healing [22]. These wounds may remain open for months with cycles of breakdown and repair. Chronic wounds often contain necrotic tissue and show persistent inflammation. Treatment requires addressing underlying impediments to initiate healing. DFUs are one significant chronic wound concern. Diabetes causes peripheral neuropathy and arterial disease, often with foot deformities that increase skin pressure [23]. Minor injuries can go unnoticed due to neuropathy and heal poorly from ischemia [23]. These ulcers typically occur on feet at pressure points. Approximately 15-34% of diabetics develop foot ulcers [23]. Men have 1.5 times higher incidence than women and the risk increases with age and diabetes duration [23].

DFU development requires multiple risk factors, including peripheral neuropathy (PN), peripheral arterial disease (PAD), and foot deformity or trauma [23]. Peripheral neuropathy causes loss of protective sensation, making patients unaware of injuries. Approximately 20% of DFU patients require lower extremity amputation, with DFUs preceding 85% of diabetes-related amputations [22], [19]. Despite conservative care, 15% of diabetic ulcers fail to heal, leading to amputation within 6-18 months, and over half recur within five years [23]. DFUs can be classified as neuropathic, ischemic, or neuroischemic [23]. Neuropathic ulcers occur with sufficient blood flow but nerve damage, forming on weight-bearing areas due to unnoticed trauma [23]. Autonomic neuropathy reduces sweating, causing dry skin and calluses [23]. Ischemic ulcers, caused by peripheral arterial disease, appear on toes or foot margins with poor perfusion signs [24], [25]. Neuroischemic ulcers combine both of these conditions [23]. Regardless of the type, infection can complicate any diabetic ulcer, but

ischemic ulcers have a high risk of progressing to gangrene [24]. Understanding these different chronic wound types is important from the perspective of AI classification.

DFUs are assessed by severity or stage, which guides treatment and predicts outcomes. Several wound classification systems exist, including the Wagner and University of Texas (UT) classifications. These systems provide standardized descriptions of diabetic ulcer advancement, and they are highly relevant for AI models that categorize wound severity from images or clinical data.

Wagner classification grades wounds from 0 to 5 based on depth and gangrene presence [26]. Grade 0 indicates intact skin with high-risk foot. Grade 1 is a superficial ulcer. Grade 2 denotes a deep ulcer penetrating to tendon, bone, or joint capsule. Grade 3 includes deep ulcer with abscess or osteomyelitis. Grade 4 indicates partial foot gangrene, while Grade 5 shows extensive foot gangrene [26]. Lower grades (1-2) are open ulcers, while grades 3+ indicate severe infection or necrosis. Higher grades correlate with amputation risk. The system focuses on depth and necrosis, remaining useful in research and clinical documentation [17], [26]. The University of Texas (UT) classification system combines grade and stage [26]. Grades 0-3 indicate wound depth, from pre-ulcerative to bone-penetrating ulcers. Stages A–D reflect infection and ischemia status: clean (A), infected (B), ischemic (C), or both (D) [26]. The system creates 16 categories and recognizes that infection and ischemia impact healing. Higher UT stages correlate with increased amputation risk and longer healing times [27].

The UT and Wagner classification have strong predictive value, making them effective for an AI wound classifier to predict healing outcomes [26]. The detail in these classification systems is important for automated wound assessments. While DUSS, SINBAD, and DEPA have performed well, Wagner and UT combine simplicity and accuracy effectively, making them more suitable for AI wound care models [26]. Despite these standardized frameworks, there is a lack of peer-reviewed research applying them to AI classification tasks.

Wound care is critical for managing acute and chronic wounds, aiming to promote healing and prevent complications such as infection or amputation. Effective wound care combines traditional measures, such as cleaning, debridement, and basic dressings, with advanced therapies to optimize healing. Removing dirt, necrotic tissue, and foreign materials is often the first step in managing chronic wounds such as DFUs [28]. Debridement (surgical, enzymatic, autolytic, mechanical, or biological) decreases bacterial burden and promotes healing [28]. In DFUs, regular sharp debridement accelerates healing and reduces amputation

risk [28]. Simple gauze or cotton wool dressings can protect wounds and absorb exudate, being also inexpensive but offer limited healing benefits [29]. Although, gauze might adhere to wound, causing trauma during removal [29]. Generally, dry dressings suit clean wounds with minimal exudate [30].

A major advancement in wound care was recognizing that wounds heal faster in moist environments. Traditional dressings allow wound desiccation, while modern dressings maintain moisture, promoting cell migration and autolytic debridement [29]. Dry gauze has been largely replaced by moist dressings in modern care [30]. Wet-to-dry gauze dressings have been used for debridement but can disrupt new tissue [28]. Clinicians favor methods that preserve granulation tissue while removing dead tissue. Standard care focuses on preventing infection through proper cleaning, antiseptics, and systemic antibiotics when needed [31]. Topical antimicrobials may treat superficial infections, while deep DFU infections require systemic antibiotics [28]. Overall, traditional wound care provides the foundation of clean, infection-free wound bed for moving on to advanced therapies.

Modern wound care uses advanced dressings to accelerate healing in chronic wounds by maintaining ideal wound conditions and stimulating tissue regeneration [30]. Unlike gauze, these occlusive dressings interact with wounds to promote healing while protecting against contamination [30]. Films and foams are transparent polyurethane sheets that allow oxygen exchange while blocking bacteria. Films suit shallow wounds, while foams absorb moderate exudate [30]. Both maintain moisture and support autolytic debridement [30]. Hydrocolloid dressings contain gel-forming agents that promote granulation in mild to moderate exuding wounds [28]. Hydrogels rehydrate tissues in dry wounds and provide cooling effects [28]. A 2024 meta-analysis by Zhao et al. showed that hydrogels improved DFU healing compared to conventional dressings [32]. Dressing selection depends on wound characteristics such as depth, exudate level, and infection; no single product suits all wounds [28]. Clinicians select dressings based on wound needs, cost, and patient comfort [33]. Generally, maintaining a moist wound bed with appropriate dressing improves healing [29], [33].

Advanced therapeutic modalities can enhance wound healing, particularly for chronic wounds characterized by growth factor deficiency [34], [35]. Exogenous growth factors promote cell proliferation and tissue repair [34], [35]. Among investigated growth factors, platelet-derived growth factor (PDGF) has shown notable success. Recombinant human PDGF-BB gel (becaplermin) became the first FDA-approved growth factor therapy for DFUs, with clinical

trials showing that 50% of treated DFUs achieved full closure versus 35% with placebo [28]. However, its high cost and daily application requirements limit its widespread use [28]. Platelet-rich plasma (PRP), a concentrate of patient's platelets and growth factors, can be applied as gel or injected into wounds [36], [37]. Systematic review by OuYang et al. found that PRP significantly increases DFU healing rates compared to standard care [38]. Recent research by Koivunotko et al. demonstrated that NFC combined with PRP enhanced wound healing through controlled PRP release, increasing angiogenesis and supporting re-epithelialization [39]. NFC is emerging in advanced wound care, with its dressing form showing clinical success in treating skin graft donor sites [7]. A clinical trial comparing NFC dressing with polylactide-based copolymer dressing in 24 skin grafting patients showed NFC advantages in scar appearance and skin elasticity [7].

Advanced wound therapies utilizing mechanical treatment or environmental factors can supplement standard care but are not standalone solutions. Negative pressure wound therapy (NPWT) applies controlled suction via a sealed foam dressing with vacuum pump [28]. NPWT removes excess exudate, reduces bacterial colonization, and increases blood flow and granulation tissue formation [28]. Studies have shown that NPWT leads to better healing outcomes compared to standard dressings, with shorter healing times and increased wound closure [28]. For post-surgical diabetic foot wounds, NPWT shows faster healing than moist gauze dressings [28]. Hyperbaric oxygen therapy (HBOT) is another environmentally applicable treatment, that delivers 100% oxygen in pressurized chambers to help hypoxic wounds [40]. While evidence is mixed, a review by Stoekenbroek et al. found that HBOT improved short-term diabetic ulcer healing [40]. Meta-analysis by Oley et al. shows HBOT enhances healing rates and reduces amputation risk in moderate-to-severe DFUs, particularly with peripheral arterial disease [41]. However, HBOT requires daily sessions for weeks and is costly. Given wound care's complexity and numerous variables, AI-based tools could assist in wound assessment, treatment selection, monitoring treatment response and predicting healing outcomes. The proposed AI-based tool could assist clinicians in these complex wound care decisions and shorten the trial-and-error period, and reducing the risks that come with prolonged healing times.

2.2 Artificial Intelligence

AI refers to the simulation of human intelligence processes by machines [42]. The field was formally born in 1956 when the term ‘artificial intelligence’ was coined at a Dartmouth

College workshop [42]. Early AI research focused on automatic reasoning and logic, such as theorem-proving programs, and it achieved some success, but also faced setbacks, as many problems proved harder than expected [42]. In recent decades, AI has moved from slideshows to real-life applications such as computer vision, speech recognition, and large language models (LLMs).

Machine learning (ML) is a subfield of AI that focuses on algorithms that learn from data to improve task performance [42]. Instead of being explicitly programmed with fixed rules, ML systems are adapted based on example data. ML has become the method of choice for many AI applications because it can automatically learn how to perform complex tasks by generalizing past experiences [42]. There are three main ML paradigms, distinguished by how they learn [42]. The first one is supervised learning. In supervised learning, algorithms learn from labeled examples [43]. Each training data point comes with a correct output (label), and the model adjusts to predict those labels for new inputs. This approach learns mapping from inputs, such as images, to desired outputs, such as categories. Classic supervised tasks include classification, which means predicting a discrete label, such as “cat” or “dog”, and regression, predicting a numeric value [43]. When algorithms discover patterns in unlabeled data, it is called unsupervised learning [43]. The goal is to find intrinsic structure or groupings without any predefined labels. For example, clustering methods can group similar images, and techniques, such as principal component analysis (PCA) can reduce data dimensionality [44]. The third approach between these two paradigms is reinforcement learning. In reinforcement learning algorithms learn by interacting with an environment and receiving rewards or penalties for actions, rather than explicit labels [43]. The system tries actions and learns a policy to maximize cumulative reward [43]. For example, a reinforcement learning agent can learn to play a game such as chess by trial and error, receiving a positive reward for winning moves and a negative reward for losing moves. Over time, it improves its strategy to achieve long-term goals.

Deep learning is a subfield of machine learning that employs neural networks with many layers (deep neural networks) to learn complex representations of data [42]. Although the concept of neural networks dates back to the 1950s, practical implementations for many years were limited to shallow architectures with only one or two hidden layers, largely due to computational constraints and the vanishing-gradient problem that complicated the training of deeper models [45]. It was not until the 2000s that multiple hidden layer deep networks became feasible. This shift was driven by the ability of large-scale labeled datasets such as

ImageNet, parallelizable training by graphics processing units (GPUs), and algorithmic innovations such as rectified linear units, batch normalization, and advanced regularization techniques [46], [47]. However, in limited data situations, such as this thesis, the theoretical advantages of increased depth may not be fully realized without strategies such as transfer learning and data augmentation, to mitigate overfitting and improve generalization [48], [49]. Instead of relying on handpicked features, deep learning enables a system to automatically learn features at multiple levels of abstraction from raw inputs [43]. For example, in image analysis, a deep network's early layers may learn low-level features, such as edges, while later layers capture high-level concepts, such as object parts [43]. The key idea is that these layers of representation are learned from data rather than designed by humans. Over the past decade, deep learning has led to significant breakthroughs in AI. It has reached cutting-edge outcomes on tasks that have long been challenging for previous AI techniques, often equaling or surpassing human-level performance [43]. One landmark moment was the 2012 ImageNet competition, where a deep CNN halved the error rate for image classification compared to the best earlier approaches [50]. Due to large datasets and increased computing power, especially GPUs, deep learning systems can be trained on millions of examples to achieve high accuracy [43].

CNNs are a class of deep learning models that are particularly well suited for image and visual data [43]. A CNN is inspired by the visual cortex in animals, it uses layers of artificial neurons with local receptive fields, analogously scanning small patches of an image, to detect features such as edges or textures [43]. Local connections and shared weights, with each filter or kernel applied across the entire image, take advantage of the spatial organization of images. This method substantially reduces the number of parameters when compared to fully connected neural networks [43]. This makes CNNs easier to train and more efficient because the same feature detector is used at multiple locations. In essence, a CNN learns a hierarchy of visual features; for example, one filter might activate diagonal edges, a higher layer might combine edges into corners or shapes, and deeper layers assemble these into object parts and ultimately complete object representations [43]. CNNs automatically learn which visual features that are important for classification without manual feature engineering [43].

As shown in **Figure 2**, an image classification CNN typically features several convolutional layers, each followed by non-linear activation functions, along with pooling layers that reduce the size of feature maps to achieve spatial invariance [43]. Each layer detects low-level features, such as edges, colors, and textures, where later layers capture high-level patterns,

such as object parts and shapes [43]. After several convolution/pooling stages, the network usually includes one or more fully connected layers that integrate all extracted features to make a final prediction, for example assigning a probability to each image class [43]. Similar to many other deep learning architectures, CNNs are also trained end-to-end with large labeled datasets, and the training process uses backpropagation and gradient-based optimization to adjust the network weights to minimize classification errors in the training examples [43]. These type of models often have millions of parameters (weights), and stochastic gradient descent or related optimization algorithms are used to gradually improve the performance of the network, iterating over many examples [43].

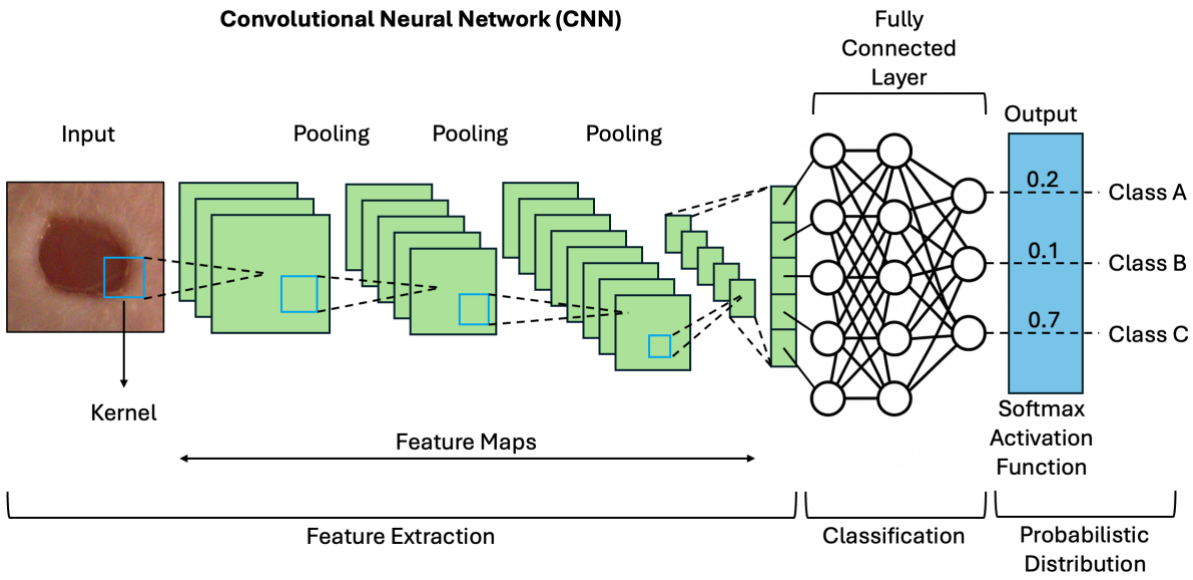

Figure 2. Architecture of a general CNN, featuring three convolutional layers (green), followed by pooling layers in between them, followed by three fully connected layers that integrate all extracted features to make a final prediction, assigning a probability to each class (A, B, C) through softmax activation function (blue).

Over the years, CNN architectures have evolved significantly. One milestone was AlexNet (2012), a deep CNN that was the first to win the ImageNet challenge; it introduced the use of the ReLU activation function for faster training and dropout regularization to reduce overfitting, allowing a much deeper network than previously feasible [43], [51], [52]. Subsequent models, such as Visual Geometry Group (VGG) (2014), showed that pushing to an even greater depth (16–19 layers) can improve accuracy, although at the cost of more computation [43]. A breakthrough was achieved with Residual neural network (ResNet) in 2015, which introduced residual learning and skip connections to enable ultra-deep networks (over 100 layers) without vanishing gradients [43]. ResNet, at 152 layers, won the 2015 ImageNet Large Scale Visual Recognition Challenge (ILSVRC) championship; despite being

eight times deeper than VGG, it had lower computational complexity [43]. These and other architectural innovations have continually improved CNN performance. In addition to architecture, optimization strategies, such as batch normalization, data augmentation, and advanced optimizers, have further boosted CNN capabilities [43]. As discussed, today's CNNs are highly optimized and scalable, benefiting from years of research into layer designs and training techniques.

CNNs continue to be the cornerstone of image-based AI; however, new architectures have been developed for the scene. Vision Transformers (ViTs) have recently emerged as an alternative approach to image recognition. Introduced in 2020, ViTs apply the transformer architecture, originally from natural language processing (NLP), to image patches using self-attention mechanisms instead of convolutions to learn visual representations [53]. Early research indicated that ViTs can achieve a performance comparable to that of state-of-the-art CNNs in image classification tasks, and in some cases, they even outperform CNNs in terms of accuracy or robustness [54]. This has encouraged interest in hybrid models and comparisons between convolutional- and attention based vision models. However, CNNs remain highly effective and are deeply rooted in existing applications. As discussed earlier, they benefit from decades of development and optimization for visual data. In practice, current state-of-the-art vision systems can combine ideas from both paradigms, for example, using a CNN backbone with attention modules. Looking forward, CNNs are expected to continue evolving with more efficient architectures and unsupervised learning techniques, even as ViTs and other innovations gain traction.

Researchers have been expanding AI models beyond single-modality inputs such as images or text alone. This approach fuses information from different datatypes in unified models, for example, by combining vision and language or merging images with numeric features. Early studies demonstrated that leveraging complementary modalities can improve understanding and performance; for example, integrating visual and textual data enabled tasks such as visual question answering and image captioning that were infeasible with one modality alone [55]. By exploiting the complementary characteristics of each modality, multimodal models achieve more accurate predictions than unimodal models [55]. Complex decisions often rely on heterogeneous data, such as medical images, patient demographics, laboratory results, genomics, and previous discussions with patients. Rather than analyzing X-ray or laboratory results in isolation, multimodal models integrate imaging with patient metadata and clinical biomarkers to mirror a physician's holistic approach. As discussed earlier, the performance of

AI models seems to improve as the types of information increase. Studies across oncology, radiology, neurology, and other fields have reported that such data fusion improves diagnostic and prognostic accuracy [56]. A 2022 scoping review of 128 studies by Kline et al. on health AI observed a clear trend: combining modalities provided better predictive performance than single-source models in most cases [56]. Notably, among studies that directly compared multimodal versus unimodal models, the fused approaches achieved on average a 6.4% higher accuracy in prediction tasks [56]. This aligns with other studies showing that multimodal systems are usually more robust and precise than unimodal systems. Indeed, a study by Benani et al. (2025) analyzed 97 studies across 12 medical specialities, with oncology being the most represented [57]. The study concluded that multimodal algorithms outperformed unimodal algorithms in over 90% of comparative studies [57]. Although not yet peer-reviewed, this study strengthens the hypothesis that multimodality is a more accurate approach in most cases. In radiology, one team fused chest CT scan images with corresponding electronic medical record (EMR) data to detect pulmonary embolisms, achieving an area under the receiver operating characteristic curve (AUROC) of ~0.95 with the best fused model significantly outperformed both of best performing single modality models: 0.036 AUROC higher than EMR only-model and 0.156 AUROC higher than imaging-only model [58]. In oncology and pathology, integrating histological images with patient clinical features has improved classification accuracy. Yan et al. reported that combining breast tumor histopathology images with patient data increased the diagnostic accuracy to 87.9% versus 83.6% using images alone, and only 78.5% with clinical data alone [59]. Likewise, for cancer prognostics, merging pathology slides with genomic profiles improved risk predictions, for example combined models achieved higher concordance index in survival prediction than either modality by itself [59]. Even in multi-imaging scenarios, using different imaging modalities together with tabular biomarkers can provide additional benefits; for example, combining a patient's magnetic resonance imaging (MRI) and positron emission tomography (PET) scans with cognitive test scores improved Alzheimer's disease classification accuracy compared to any single input [59]. Across various medical applications, these studies demonstrate that multimodal AI systems often attain increased accuracy and diagnostic power compared to a single modality, leveraging the strengths of each data source to support clinical decisions more reliably. Thus, multimodality is a promising direction for research on the medical context of artificial intelligence.

2.3 Application to Wound Care

In the last decade, before moving to multimodal approaches, image-only models have been introduced for wound care. Wang et al. (2017) used a two-stage support vector machine (SVM) with conditional random fields to automatically detect wound boundaries on smartphone images [60], [61]. This approach achieved good specificity (~94.6%) but moderate sensitivity (~73.3%), highlighting the need for more accessible, robust models and larger datasets [60], [61]. Around 2018–2019, deep CNNs emerged in wound care tasks. Ohura et al. (2019) trained CNN-based segmentation models (SegNet, U-Net variants) on chronic wound images, achieving high accuracy (~97–99%) in delineating wound areas [60], [62]. Goyal et al. (2020) tackled DFU monitoring by classifying signs of ischemia and infection from images using deep networks and ensemble models; they reported ~83% accuracy for ischemia detection but lower (~66%) for infection [60], [63]. These early deep learning studies were mostly unimodal (image-only) and underscored challenges such as limited datasets and variability in imaging conditions [60]. However, they laid the groundwork for incorporating additional datatypes to improve performance.

As discussed in chapter 2.1, wound healing is influenced by several factors. Recent research has shifted towards multimodal models that fuse these factors, such as image data with wound size, location, or patient health [10], [11], [64], [65]. These studies introduced highly tailored AI systems that combine multiple inputs. For example, Anisuzzaman et al. (2022) developed a deep multimodal classifier that takes a wound image and its body location as inputs to automatically categorize the wound type [11]. By concatenating image features with location information, this system achieved a mixed-class accuracy from ~82% to 100% (when including background/normal skin) and a wound-type classification accuracy of ~73–97% [11]. As discussed with broader scope of medical AI in chapter 2.2, multimodal AI has also improved performance in wound care tasks when compared to image-only approaches [56]. A clear example is in burn wound assessment: Rambhatla et al. developed a system (DL4Burn) that integrates burn wound photographs with patient demographics/injury data to predict the need for surgical intervention [64]. The inclusion of patient context dramatically boosted accuracy, and their multimodal model was approximately 93.8% accurate, where a state-of-the-art CNN using images alone achieved approximately 81.0% on the same task [64]. In other words, the model's ability to correctly decide if a burn required a surgery improved substantially by adding patient data, such as age or burn mechanism, to the image analysis. Similarly, the wound-type classifier using the images and locations mentioned above

significantly outperformed previous models that lacked location input [11]. These findings underscore that in wound care, similar to broader medical AI, combining visual data with complementary modalities provides more accurate and reliable outcomes. Multimodal AI systems in this domain have started to match expert-level assessments; for example, a recent multimodal tool for surgical site infection monitoring performed evenly with clinician triage in identifying infections [65]. At the same time, it potentially reduces the clinician's workload.

Beyond simple numeric metadata, recent studies have incorporated text-based clinical data along with images. The Deep Multimodal Wound Assessment Tool (DM-WAT), developed by Busanur et al. (2025), is a novel framework that analyzes smartphone-captured wound images together with electronic health record (EHR) notes to determine whether a chronic wound patient needs specialist referral [10]. DM-WAT uses a Vision Transformer for image feature extraction and a BERT-based model for text embeddings to combine visual and textual cues via intermediate fusion [10]. On a small dataset of chronic wounds, DM-WAT achieved ~77% accuracy (F1 score ~70%) for referral recommendations, significantly outperforming models that used only images or text [10]. This indicates that combining wound images with clinical descriptions provides more accurate decision support. Another cutting-edge approach to wound infection detection augments images with text in an explainable way. Busaranuvong et al. (2025) introduced SCARWID, which generates a synthetic caption for a given wound image, using a GPT-4-based vision-language model, and then fuses the image and caption through cross-attention for infection classification [66]. In DFU images, this image + caption model reached ~81% accuracy (sensitivity 0.85, specificity 0.78) in detecting infections [66]. Importantly, captions improve transparency by describing visual findings, which nurses can review alongside the AI's output [66]. These studies demonstrate how combining several architectures, such as CNNs and transformers, can leverage unstructured clinical data and improve wound assessment beyond what either modality alone achieves.

Thermal and hyperspectral imaging can reveal physiological aspects (such as perfusion and temperature changes) that are not visible in standard images. For example, thermal imaging has been used to detect early wound complications. Fletcher et al. (2021) collected smartphone thermal camera images from post-surgical wounds (caesarean incisions) and trained CNN models to predict surgical site infections [67]. Their best model (transfer-learned CNN) achieved an AUC of 0.90 (with 95% sensitivity and 84% specificity) in distinguishing

between infected and normal wounds [67]. This was one of the first demonstrations that thermal images alone can power an AI model for wound infection detection in a real clinical context. Similarly, in diabetic foot care, researchers have applied deep CNNs to infrared foot thermography. Verma et al. (2024) used a ResNet/EfficientNet model on thermal foot scans to classify the presence of ulcers, reporting accuracies around 96–99% for detecting DFUs versus healthy tissue [68]. Some projects go further by fusing visual-light images with thermal data; research prototypes have used mobile devices and portable thermographic cameras to capture both modalities and overlay the information for analysis [69]. The rationale is that aligned multimodal images can feed a network that learns both color/textural cues and heat distribution, potentially improving diagnostic accuracy for infection or perfusion status. Likewise, hyperspectral imaging (capturing oxygenation or hemoglobin levels in tissue) paired with AI has shown 80-90% sensitivity and 74–86% specificity in predicting which diabetic foot ulcers will heal versus not heal [60]. These multispectral approaches, which often involve deep learning algorithms to interpret complex image data, illustrate the expansion of wound AI from simple RGB images to a multimodal sensor perspective. While many thermal/hyperspectral studies are still experimental, they highlight AI's potential to integrate diverse data, such as temperature maps or spectral signatures, for comprehensive wound assessment.

The translation of AI models into clinical practice requires validation in real-world settings. Several multimodal AI systems have been subjected to clinical testing. Tan et al. (2021) performed a prospective study to validate an AI-powered wound imaging application (CARES4WOUNDS) for diabetic foot ulcers [70]. This smartphone-based system uses computer vision to automatically measure wound dimensions (length, width, and area) and classify tissue types. In trials on 28 DFU patients, the app's measurements showed excellent intra- and inter-rater reliability (intraclass correlation >0.93) for repeated measures, and inter-device agreement of 0.923–0.965 for key metrics (length, width, and area) [70]. In other words, the AI app was as consistent as expert nurses using traditional rulers, demonstrating clinical grade accuracy in wound sizing [70]. Such validation is crucial for regulatory approval and adoption. Another example is the work by Swift Medical, Ramachandram et al. (2022), who trained deep CNN models on an unprecedented 450,000+ wound images to segment wound boundaries and tissue composition [71]. Their system, deployed on mobile devices, can objectively quantify percentages of granulation versus slough versus eschar tissue, addressing the high variability observed among clinicians in visual estimations [71].

By demonstrating that AI can match or exceed human consistency in wound assessment, these studies build trust in AI-enabled wound care tools. Many commercial wound platforms now integrate AI for imaging and EHR data, although not all have been published in peer-reviewed literature [60]. Overall, the trend in the last ten years has moved from proof-of-concept to clinical validation of AI models, especially in diabetic foot ulcer management, where the early detection of non-healing wounds or infections can prevent amputation. As discussed in this chapter, many of the platforms were also developed for mobile devices, which is crucial for making advanced wound care even more accessible also to areas with limited availability of specialists.

2.4 About the Data

In any AI-driven medical application, the quality of data fundamentally limits the quality of the results. Thus, AI models are limited to the quality of data they are trained on [72]. Studies have warned that using poor-quality or biased data to train medical AI can lead to erroneous or unreliable conclusions in clinical decision-making [73]. In the context of patient care, such errors could compromise diagnostic and recovery follow-up accuracy and safety. Indeed, many reported issues with AI models (lack of clinical usefulness, unreliability, unfair bias, etc.) have been attributed to suboptimal dataset quality [74]. Therefore, ensuring high-quality representative data is crucial for developing wound care models. Recent literature has emphasized that data quality ”dictates the behavior” of machine learning systems in medicine [75].

2.4.1 RGB Images

In this study, RGB images of wounds were used as the main input. High resolution RGB images of wounds provide rich visual information about the wound’s appearance. Such images capture color, shape, texture, and tissue composition, which are key indicators of wound status. RGB images have been widely used in research to classify wounds, monitor healing, and assist in diagnosis. Secco et al. grouped wound-image pixels into color categories to identify different tissue components [76]. RGB sensors, such as CCD or CMOS sensors in smartphones or system cameras, are commonly used in wound monitoring because they are inexpensive and easy to deploy. For example, smartphone cameras were used in the study by Kaselimi et al. [77]. RGB sensors have been used in numerous AI models for tasks, such as wound segmentation and ulcer detection [77]. As discussed in the previous section, Anisuzzaman et al. developed a deep neural network classifier that uses wound images to

automatically categorize wound types (diabetic, pressure, surgical, and venous ulcers) [11]. These studies demonstrate that RGB data can effectively capture relevant features diagnostically and monitor wound healing. Although many previously covered early wound image models were trained on only tens to hundreds of images, they still demonstrated the feasibility of automated wound detection and segmentation. One quite extreme example is a study by Abubakar et al. that used only 60 wound images (29 pressure ulcers and 31 burns) and expanded this to hundreds of training images through augmentation, enabling an SVM-based model to distinguish burn versus pressure wounds [78]. This and previously introduced studies underscore that even modest collections of RGB images, when carefully annotated and preprocessed, contain enough signal for AI models to learn clinically relevant patterns.

2.4.2 Thermal Images

Thermal images were used as an additional input for this thesis model. Thermography produces images of the temperature distribution across the wound and surrounding skin [79]. The scientific rationale is that skin temperature correlates with physiological processes such as inflammation, infection, and blood flow [79]. In wound care, abnormal heat patterns can be early indicators of complications. For example, an infected or highly inflamed wound area often exhibits increased heat due to elevated metabolic activity and blood perfusion, where a healing wound might gradually cool down as inflammation resolves [79], [80]. For diabetic foot ulcers, comparing thermal images of both feet can reveal temperature asymmetries; a significantly warmer region on one foot may indicate ulceration or infection [77]. A recent scoping review by Fridberg et al. (2024) confirmed that thermography has been used to detect and predict wound infection and inflammation by capturing temperature changes [79]. In the same review, with surgical wounds researchers observed a characteristic pattern: a moderate temperature elevation in the first 1–2 weeks post-surgery (reflecting normal inflammatory healing), followed by a return to baseline over the next few months [79]. If a secondary temperature spike occurs during healing, it likely signals an infection developing in the wound [79]. This shows the relevance of thermal imaging, which provides a quantitative visual map of wound pathophysiology that is not visible in standard images. Thermal wound imaging is a relatively new modality in AI models, but its value has been demonstrated in several studies. For example, Li et al. (2024) processed 1,189 thermal images of surgical incisions with a computer vision model to assess healing status [81]. Their model automatically identified the incision region and classified healing versus non-healing with

high accuracy, suggesting that thermal cues can effectively indicate wound healing progress [81].

2.4.3 Wound Area

The third input for this thesis model was wound area in square millimeters. Wound size is a straightforward measure that is directly associated with healing; as a wound heals, its area typically decreases [82]. Clinically, a common heuristic is that a significant reduction in area over a few weeks is a positive sign, whereas stagnant or increasing area is alarming [82]. A recent machine learning study by Berezo et al. analyzed over 1.2 million chronic wound cases from electronic records and found that wound area was among the top predictive features for whether a wound would fail to heal in 12 weeks [13]. In that study, factors such as wound depth, location, and area strongly influenced the model's predictions of non-healing wounds. This aligns with clinical intuition that larger and deeper wounds generally take longer to heal or may not heal without advanced interventions. Because of this strong predictive value, documenting wound size is routine in care and is often included in prognostic models [82]. While this thesis model treats the wound area as an additional data channel, evidence from multimodal studies supports this approach. For example, Rambhatla et al. developed a model for burn wound assessment that combined burn images with patient metadata, including total burn size, to predict the need for surgery [64]. Their ResNet-50 based multimodal network, which used both image pixels and numeric injury parameters as input, outperformed image-only models in matching expert decisions [64]. This suggests that giving the model explicit size measurements helps it contextualize the visual information, and helps dealing with low quality or non-standardized data, where the image distance from the wound might vary. Thus, incorporating wound area along with images in a multimodal model is a natural choice, reflecting how human clinicians consider both what a wound looks like and how big it is.

2.4.4 Amount of Data

Most of the models from the studies covered in the previous chapters were trained on only tens to few hundreds of images. This highlights a major challenge in applying AI to wound care: the limited availability of large, diverse datasets. The lack of labeled data was also noted in a survey by Cheplygina et al. [83]. They pointed out that a large amount of unlabeled data is widely available, but without detailed labeling, advanced models are challenging to build [83]. Manual labeling of images is an expensive and/or time-consuming process, and similar labeling is not needed in clinical practice as in ML studies; therefore, labeled datasets do not

appear “naturally” from clinical work [83]. Unlike domains such as general object recognition, which have millions of images available (popular datasets such as CIFAR-10/100 and ImageNet), medical data tend to be siloed and small in scale or constrained by various regulatory anonymity requirements [84], [85]. For example, in the context of diabetic foot ulcers, there is one open source dataset, but even that has only 493 original images and a 167 image test set [86]. Without significant preprocessing with a wound care expert, the dataset is suitable mostly for binary classifications between abnormal (ulcer) and normal (healthy) skin. Researchers have developed strategies for overcoming these limitations. One common approach is data augmentation, which artificially expands the dataset by creating modified copies of images (e.g. rotating, flipping, scaling, and adding noise). Candermir et al. found that reasonable augmentation can improve model performance by 5–15% in medical imaging tasks [87]. It effectively acts as a regularizer, reducing overfitting by preventing the model from memorizing exact images. Augmentation effectively generates new training examples that retain the characteristics of the original data while introducing variety. However, it is important to ensure that augmented images are still realistic; excessive or inappropriate augmentation can introduce bias or unrealistic patterns, potentially confusing the model [87]. For example, simple augmentations, such as horizontal flips or slight color shifts, can expose a CNN to different plausible views of wounds without needing new patients [88]. Hussain et al. reported that using such augmentation techniques substantially boosted the validation accuracy from ~66% to ~84–88% [88]. In wound imaging studies, augmentation is almost a standard practice; several studies covered in the earlier chapters utilized it in their models [10], [11], [65], [66], [68]. Another effective technique is transfer learning. Instead of training a deep network from scratch on limited wound data, in almost all the studies covered previously , researchers started with a model pre-trained on a large dataset and then fine-tuned it on wound images [10], [11], [64], [65], [66], [67], [68]. These models leverage previously learned features (colors, edges, textures, and shapes) that are general and useful in wound assessments. Despite challenges, these previously covered studies suggest that with these techniques, feasible performance can be attained even with relatively small datasets. Numerous studies on wound AI have reported surprisingly good accuracy, with limited data. For example, Rostami et al., referenced in Anisuzzaman’s related work, built a wound classifier on only 538 images and achieved approximately 92% classification accuracy in a three-class task [11], [89]. In the same review, Anisuzzaman et al. reported that Aguirre et al. achieved 85% accuracy with ~300 images by leveraging heavy augmentation and a pre-trained VGG-19 network [11]. These results illustrate that small datasets, if properly handled,

can produce models that perform well on their test sets. However, it is important to restrain optimism with caution; high accuracy on a narrow test set does not guarantee generalization to broader real-world settings. Authors often acknowledge this; for example, in the DL4Burn study, the researchers noted that their model's performance was based on a relatively small dataset and showed signs of slight overfitting, and they planned to collect more data via their mobile application to improve robustness [64]. In general, small-data models must be rigorously validated to ensure that they are truly generalizable and not over-tailored to the training distribution. More technical approaches, such as cross-validation and regularization, are also solutions to reduce overfitting; however, these are discussed later in this thesis.

3 Hypothesis and Aims

A multimodal AI model integrating RGB images, thermal images, and wound area measurements would succeed in classifying the post operative day (POD, also represented as “day”), of an input wound. Contribution of different modalities would vary, RGB images providing the most predictive value for the model, and the thermal images adding confusion due to their lower quality and homogeneous nature in the later PODs. Still the multimodal model would achieve higher accuracies in predicting treatment outcomes compared to the unimodal approach.

To test this hypothesis, the following aims were defined:

- Development of a POC multimodal AI model combining three parallel branches for RGB, thermal, and size inputs.
- Evaluation of each modality’s individual contribution and their combined effects.
- Model interpretability evaluation with Grad-CAM to visualize the model’s decision-making.

This study should also lay a foundation for future possibilities with multimodal AI in wound care, revealing common pitfalls, guiding for best practices and opening areas for improvement, such as quality of data, dealing with data leakage and designing an architecture that enables versatile predictive value that is based on real medical features.

4 Materials and Methods

4.1 Data

Data for this study was received from a study by Elle Koivunotko, PhD. In the study two different NFC hydrogel formulations (0.8% m/w, medical-grade) with or without glucose excipients and 0.9% NaCl as a control was evaluated using an in vivo diabetic wound model in 10 SKH1 nude mice (Charles River). Diabetes was induced by intraperitoneal Streptozotocin injections. GLP-compliant animal studies were conducted in collaboration with Mari Madetoja, MSc., from Madeconsulting Oy (Finland). Each mouse received two full-thickness wounds on its dorsal side: one treated with NFC hydrogel and the other with 0.9% NaCl as a control. Five mice were meant to be used per treatment formulation, but mouse number 6 from the "NFC without glucose" group had too low blood glucose levels, so it could not be wounded and was not included in this dataset. The primary endpoint was planned to be wound healing by Day 21, but the imaging was stopped on day 14 when no significant improvement on wound closure was observed. Wounds were photographed using a high-resolution camera (Canon EOS 600D), and temperature changes were monitored using FDA-cleared thermography (Thermidas IRT-384, Finland) equipped with an animal study software module. Data collected included wound size, mouse weight, blood sugar levels, and urine biomarkers. In addition to images, only wound size was used in this study.

System camera images (RGB images) were received in Canon RAW (.CR2) format. **Figure 3** shows an example image of an original image from day 0, Mouse 2. This image is already converted from RAW to JPEG. The images were captured from different angles. **Figure 3** shows an example from the same day and mouse from a different angle, and a cropped version of that angle. Thermal images were accompanied by a temperature scale of different field-of-view (FOV) regions; black-purple-blue colors encoded cooler temperatures (~10.0–21.7ºC), and yellow-red-white represented warmer temperatures (~28.7–37 ºC). The images were received in PDF format, so they had to be converted to JPEG. **Figure 3** shows an example of a thermal image and the cropped version of the same wound. The quality of the thermal images was significantly lower because they were mainly captured from further distance. Also, the thermal differences appeared mainly through days 0–6, where days 8 and 10 were identical in most of the cases.

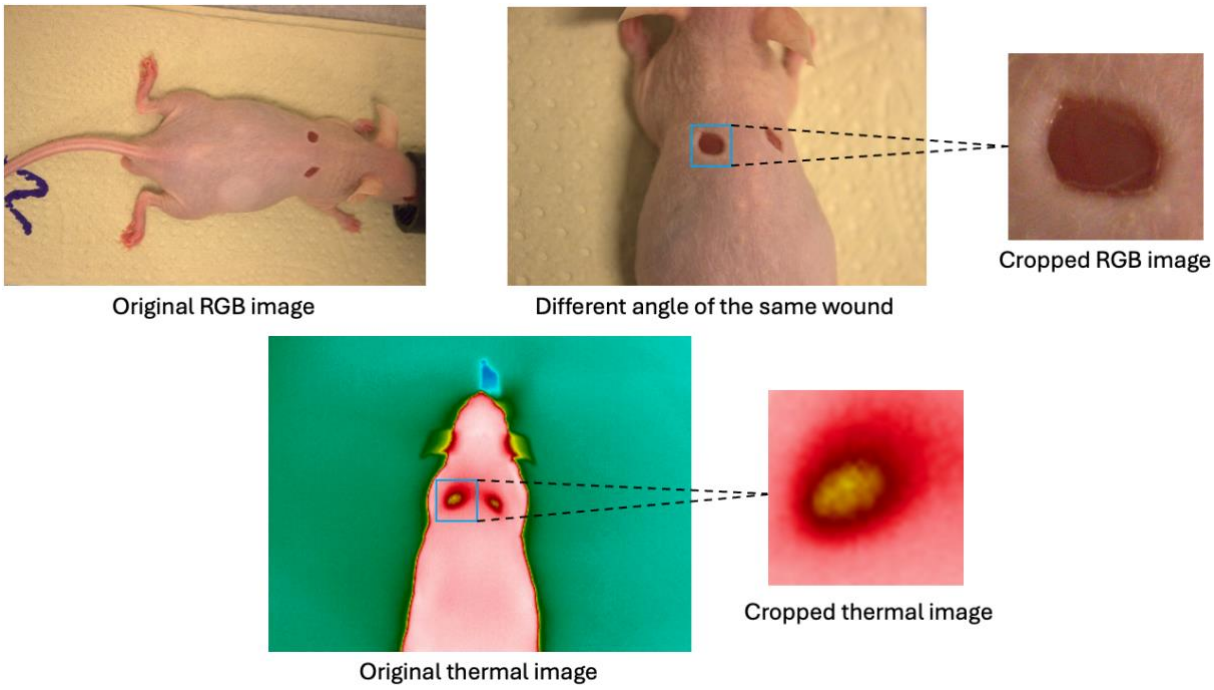

Figure 3 Top-left corner, an original RGB image from day 0, Mouse 2. Top-middle, a different angle of the same RGB image and the target area of the cropped RGB image (top-right). Bottom left, an original thermal image with the marked target area of the cropped thermal image (bottom-right).

Wound length and width were measured on each day, and the area was calculated from these measurements. **Figure 4** shows the individual wound closure trajectories for the control and treated wounds. Mice 1, 3, 5, and 7 were treated with NFC without glucose, and 2, 4, 8, 9, and 10 were treated with NFC with glucose. Mouse 6 was not wounded because it had too low blood glucose levels on the day of wounding. As seen in **Figure 4**, control wound closure seems to vary significantly more compared to treated wounds, with treated wound trajectories being noticeably more linear than control wounds, although not closing as rapidly.

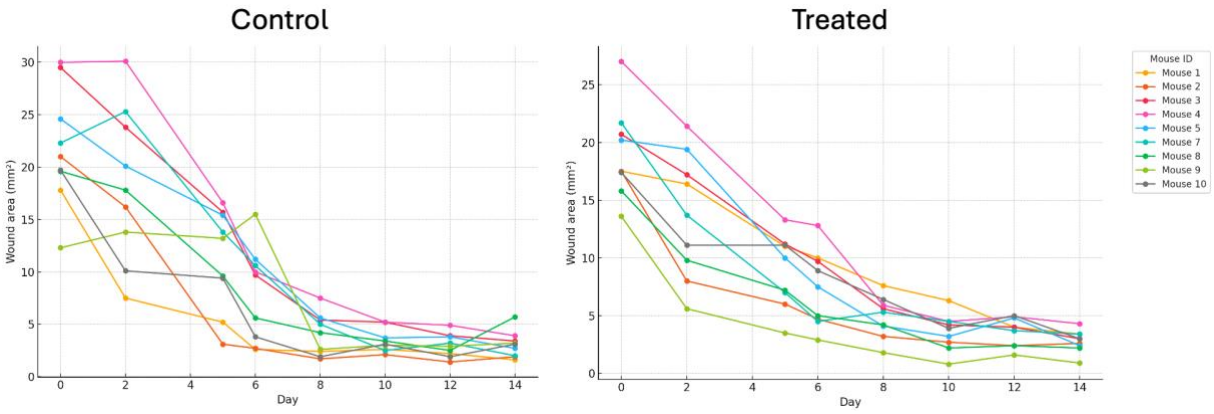

Figure 4 Individual trajectories of control (left) and treated (right) wound closures. Mouse IDs mapped with unique colors on the right side of the figure.

Each angle from all image types was utilized, so the number of samples per class varied. These class distribution variations were equalized using data augmentation. In **Figure 5**, the number in each cell represents the number of original images of the wound. In total, there were 328 RGB control images and 315 RGB treated images. Thermal images were mostly taken from the top of the mouse, where both wounds were visible in the image. Thus, there were significantly fewer original thermal images than RGB images. In total, there were 207 control and 183 treated thermal images. **Figure 5** shows also the distributions of the original thermal images from both classes, control and treated. These class distribution variations were also equalized using data augmentation.

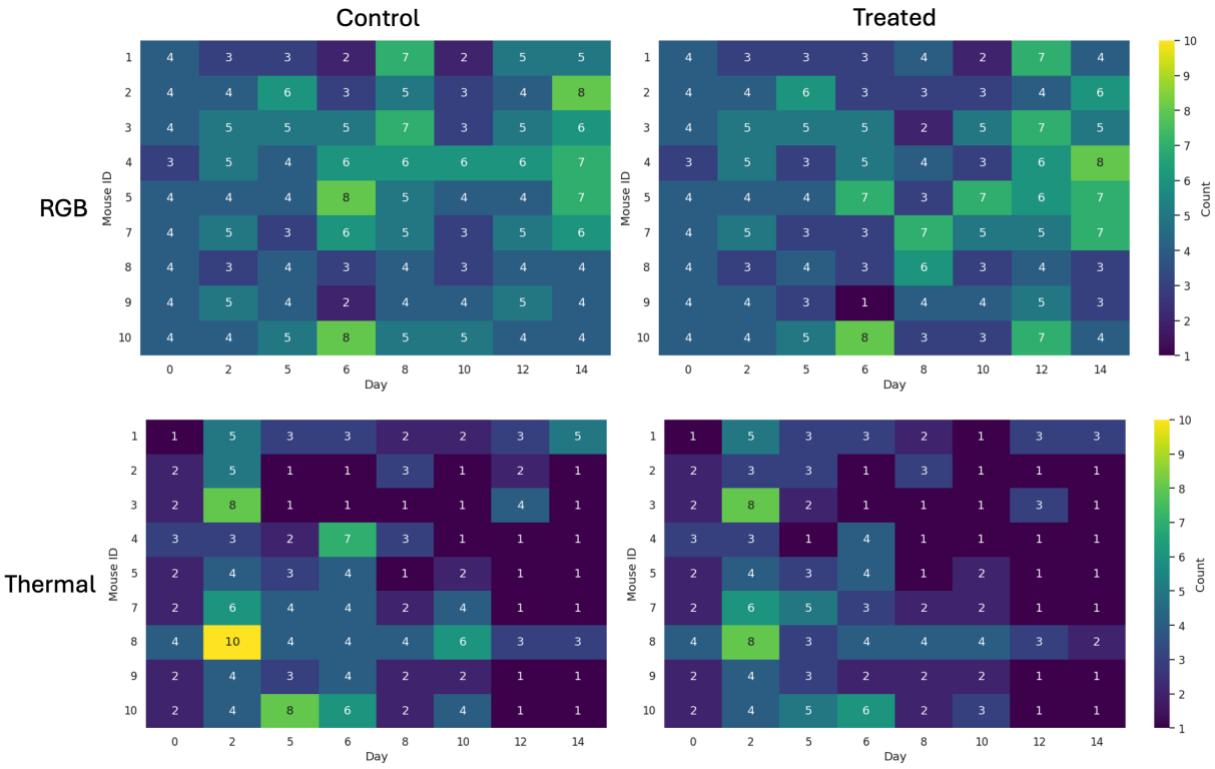

Figure 5 Class distributions for RGB (top row) and thermal images (bottom row), from both, control (left column) and treated (right column) groups. Cell sample count is color-coded with Viridis heatmap, color scale represented on the right of the table.

4.2 Preprocessing

The initial phase of preprocessing involved loading RGB images from cloud storage. These images were in the Canon RAW format (.CR2). Due to the compatibility with ML frameworks and to save space, it was necessary to convert the RGB images from RAW to JPEG format using the rawpy and Pillow Python libraries [90], [91], [92]. Appropriate white balance settings had to be embedded within the CR2 files to ensure precise color reproduction in converted images. The initial availability of thermal images was restricted to the PDF

format, necessitating their conversion to the JPEG format for further use. This conversion was accomplished utilizing the Pdf2image module from the Pillow library to facilitate the transformation of thermal images from PDF to JPEG format [91]. Python libraries OpenCV-Python, os, and shutil were utilized to perform semi-automated cropping, labeling and sorting the image files [92], [93], [94], [95].

To make the images suitable for potential use of transfer learning, they all needed to be resized. Dimensions of 224 × 224 were selected because this size is compatible with a wide range of transfer learning models aimed at classification tasks, including ResNet, VGG, and EfficientNet [96], [97], [98]. The OpenCV-Python and os were used to adjust the images to 224 × 224 pixels [93], [94]. For images that were not square, the cv2.BORDER_REPLICATE method was employed to fill in the missing pixels with the nearest pixel value, ensuring that the empty spaces were filled with extended pixels from the edges [99]. However, due to the low quality of some RGB images, this approach resulted in artificial lines appearing at the edges, which confused the model, as observed through Gradient-weighted Class Activation Mapping (Grad-CAM). Consequently, these images required additional preprocessing, which involved blurring the edges to reduce the repetition of line patterns.

The blurring process, which is visualized in **Figure 6**, was integrated into the augmentation process, starting with the original images, to ensure that multiple iterations of preprocessing did not degrade image quality. Initially, each original image was padded to a square shape and subsequently resized to 224 × 224 pixels, employing the same border replication technique as previously utilized. After resizing, a data augmentation pipeline was established to address the limitations of the dataset. This pipeline augmented the images to ten images per folder. The selected augmentation techniques included horizontal flip, vertical flip, combined flip, rotations of ±11°, zoom in/out by 10%, and translations of ±20%, in addition to the unaltered original image, resulting in ten distinct images per control/treated folder. A two-stage blurring strategy was implemented to mitigate the edge artifacts introduced by resizing, rotation, and translation. Initially, a selective blur was applied using morphological dilation of the padded-region mask and a distance-based alpha map to apply strong Gaussian smoothing over the replicated borders and a graded transition zone of ten pixels into the original image. This approach ensured that color mismatches at the image margins were effectively softened without diminishing central detail. Finally, a supplementary circular blur was applied to the borders of each image to even out blurred sections and maintain the visual coherence of the dataset. The quality of all preprocessing was visually verified by browsing through all images.

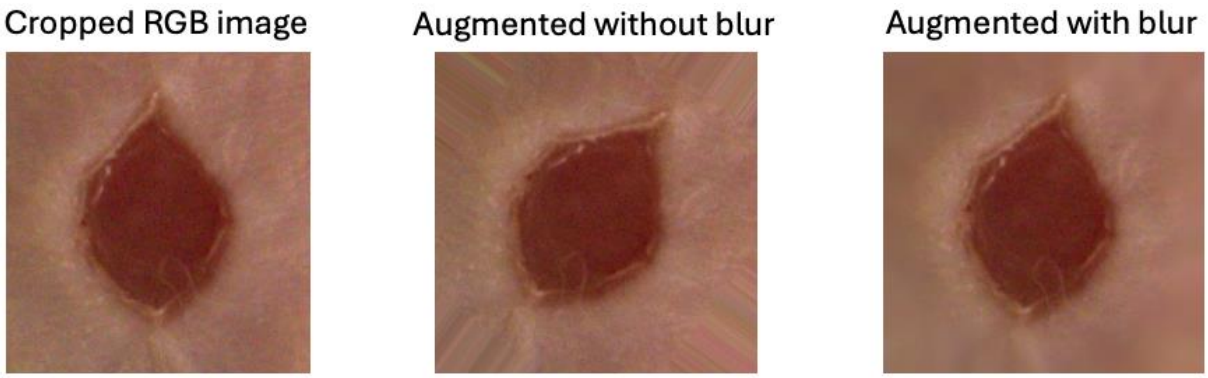

Figure 6 Two augmentation versions displayed (middle and right) from the original image (left). Middle wound image augmented with rotation (±0–11°) and right image augmented by rotation with circular blur to fade the artificial lines caused by augmentation.

4.3 Evaluation Metrics and Methods

4.3.1 Cross-Validation Strategy

Model evaluation was performed using multiple cross-validation techniques, with the aim of preventing data leakage. K-fold cross-validation (CV) was explored to maximize the data usage for training and validation. In K-fold CV, the dataset is randomly split into K folds, and the model is iteratively trained on K–1 folds and validated on the remaining fold, rotating until each fold serves as a validation set once [100]. However, in this dataset of serial wound images, the K-fold splits led to data leakage. Specifically, multiple images of the same wound, from different time points, were split across training and test folds, which violated the independence of training and validation data. Data leakage refers to the unintended inclusion of test information in training, which can significantly inflate performance estimates [100]. Rosenblatt et al. demonstrated that when the same subject data appear in both training and testing, models can achieve unrealistically high accuracy [101]. In medical imaging contexts, an improper slice-level split instead of a subject (patient)-level split has been shown to boost measured accuracy by over 30–50% due to leakage [100]. In the end, the nature of the data forced the choice to leave-group-out cross-validation. This highlights the need for careful validation design.

To address this, a Leave-One-Group-Out (LOGO) CV strategy was implemented. LOGO CV is a variation of the more known Leave-One-Out CV (LOOCV) [102]. Scikit-learn has a method for LOGO CV, which was utilized during the implementation of LOGO CV in this model [103]. Utilizing LOGO CV through that method, all images belonging to one group are left out as the test set in each fold, training the model on the remaining groups. In this study,

the natural grouping was by subject and wound: the dataset consisted of nine mice (subjects) with two wounds each (18 wound sites total), with partially augmented 10 image sets per wound. Thus, each fold left out all images of one wound (or one mouse) for validation, training on images from the other wounds. LOGO CV ensures that no wound images appear in both the training and testing sets in the same fold, eliminating data leakage. As shown by Yagis et al., this subject (mouse)-level split CV can eliminate unrealistically high accuracies and provide an unbiased performance estimate [100]. Although a standard K-fold split was initially attempted, it was replaced by LOGO CV once leakage was identified. By using LOGO CV, the evaluation better simulated a scenario of predicting an entirely new wound, which was the true test of generalization in this study. This strategy aligns with the best practices in medical AI development, which advise that all images from a given patient (or lesion) be kept within a single fold to prevent train-test contamination [101].

4.3.2 Performance Metrics

The model performance was quantified using classification metrics derived from the confusion matrix: accuracy, precision, recall, and F1-score. A traditional confusion matrix summarizes the outcomes of a binary classifier by organizing the predictions into true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN) [104]. These four outcomes completely describe the classifier's performance for binary classification. In this study, the model was a multi-class classification model, so the traditional 2 × 2 confusion matrix was not suitable. A confusion matrix for a multi-class classification problem is an N × N contingency table, where each row *i* corresponds to the true class, and each column *j* corresponds to the predicted class [104]. The main diagonal entries thus represent correct predictions for each class, while the off-diagonal entries reveal inter-class confusions. The scikit-learn implementation was used in this study [105]. It defines the confusion matrix C such that C_{ij} counts observations of true class i predicted as class j, with an option to normalize over true labels (rows), predicted labels (columns), or the entire matrix to facilitate interpretation under class imbalance [105].

Accuracy is the proportion of all predictions that were correct, define as [104]

$$Acc. = \frac{TP + TN}{TP + TN + FP + FN}$$

This metric indicates overall correctness, but can be misleading if the data is imbalanced [104]. To solve this, precision and recall can be added to the evaluation metrics. Precision, also called positive predictive value, is calculated as

$$Pre. = \frac{TP}{TP\ +\ FP}$$

and represents the fraction of positive predictions that were actually positive [104]. This reflects the efficacy of the model in avoiding false positives. Recall, also known as sensitivity, is calculated as

$$Rec. = \frac{TP}{TP\ +\ FN}$$

and it measures the fraction of actual positives that the model correctly identified [104]. A high recall indicates that the model misses few positive cases (low false negatives), which can be critical in biomedical applications.

F1-score is the harmonic mean of precision and recall, given by [104]

$$F1 = 2 \times \frac{precision \times recall}{precision + recall}$$

The F1-score provides a single balanced metric that is high only when both precision and recall are high.

These metrics were chosen to provide a comprehensive evaluation: accuracy indicates the overall performance, while precision and recall reveal the trade-off between false alarms and missed detections, and F1 offers a balanced summary.

The use of accuracy, precision, recall and F1 is consistent with evaluation protocols in previously covered literature, similar medical studies. For example, Rambhatla et al. evaluated a burn wound classification model (DL4Burn) with overall accuracy as the primary metric, reporting an accuracy of ~ 93.8% for a multimodal CNN combining images and patient data [64]. Similarly, a recent multimodal tool for chronic wound assessment (DM-WAT) reported an accuracy of ~77% and an F1-score of ~0.70 on its test set [10]. Recall (sensitivity) was used as a core metric in a two-stage SVM by Wang et al., who reported a sensitivity score of 73.3% [60], [61]. Alzubaidi et al. used all three, precision, recall and F1 score while evaluating their DFU_QUTNet in classification of normal (healthy) skin class

versus abnormal skin (DFU) class [12], [106]. In line with this literature, these metrics provide a clear understanding of the model's strengths and weaknesses in this study's multi-class classification task.

4.3.3 Model Interpretability

Although CNNs are powerful, they operate in a "black-box" manner, which means that their internal decision-making processes are not directly interpretable by humans [107]. This opaqueness introduces a challenge in the medical domain, where understanding why a model makes a certain prediction is crucial for trust, accountability, and clinical adoption [107]. Transparent reasoning and highlighting the image regions relevant to the model's prediction likely enhance the chance of medical professionals accepting AI assistance instead of blindly trusting new AI model decision-making. To introduce interpretability to this thesis CNN model, Grad-CAM was implemented as a post hoc explanation technique. Grad-CAM is an approach that generates a visual heatmap over the input image, indicating which areas are most influential in the model's prediction [108]. Grad-CAM works by back-propagating the gradient of a target class score with respect to the final convolutional feature maps [108]. These gradients indicate the importance of each spatial location in the feature maps for that class [108]. The gradients are global-average-pooled to compute the importance weights for each feature map [108]. These weights are then used to perform a weighted sum of the forward-activated feature maps [108]. The resulting map is passed through a rectified linear unit (ReLU) activation, producing a heatmap (class activation map) that highlights the spatial regions in the input image that are most relevant to the model's prediction [51], [108]. In other words, regions with higher intensity in the Grad-CAM heat map are those that the CNN considers important for its decision.

Grad-CAM was selected for this project because it is applicable to any CNN architecture without requiring re-training and is class-discriminative, producing localized explanations tied to the specific class output [108]. This method has been used in recent medical imaging studies to provide insights into model decisions. For example, in neuroimaging studies, Grad-CAM visualizations have successfully highlighted disease-relevant brain regions on MRI, helping researchers and clinicians verify that the model's focus aligns with the known pathology [107], [109]. In this study, in the context of wound image analysis, Grad-CAM generates an attention heatmap over each wound image, showing where the network "looks" to determine the wound healing status. This added interpretability allows verification that the

model is attending to appropriate clinical features (e.g. the wound bed or edges, rather than possibly irrelevant background). By inspecting the Grad-CAM outputs, the black-box nature of the CNN can be mitigated, and the model's predictions can be combined with a visual explanation.

4.4 Technologies

4.4.1 TensorFlow

TensorFlow (TF) is an open-source framework for machine learning and deep learning developed by Google Brain, which is designed to execute complex numerical computations on heterogeneous distributed systems [110]. It uses dataflow graphs to represent computations, enabling efficient execution across central processing units (CPUs), GPUs, and tensor processing units (TPUs) [110], [111]. TF provides low-level control and high-level APIs through its tf.keras interface for constructing neural networks, training algorithms, and deploying models [112].

TF was selected for this project mainly because of the support materials provided by the University of Turku and its compatibility with Google Colaboratory (Colab) [16]. Additionally, TF is known for its effective handling of deep neural networks and extensive ecosystem of tools, all of which are accessible within the Google Colab environment [16]. It has been widely used in medical imaging and in biomedical research. For example, Goyal et al. employed TF to create a model for detecting ischemia and infection in DFUs [63]. Fletcher et al. utilized TF to develop two CNNs for identifying surgical site infections [67]. Barata et al. used TF to build a deep learning system for diagnosing skin lesions [113]. These examples highlight the framework's suitability for image-based CNNs in wound care. The broad use of TF in these studies is due to its reliability and trust in its performance, as demonstrated by the high evaluation metrics across these various wound care applications.

4.4.2 KerasTuner

Keras was built to run on top of low-level backends, such as TF, and is now tightly integrated with TF through its tf.keras module [112]. Keras offers modular building blocks for neural network layers, loss functions, and optimizers, enabling users to create complex models with minimal code [114]. In the context of this project, Keras was chosen because its potential was noticed to significantly accelerate the model development cycle while leveraging TF's powerful computation engine. KerasTuner (KT) is a hyperparameter optimization framework

that automates the process of tuning model parameters [115]. It provides a high-level interface to define a search space of hyperparameters, such as network layer sizes, learning rates, and activation functions, and supports various search algorithms, such as random search, Bayesian optimization, and Hyperband [115]. KT integrates with Keras models by allowing users to specify a model-building function that accepts hyperparameters [115]. KT then explores various hyperparameter combinations based on the same underlying architecture to identify those with the best validation performance [115]. This tool was selected for the project to effectively address the challenge of hyperparameter selection, which can significantly influence the model performance.

Although KT is a more recent addition (released in late 2019), it has been adopted in biomedical research [116]. None of the previously mentioned studies mentioned the use of KT, but a notable example is a breast lesion detection study in MRI, where model hyperparameters were tuned using KT to maximize tumor classification performance [117]. In that study, KT's random search algorithm helped configure deep networks (U-Net variants) for optimal lesion detection accuracy [117]. The successful use of KT in such a context confirms the assumption about its practicality for medical imaging problems, where appropriate hyperparameter settings can be the difference between a mediocre model and a state-of-the-art model.

4.4.3 Environment and Other Libraries

All experiments were implemented in Python (v3.11) in a cloud-based Google Colaboratory environment (Colab, versions 2024-11-11–2025-04-09) [16], [92]. Colab is a hosted Jupyter Notebook service that provides free access to CPU computing resources and with a paid subscription to GPU and TPU resources [16]. Integration with Google Drive makes the user completely free from local systems, while the training data can be stored in the cloud, where it is easily accessible from the Colab workspace. Python was the default choice for the core language because of its use with TF and Keras, and the general use of Python libraries such as NumPy and Pandas in data science and machine learning [118]. Colab's convenience with pre-installed libraries, GPU acceleration, automated dependency management, and Python's match with TF, Keras, and all other data science and ML libraries, the development environment basically "built itself."

NumPy (v2.0.2) and Pandas (v2.2.2) were used to handle the data [119], [120]. NumPy provides an N-dimensional array object that underlies most numerical computations in Python

[118]. It is regarded as the *"primary array programming library for the Python language"*, laying the foundation of the scientific Python ecosystem [118]. Virtually all the training data in this project were represented as NumPy arrays. Pandas complements NumPy by introducing DataFrame structures for labeled tabular data [120]. Pandas was used to preprocess, manage, and analyze the tabular data during implementation.

Matplotlib (v3.10.0) and Seaborn (v0.13.2) were used for data visualization and presentation [121], [122]. Matplotlib is a Python plotting library that creates static, animated, and interactive visualizations. In this project, Matplotlib generated figures including training accuracy and loss curves, confusion matrices, and example images with model predictions. Building on Matplotlib, Seaborn provides a high-level interface for statistical graphics and themes, making graphs more readable and representable. Seaborn was used to visualize data distributions and style confusion matrices and classification reports.

Throughout the implementation, tqdm (v4.67.1) was used to display progress bars for iterative operations [123]. Tqdm is a lightweight tool that integrates with Python loops and TensorFlow operations to show progress, which helps monitor long training times and data preprocessing. While tqdm does not affect analytical outcomes, it improves workflow by providing code execution progress information, making it easier to plan work during long training times.

Scikit-learn (v1.6.1) was used for certain machine learning utilities, notably cross-validation and performance evaluation [124]. Scikit-learn is a Python machine learning library offering algorithms and validation techniques through an API [124]. While TensorFlow/Keras handled neural network training, scikit-learn was used for evaluation metrics and cross-validation. The LOGO CV technique was highly impactful, ensuring samples from one mouse were held out together in training and validation folds, preventing data leakage. Scikit-learn metrics like accuracy, precision, recall, F1-score, and confusion matrix evaluated model predictions.

The development environment roles were defined: Google Colab and Python provided the platform, TF and Keras provided the ML framework, NumPy and Pandas managed data, Matplotlib and Seaborn enabled visualization, tqdm assisted with iteration tracking, and scikit-learn provided model performance evaluation. These tools were selected for their popularity, comprehensive documentation, and proven success in scientific research; for example, the previously mentioned study by Barata et al. used most of this technology stack in their implementation [113].

5 Results

5.1 Healing Scale Model Architecture

Healing Scale Model (HSM) was an architecture designed for a 6-class classification task, six classes corresponding the first six post operative days (PODs). The last two days (12 and 14) were removed because healing did not progress significantly after day 10 (as seen in chapter 4.1, **Figure 4**, in the wound closure trajectories), leading to unnecessary confusion in the model. **Figure 7** explains the main logic behind HSM, which created the "Healing Scale". The HSM was trained on the control data and used on the treated data. While the model learned the natural healing process of a control wound, it could be used as a "Healing Scale" for treated wounds. The model used an RGB image, a thermal image, and the wound area measured in square millimeters as inputs. The input day was designed to be from days 2 to 8, because day 0 wound is not treated yet and day 10 wounds have only one possible outcome; responded poorly, because the model's prediction cannot be greater than 10. The two outputs are presented in **Figure 7**, with the following explanations:

1. **Prediction of the treatment response**

   The HSM was used to estimate how the wound has responded to the treatment by predicting the POD: a predicted day later than the true day indicated a good response (the treated wound appeared better healed than controls), whereas a predicted day equal to or earlier than the true day indicated a poor response.

2. **Recommendation of NFC hydrogel composition**

   A recommendation was made based on the HSM output. If the treated wound was responding well, that treatment was recommended to be continued; if the wound was responding poorly, it was suggested to change to the other NFC type or the standard of care.

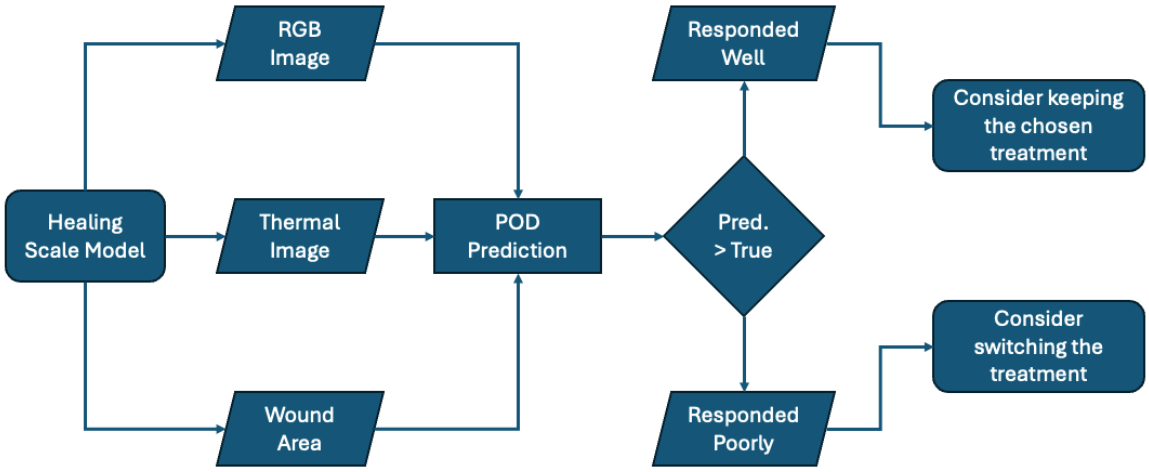

Figure 7 Healing Scale Model architecture. Rounded rectangles indicate the start or end of the whole process, trapezoid indicates an input or an output, sharp cornered rectangle indicates a sub-process, and tilted square indicates a decision point. The model used three different inputs: RGB image, thermal image, and wound area to make a POD prediction for the input data, which were from **treated** wounds. If the prediction was higher than the true label, the input wound was classified as responding well to the treatment, and the chosen treatment was recommended to be kept. If the prediction was lower than or equal to the true label, the wound was classified as responding poorly to the treatment, and the treatment was recommended to be switched.

5.2 Base Model

In the initial version of the HSM, the results appeared unrealistically accurate. Even with simple early versions, a 100% accuracy rate was achieved, including for the most challenging classes, only adjacent days 5 and 6. After examining the Grad-CAM heatmaps, which indicated that the model was focusing on nearly everything except the wound itself, concerns regarding data leakage started to rise. Although no technical data leakage was identified, which would mean that the same files appear at the same time in both, training and validation sets, but subject-level data leakage was discovered. Subject-level data leakage means that although the same files do not leak, the content in two or more different files is from the same subject, in this case, the wound. The stratified k-fold split was performed at the image level rather than at the subject level. Specifically, the dataset consisted of multiple images of the same wound captured from various angles and at different times. Although the data was stratified, the splitting occurred at the image level instead of the subject level. Consequently, images of the same wound could be included in both the training, validation and test folds.

This allowed the model to effectively "see" and "memorize" a particular wound during training and again utilize this memorization during testing, thereby violating the independence of the test set. In other words, the model's 100% accuracy did not represent true generalization to new wounds but rather the memorization of wound-specific features that unintentionally entered the test partition. It became clear that the appropriate method for evaluating this wound classifier was partitioning by subject. LOGO CV was implemented, where all images from a single subject were held as a unit for testing, thus ensuring that no wound appeared in the training, validation, and testing sets simultaneously.

A new version of the HSM was developed from scratch. The model used technologies such as Keras v3.8.0, TensorFlow v2.18.0, and KerasTuner v1.4.7 [114], [115], [125]. All model development was implemented in Python v3.11 using Google Colab [16], [92]. Colab Pro was utilized to get more computational resources, and the NVIDIA A100 40GB GPU was mainly used to train the model. The wound dataset was stored on Google Drive and integrated into the Colab environment for easy accessibility. The model accepted three inputs: an RGB image, a thermal image, and a wound size value. It consisted of three parallel branches that were fused into a single classification head. Two branches were CNNs for image inputs, and the third was a fully connected branch for the scalar input. The RGB image and thermal image CNN branches each included a series of convolutional layers with ReLU activation and max-pooling, followed by a flattening layer [51], [126], [127], [128]. These branches then included a dense layer and dropout regularization to produce compact feature vectors from each image modality [52], [129]. The wound size branch took the wound area as a single input feature and passed it through a dense layer (with ReLU activation) to embed the size into a comparable feature dimension [51], [129]. The outputs of all three branches (RGB image features, thermal image features, and wound size features) were concatenated and fed into a fusion dense layer, followed by a final output layer. The output layer used softmax activation to classify the wound healing stage into one of six classes, corresponding to different PODs (days 0, 2, 5, 6, 8, or 10) [130].

To optimize the architecture and training configuration, the model was implemented as a KerasTuner custom HyperModel [131]. This approach allowed to define tunable hyperparameters for the network architecture and training process. The key hyperparameters included the number of convolutional blocks in each image branch (2–4 CNN layers), the number of filters per convolution (e.g. 16–64), convolutional kernel sizes (3 or 5), the size of dense layers in each branch, and dropout rates. Training parameters, such as batch size (16–

128), number of epochs (10–50), and learning rates (1e-4–1e-2), were tuned as hyperparameters. The build() method of HyperModel constructed the model according to a given set of hyperparameters and returned a compiled Keras model [131]. The model was compiled with Adam optimizer (with the learning rate tunable on logarithmic scale) and categorical crossentropy loss, which is appropriate for multi-class classification [132], [133]. Validation accuracy was tracked as the primary performance metric during training.

KerasTuner with Bayesian optimization was used to automatically tune the HSM model's hyperparameters [134]. Given the limited size of the dataset and the fact that each mouse contributes multiple samples, a LOGO CV was integrated to the hyperparameter tuning process. In each fold of the LOGO CV, all data from one mouse were held out as the validation set, and the model was trained on data from the remaining mice. By splitting at the mouse ID level, it was ensured that the validation data were truly independent of the training data, which improved the generalizability of the model across individuals [135], [136]. This approach mainly guarded against data leakage, which was the main problem with the earlier versions. In order to reach reliable results but save GPU resources, a true nested LOGO CV approach was mimicked, where one mouse at a time was reserved from the LOGO CV hyperparameter tuning process and used as an unseen test group for the best hyperparameters found from the LOGO CV hyperparameter tuning process. With this approach, the LOGO CV hyperparameter tuning process acted as the inner loop of traditional nested LOGO CV and the manual repetition of this process for each mouse at a time acted as the outer loop.

To integrate the custom LOGO CV into the tuning process, a subclass of keras_tuner.Tuner called CVTuner was created that override the run_trial method [137]. In this context, one trial refers to one round of the following processes with one set of hyperparameters. In each trial, the CVTuner performed an inner loop of the LOGO CV on the tuning set. Specifically, for each mouse ID among 1–9 (excluding 6), a model was built with the given hyperparameters and trained on seven of the mice (training fold) and evaluated on the one held-out mouse (validation fold). This provided a validation accuracy for that fold, and after iterating through all nine possible validation mice, the mean validation accuracy was computed for each trial. This mean validation accuracy was reported back to KerasTuner as the objective value for that trial. Console printouts were implemented in CVTuner to track the process, logging the trial number and the mouse ID used as the validation set in each fold. The tuner had to be set to explore a maximum of 30 trials because of memory issues. Sometimes, the GPU ran out of memory on trials 27–29, and the final training had to be performed from the best

hyperparameters from that point. Throughout this process, no data from the reserved testing mouse was used, saving it as a completely separate test set for the final training after the best hyperparameters were found. The tuning phase resulted in a set of best hyperparameters, meaning those that provided the highest average validation accuracy across the LOGO CV folds.

For the base model, only Mouse 10 was used as the unseen test set to lighten the early development process. True nested LOGO CV was implemented later. Mouse 10 was chosen to be the test set only by the numerical order. The best mean validation accuracy across folds was 0.59, and the chosen main hyperparameters were four convolutional blocks for RGB image branch, two for thermal image branch with a learning rate of 1e-4 and 50 epochs. After the optimal hyperparameters were determined, the final HSM was trained on the entire tuning dataset (mice 1–9, excluding 6). This final training run used the selected hyperparameter values and was executed without further validation splitting, since the hyperparameter selection was already complete. The trained model was then evaluated on the held-out Mouse 10 test set to get unbiased performance metrics. The test set predictions were received by feeding the RGB images, thermal images, and size inputs for Mouse 10 through the model to get predicted class probabilities. From these, the predicted class labels were computed for each sample. The test accuracy, precision, recall, and F1-score were calculated by comparing the predicted class labels to the true labels of Mouse 10. The Base Model reached test accuracy of ~0.73, precision of ~0.77, recall of ~0.73 and F1-score of ~0.72, also seen later from the **Figure 8** (all modalities). Compared to the mean validation accuracy across folds of 0.59 the results were quite high, but the prior knowledge about the tuning dataset explained this difference and will be covered later in this chapter.

To get more information on the effect of different modalities, an ablation experiment was done to examine different modality combinations on the classification performance. As shown in **Figure 8**, models integrating multiple modalities outperformed those using any single input alone. For example, combining RGB images with wound size input significantly improved accuracy relative to RGB only, or size only, with RGB + size being 0.75, RGB only 0.70, and size only ~0.17. Notably, RGB and size combination performed better than all three modalities, but still RGB and thermal combination was better than RGB only and reaching even result with all three modalities. Because the different modalities did not seem to contribute equally to decision-making, this sparked the idea of moving towards weighted fusion or even learnable weights for each modality branch.

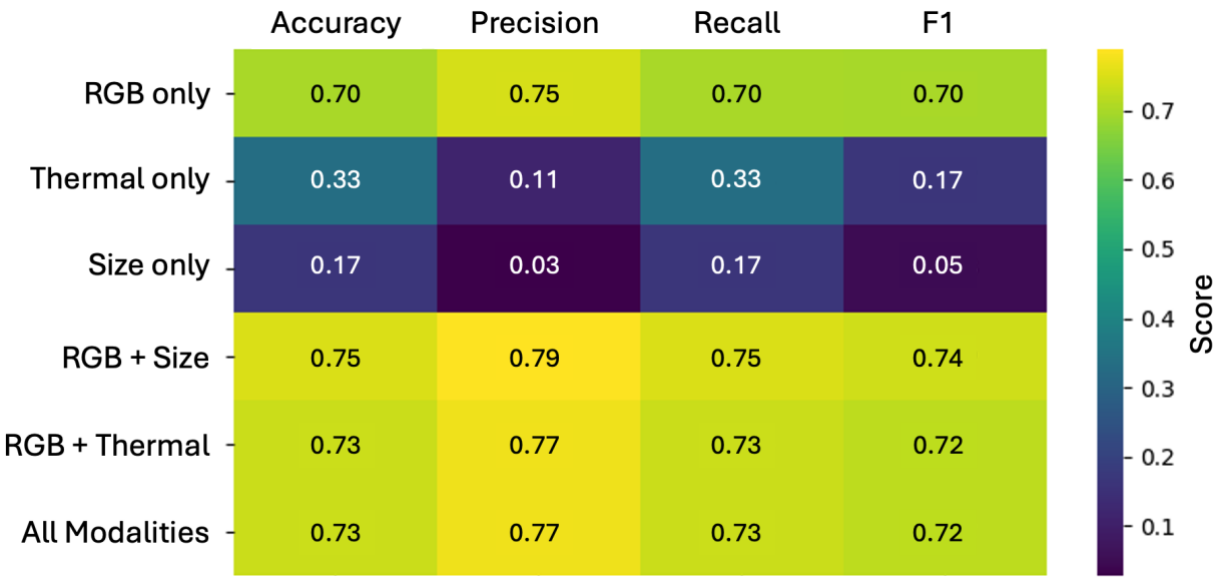

Figure 8 Ablation study, where different branches were shut down by empty inputs to experiment with the contributions of each modality for the Base Model. Accuracy, Precision, Recall and F1 defining the columns of the table and modality combinations defining the rows. The score is color-coded with Viridis heatmap, color scale represented on the right of the table.

5.3 Weighted Fusion Strategies

Based on the results from the ablation experiment (**Figure 8**, Chapter 5.2), the base model was modified to include learnable attention weights for each modality, with the goal of allowing the network to automatically determine the importance of each input type. Instead of simple concatenation, the output of each branch was first passed through a small dense "attention score" head, producing an unnormalized weight (score) for that modality's features [129]. These scores were then combined and normalized using a softmax function to produce attention coefficients for the RGB, thermal, and size branches [130]. Each branch's feature vector was multiplied by its learned weight before fusion, so that the model could dynamically emphasize or de-emphasize modalities for each sample. This tunable-weight HSM architecture introduced additional trainable parameters in the form of an attention mechanism, while the rest of the network architecture remained similar to the base model.

However, introducing learnable modality weights did not improve the performance on the test data. The attentional fusion model performed remarkably worse on Mouse 10. It achieved only ~0.48 accuracy, with a precision of approximately 0.43 and a recall of approximately 0.48 (F1-score of approximately 0.38). These metrics represented a significant drop compared to the ~0.73 accuracy of the base model. However, the results were poor enough to move back to more static approach, such as the base model, but still weighting the different modalities somehow.

Given the difficulties with freely learned weights, a statically weighted fusion was the following approach. In this weighted static fusion model, fixed scalar weights (0–1), later referred to as α values, were assigned to each modality branch's features prior to concatenation, instead of learning attention dynamically. These weights were treated as tunable hyperparameters rather than trainable network parameters, and they were optimized through the hyperparameter search. However, in the first tuning rounds, it was noticed that the tuner had too much "moving space" and it got stuck to very low weight, for example giving the weight of 0.2 for every branch and the results were worse than with the base model. The solution was to guide the tuner with prior knowledge, tightening the tunable space. The static weights were constrained to prevent any modality from being entirely ignored and thermal images were guided to be in a more supportive role; the RGB image and wound size branches were each constrained to a minimum weight of 0.5, while the thermal branch's weights were cut at a maximum of 0.5, reflecting the prior knowledge from the ablation experiment. During the model tuning, the α coefficients were adjusted within these boundaries to find an optimal fusion balance. Aside from these scaling factors for each branch output, the network architecture remained the same as that of the base model. The hyperparameter tuning provided a similar main architecture as the base model tuning, but the new α coefficients were as follows: RGB 0.7, thermal 0.4, and size 0.6. RGB having a higher weight than size and thermal not being in the highest limit of 0.5 was in line with prior expectations. With these new fixed tuned weights, the following results were achieved: an accuracy of 0.83, which was significantly better than the concatenation fusion of the base model (0.73). The precision, recall and F1 score were also 0.83, which were higher than previous fusion strategies.

A confusion matrix was generated for each fusion test results, summarizing how often the model correctly or incorrectly predicted each POD. The confusion matrix was plotted as a heatmap, with the true day on one axis and the predicted day on the other, so that each cell (*i, j*) indicated the count of samples from the true day *i* that were classified as day *j*. In **Figure 9**, the first confusion matrix from the left presents the confusion matrix for the Base Model with Mouse 10 as the test set. The samples were focused around the main diagonal, which was the most important result for the Base Model. Greater than one class mispredictions were not observed and all the misclassifications focused between days 0 to 2 (classes 0 to 1) and days 5 and 6 (classes 2 and 3). Especially, misclassifications between days 5 and 6 were expected because they are the only adjacent days, and it might not be realistic to expect perfect classification within one day difference, i.e. 1-day temporal resolution. The middle confusion

matrix in **Figure 9** for this Attention Based Fusion shows more frequent misclassifications across several classes. The model clearly succeeded in distinguishing three days from each other, day 0, day 6, and day 10, but the days in between were not separated well. The right-side confusion matrix represents the Weighted Static Fusion Model, which substantially improved the performance. The confusion matrix in **Figure 9** also shows major improvements, while the misclassifications focused only on the historically hardest days, the only adjacent days (days 5 and 6).

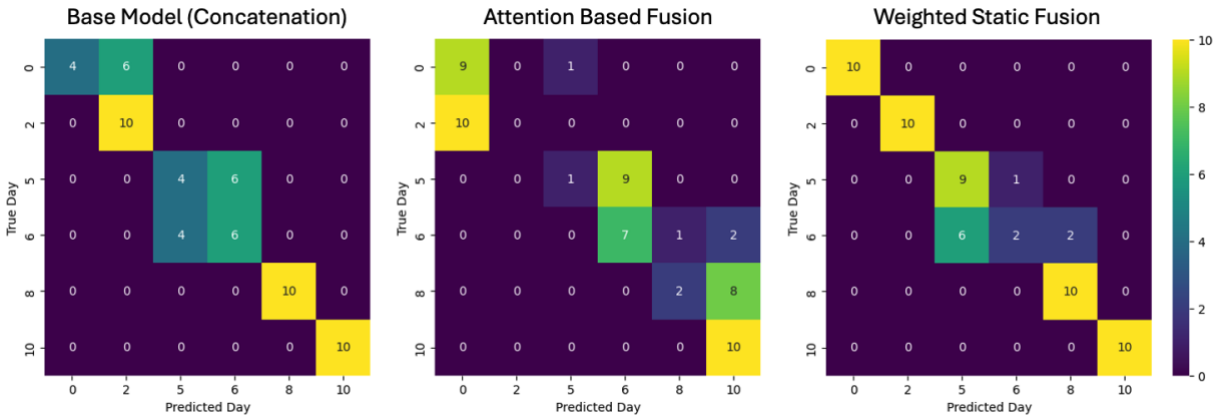

Figure 9 Confusion matrices from left to right: Base Model (Concatenation as fusion strategy), Attention Based Fusion Model and Weighted Static Fusion Model. Cell counts color-coded with Viridis heatmap, color scale represented on the right of the table.

Until this point, the Mouse 10 had been the only unseen test mouse for the Weighted Static Fusion Model, so the same LOGO CV KerasTuner hyperparameter tuning process had to be repeated for all other mice, forming the outer loop of nested LOGO CV. This approach ensured that the good results received from Mouse 10 were not coincidental. However, some variation occurred. As seen from the **Figure 10**, outer-loop test accuracy was ranging from 0.38 to 0.92 and F1 score from 0.30 to 0.91. The mean test accuracy over the nine mice was ~0.71 with a comparatively large sample standard deviation of 0.16, indicating substantial between-subject variability. In contrast, the best mean validation accuracy recorded during the inner KerasTuner LOGO CV loop was tightly clustered around ~0.70, with a standard deviation of ~0.05. This pattern shows that the inner-loop performance was relatively stable across training folds, where generalization to a held-out subject depended strongly on which mouse was excluded in the outer loop, as summarized in **Figure 10**.

At the individual level, the highest test accuracy was observed when Mouse 5 was held out (accuracy 0.92; F1 0.91), followed by Mouse 10 (accuracy 0.85; F1 0.82), and Mouse 3 (accuracy 0.80; F1 0.76). The lowest test accuracy occurred for Mouse 2 (accuracy 0.38; F1

0.30), followed by Mouse 1 (accuracy 0.55; F1 0.50), and Mouse 9 (accuracy 0.63; F1 0.60). These subject-specific outcomes were noted to align with the wound-closure trajectories in **Figure 4** from chapter 4.1 and with the R^2 scores from those trajectories displayed in **Figure 10**. All mice, with over 0.90 R^2 score had accuracy over 0.70, Mouse 10 being an exception with R^2 of 0.82, but with the second highest accuracy of 0.85.

The inner-loop best mean validation accuracy varied from 0.65 to 0.78 depending on which mouse was excluded from training, as displayed in **Figure 10**. Notably, excluding an "easy" subject, e.g. Mouse 5, coincided with lower inner-loop means: 0.66 when Mouse 5 was held out and 0.65 when Mouse 10 was held out. Conversely, excluding the "hardest" subject produced the highest inner loop mean: 0.78 when Mouse 2 was held out. Quantitatively, the per-mouse best mean validation accuracy was negatively correlated with the corresponding outer-loop test accuracy (Pearson $r = -0.67$; Spearman $\rho = -0.51$). This pattern suggests that, in this small group, removing an easy, well-performed subject from the training pool reduced the ease of the inner validation task, where removing a difficult subject increased the inner validation, even though the ultimate generalization to that difficult subject remained poor.

From a CV standpoint, the conflict in dispersion between inner-loop validation and outer-loop testing was found to be informative. As seen from **Figure 10**, the inner-loop best mean validation accuracy was distributed around ~70, yet the outer-loop test accuracy spanned 0.54 points across subjects. This gap indicates that the specific composition of the training set (especially the presence or absence of easy, representative subjects) affected generalization. In a dataset of this size, excluding a highly informative subject, such as Mouse 5 or 10 reduced the average inner-loop mean validation accuracy by ~0.12–0.13 relative to the setting where Mouse 2 was excluded, while simultaneously providing strong outer-loop test performance on those easy subjects. These results imply that under a limited number of subjects, inner-loop metrics alone can overestimate the ability to generalize to the hardest held-out subject and underestimate the contribution of easy but representative subjects to stable learning.

Figure 10 also shows that across subjects, the mean absolute error (MAE, in days) was heterogeneous but generally small relative to the two-day spacing of most class boundaries in the six-class task. The largest MAE occurred for Mouse 2 (1.73) and Mouse 1 (1.43), where the smallest MAE was observed for Mouse 10 (0.18) and Mouse 5 (0.23). Intermediate errors were recorded for Mouse 3 (0.37), Mouse 4 (0.57), Mouse 7 (0.97), Mouse 8 (0.52), and Mouse 9 (0.85). The mean for these results was 0.76 days with a sample standard deviation of

0.54 days. Because a single class step typically corresponds to two days (except the only adjacent days 5 and 6), even the worst-case mean error remained below one class average. The MAE aligned closely with the previously reported nested LOGO CV results. Numerically, MAE showed a strong negative correlation with test accuracy (Pearson r = –0.93 and Spearman ρ = –0.89), indicating that subjects with low MAE were the same subjects with high classification performance. MAE also tracked the linearity of wound closure reported earlier, and in **Figure 10**. A moderate negative association between MAE and the R^2 of the closure curve (r = –0.37 and ρ = –0.48), suggesting that more linear trajectories tended to provide smaller absolute timing errors, although linearity did not fully explain the variance across subjects. A mean MAE of 0.76 days was well below the two-day spacing of most class boundaries, meaning that on average, predictions deviated by less than one day, less than a half class step. However, a standard deviation of 0.54 reflects notable subject-level differences in the model performance, but this has been the trend with the model's performance, due to the nature of the small and heterogeneous dataset.

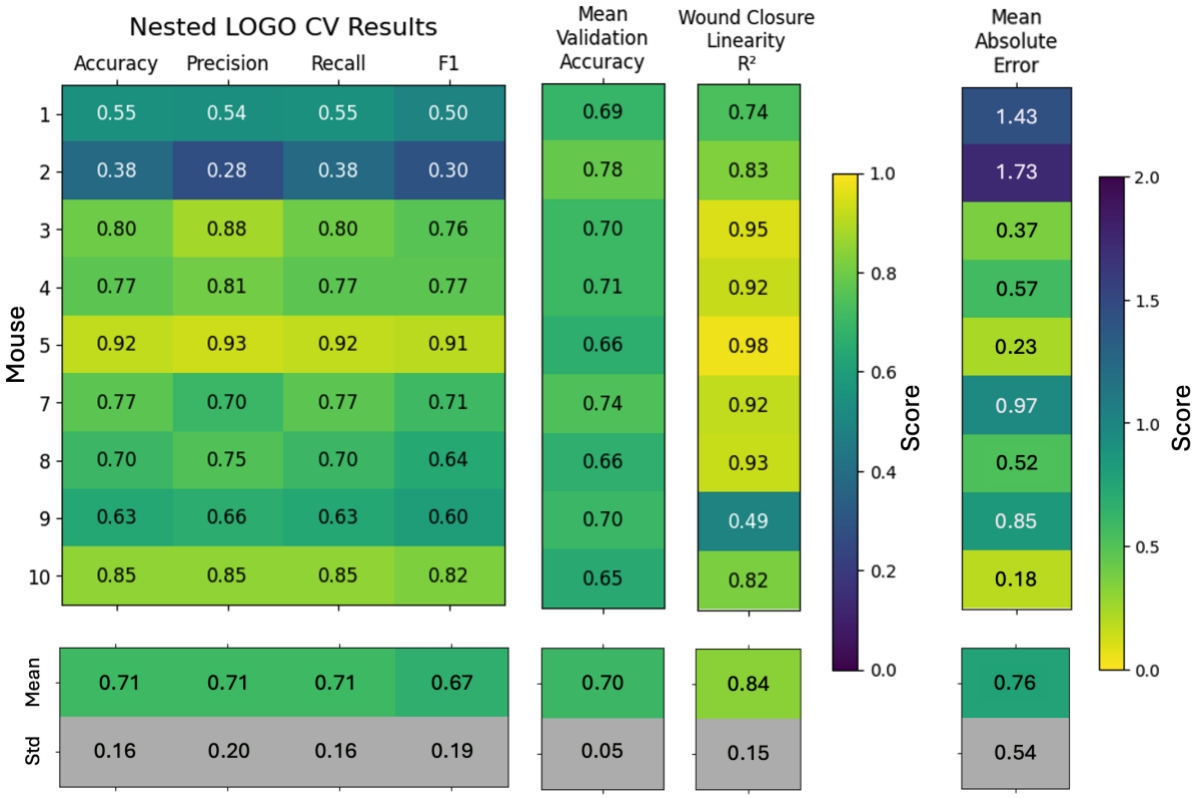

Figure 10 Nested LOGO CV results (left), Mouse IDs in rows and accuracy, precision, recall, and F1 score in columns. The best mean validation accuracy from the KerasTuner trials (middle left) and wound closure linearity (R^2) score (middle right). Mean absolute error for each Mouse (right). Below each column are separated rows for mean and standard deviation for calculated for each column. Cell scores are color-coded with Viridis heatmap, color scale represented on the right of the table.

As displayed in **Figure 10**, the best-performing Mouse 5 combined the highest linearity (R^2 = 0.98) with the highest accuracy and F1, and as mentioned earlier: all mice, with over 0.90 R^2 score had an accuracy over 0.70. However, linearity did not fully determine predictability. Mouse 2 showed relatively high linearity (R^2 = 0.83) but very low test accuracy. In the Mouse 2 case, the R^2 score did not tell the whole truth. When looking at **Figure 4** from chapter 4.1, or **Figure 11**, Mouse 2 closed extremely rapidly compared to the other mice. High linearity score did come mainly from days 5–10, where the Mouse 2 wound was almost fully closed, making especially the days 5–8 hard to predict due to the little change in wound closure between the days. These details indicate that differences in early phase healing and inter-mouse wound size differences can confuse a POD-based classifier, even when the target wound's overall healing trajectory is approximately linear.

As seen earlier from **Figure 10**, the Mouse 2 test iteration was by far the worst performing iteration. In addition to the observations based on wound closure linearity, visual cues on poor performance were also found, and the model's reasoning became clearer. **Figure 11** displays the Mouse 2 confusion matrix from the Mouse 2 test iteration. For comparison, next to it is the Mouse 3 confusion matrix from the Mouse 3 test iteration. Below them are the RGB images from both mice control wounds. The Mouse 2 had a significant leap in wound closure between days 2 and 5, where it reached almost its full closure (~3 mm^2), compared to the last days (~2 mm^2; **Figure 4**, chapter 4.1). Where day 5 wound from Mouse 3 was approximately five times larger (~16 mm^2) on day 5, compared to the day 5 wound from Mouse 2 and did not reach Mouse 2's day 5 wound closure even on day 10 (~5 mm^2). This Mouse 2 rapid wound closure is also seen in the confusion matrix misclassifications. Mouse 3 had 100% accuracy on day 5, where Mouse 2 had 0%. When the two day 5 wounds were visually compared, it was very understandable why the model classified day 5 wounds from Mouse 2 as day 8 wounds. The misclassifications look reasonable compared to visual observations, for example days 5–10 could be from the same day, if the wound size (**Figure 4**, chapter 4.1) and visual appearance is compared to the Mouse 3 (and most of the other mice). As seen from **Figure 11**, the model has classified days 5–10 between the last two days, which can be seen objectively good predictions when considering the wound closure deviation from the other mice (**Figure 4**, chapter 4.1). Also, the day 2 misclassifications seem reasonable, the Mouse 2, day 2 wound could by visual cues be also a day 0 wound from another mice. These observations show that the model is making objectively great predictions, even though they

are penalized in the testing results. This also highlights the problematic nature of POD labeling, which is discussed later.

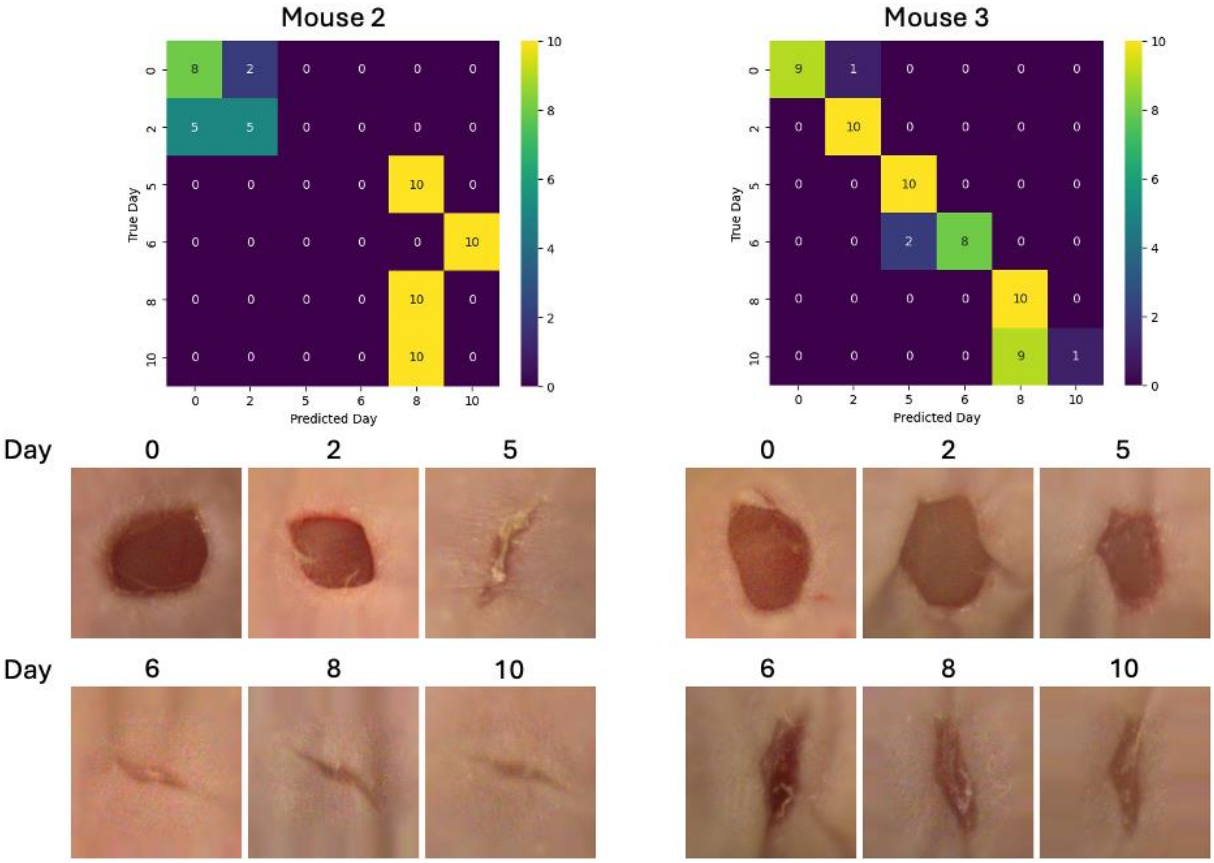

Figure 11 Mouse 2 and 3 confusion matrix and wound closure comparison. Confusion matrices in the top row, cell counts color-coded with Viridis heatmap, and color scale represented on the right of each confusion matrix. Compared with the RGB images (two bottom rows), with day labels above the individual RGB images.

In order to proceed to testing on treated data, which is ultimately the main use case of HSM, a decision was made to stop optimizing the hyperparameter selections further and leave the hyperparameter tuning processes to their current state. The median of each hyperparameter set from each test iteration was calculated, and they were collected to a final model configuration. The final model was trained with all data, so the testing accuracy could not be calculated due to the lack of unseen test data. However, the nested LOGO CV results showed that median hyperparameter selection is capable on objective generalization. The final model was trained for 50 epochs with a batch size of 64 and a learning rate of approximately 2.17×10^{-4}. The RGB, thermal and size inputs were fused with weighting coefficients of 0.8 for RGB, 0.2 for thermal and 0.6 for size, these aligned well with the hypothesis and the ablation experiment from **Figure 8**, chapter 5.2. The RGB branch had three convolutional blocks with filter sizes ranging from 32 to 48, using kernel sizes of 3 and 5, these were followed by a dropout of 0.3

and a 96-unit dense layer. The thermal branch consisted of two convolutional blocks with 48 filters each at kernel sizes of 5 and 3, followed by a dropout of 0.4 and a 64-unit dense layer. The size input, being a single scalar, was passed through a 32-unit dense layer before being concatenated with the digital and thermal features into a 128-unit fusion layer with a 0.3 dropout.

As represented in **Figure 7** in chapter 5.1, the HSM is trained on control data, and tested on the treated data, which is completely unseen for the model. **Figure 12** displays the confusion matrix for the final model tested on treated data, with some example classifications next to it. The most neutral healing stage being day 2, where the treatment might not have affected yet on the appearance of the target wound. The misclassifications increased progressively on days 5, 6, and 8, indicating that the treatment effects might have started increasingly affecting the appearance of the target wound. These predictions on the treated data were also visually observed, and the results were promising.

Figure 12 shows RGB image, thermal image, and Grad-CAM result for well and poorly responded wounds from each true class, excluding days 0 and 10. Day 0 wounds were excluded as they were untreated, and day 10 wounds were excluded since they had only one outcome: poor response. Visually, the displayed well responded wounds in **Figure 12** appeared well closed and healed compared to the true day, and significant thermal differences were not observed in the thermal images. Grad-CAM results were also promising, although not as clear as with the poorly responded ones; true day 6 prediction has focused also on the noise caused by the augmentation and true label 10 images has a lot of high activation on the surrounding skin, which is not necessarily a bad thing, because the amount of healed surrounding skin can also be a measure of the healing stage. The clearest difference was the true day 2 wound, which had no activation on Grad-CAM. The image is from a very close distance, so the model might have relied more on the size information, rather than visual cues.

The poorly responded predictions also appear reasonable in **Figure 12**. Visually, at least true day 2 could be from the predicted day 0, but true days 5 and 6 are more debatable. True day 5 has some dried skin at the edges, and true day 6 wound has some NFC hydrogel or wound exudate on top of it, which makes them less understandable to be predicted for day 0 wound. However, all the RGB images are relatively to their true label clearly more open wounds and having very notable thermal differences in the thermal images, especially when compared to the well responders in **Figure 12**. Although the poor responder day-level predictions can be

debated whether they can be classified as day 0 wounds, they still can be seen clearly poor responders. In the end, the HSM classifications validate the model's reasonable decision-making, while all the well responded classified wounds look well-healed (great closure and little thermal differences) compared to the true label, and especially compared with the same true label poor responders. As well as poorly responded predictions from **Figure 12** look significantly worse healed compared to the true labels, and their corresponding well healed pairs. Lastly, the sharp Grad-CAM images also validate the models' reasoning where it gives high activation on the biologically relevant areas, such as wound edges and the wound bed, with consistently the least activation on the image edges, where there are no relevant features regarding wound healing.

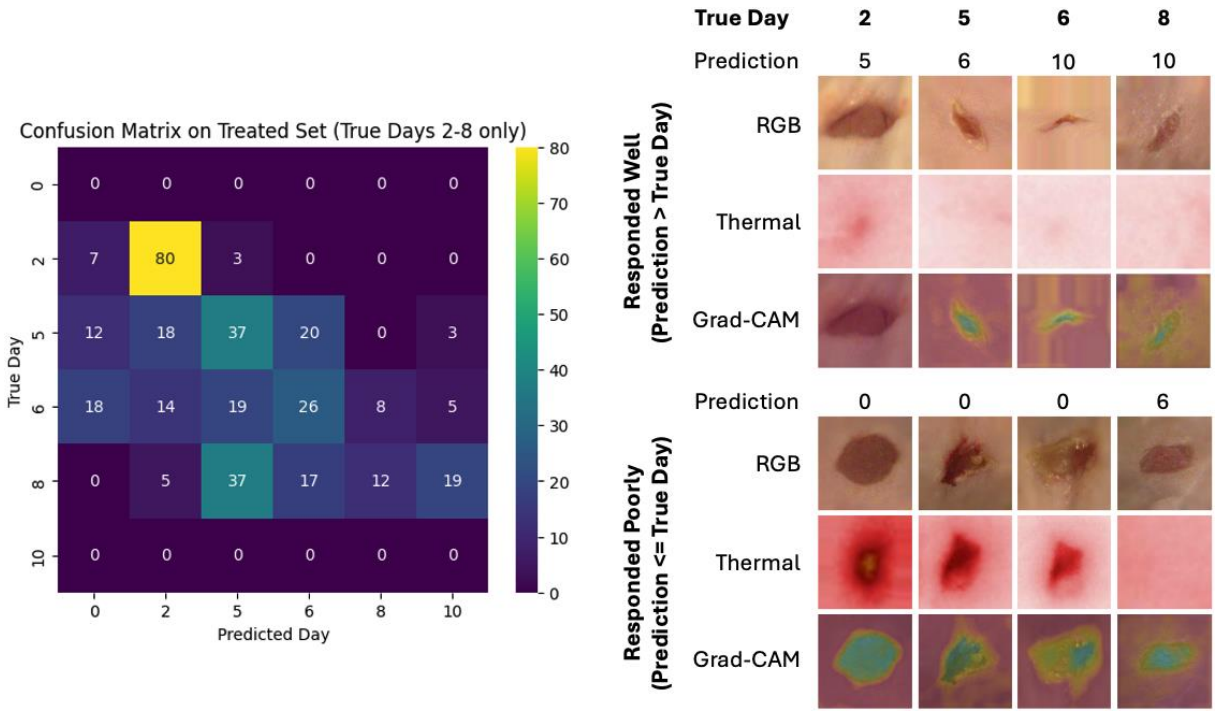

Figure 12 Treated dataset predictions using median hyperparameters from nested LOGO CV. Confusion matrix shows cell counts with Viridis heatmap color scale, with example predictions displaying RGB, thermal, and Grad-CAM images with True and Predicted labels. Grad-CAM heatmaps indicate activation levels: yellow (highest), blue (medium), purple (lowest).

Overall, these results demonstrated that these models can consistently and reasonably classify treated wounds as well or poorly healed, with visual observations supporting most of these predictions. The chosen hyperparameter pool was consistent through different training, validation and testing combinations, making the architecture trustable to be used in evaluation on treated mice data, and considerable for future development and integration for human data.

6 Discussion

6.1 Classification Challenges

The classification scheme and labeling used in this study presented several challenges that impacted model performance and implementation. One notable challenge was the binary classification nature of the treated wound classification (responded well vs. poorly), which was only based on the PODs. The treatment response outcome was basically defined by shrinkage in area, while (as discussed in the chapter 2.1) healthcare professionals assess healing through multiple factors, including granulation tissue quality, infection status, or perfusion, in addition to closure [18]. For example, regarding to treatment response, even though Koivuniemi et al. noted that no statistically significant differences were observed between NFC and copolymer dressings regarding wound healing time, epithelialization, experience of pain, or transepidermal water loss (TEWL) [7]. Still, the NFC dressing showed advantages in scar appearance, particularly in thickness and vascularity, as assessed by the POSAS, and significant improvements in skin elasticity were observed in donor sites treated with NFC dressing compared to those treated with copolymer dressing [7]. Wounds heal at variable rates; regarding this dataset, a great example was day 5 control wounds between mice 2 and 3, from **Figure 11**, chapter 5.3. Supported with **Figure 4** from chapter 4.1, Mouse 2, day 5 wound was approximately five times smaller on day 5 and the RGB images in **Figure 11** (chapter 5.3) looked to be in a completely different healing stage, Mouse 2 wound being almost completely closed and Mouse 3 wound being wide open.

The outer-loop results for Mouse 1 and Mouse 2, displayed in **Figure 10**, illustrate this limitation of the POD labels. Both closed substantially faster than the rest, and by day 6 wounds were more closed than many other mice wound at day 10. In such cases, the model produced clinically sensible outputs in the sense of ordering wound healing stages; however, these outputs were penalized when mapped strictly to POD labels. At the opposite extreme, Mouse 9's unusually slow wound closure and small day 0 wound further increased confusion; its low linearity ($R^2 = 0.49$) coincided with below-average test accuracy (0.63) and, more notably, 0% day 0 test accuracy. Taken together, **Figure 4** compared to **Figure 11** RGB image examples helps explain why subject-level differences such as abnormal wound closure progression or differences in operative day wound size, for example Mouse 9 ~12 mm^2 versus Mouse 4 ~40 mm^2 day 0 wounds increase between-subject variance in nested LOGO CV. Therefore, using the POD as a class label is "inherently noisy" and can confuse the model

because the POD number alone does not reliably indicate the healing status. This contributed to the model's "misclassification" of some well-healing wounds as poorly healed, simply because the timeline did not align with the expected size reduction. A more objective labeling strategy would be needed to align model predictions with clinical reality.

A potential improvement in defining classes would be to use established clinical wound classification systems instead of labels, such as PODs. For example, the Wagner grade or UT classification discussed in chapter 2.1. Both, the Wagner and UT classification have been studied to have great predictive value, which would make them good base for an AI wound classifier for predicting wound healing outcomes [26]. These scales provide a medically relevant categorization of wound severity and healing stage. If the model was trained to classify wounds according to such clinical grades, combined with significantly larger dataset, the model might learn more relevant features without confusion of only healing time related labels, such as POD. A small but infected wound (high Wagner grade) would be appropriately recognized by the model as a poorly healed wound, where a wound that might remain large but has a healthy granulating base (lower Wagner/UT grade) could be classified as well healed/responded. Another example, if the classification was done based on more biological factors, such as ischemia, neuropathy or neuroischemia discussed in chapter 2.1. An algorithm could distinguish a neuropathic plantar ulcer with a callused rim from an ischemic toe ulcer with dry gangrene and provide specialized treatment recommendations for both conditions. The chronic nature of diabetic ulcers is also evident histologically; they often show a prolonged inflammatory milieu, matrix metalloproteinases that hinder repair, and biofilm-producing bacteria that together prevent wound healing [18]. In the future, one branch of a multimodal AI model for wound care could be trained on histological samples of different types of wounds and healing phases to improve accuracy. Unfortunately, Wagner, UT, ischemia, neuropathy, or neuroischemia labeled data were not available for this study, and simpler, healing time-based labeling had to be implemented.

6.2 Model Performance

The model's performance across different modality combinations revealed insights into the contribution of each data modality and potential issues of modality integration. As presented in **Figure 8**, in chapter 5.2, the baseline using only RGB images achieved a classification accuracy of 0.70. When the wound size was combined with the RGB images, the accuracy improved to 0.75, which suggests that providing size context alongside the RGB image

significantly improved the predictions. These results aligned with the literature in chapter 2.3, for example, with the study by Rambhatla et al., where they developed a system (DL4Burn) that integrated burn wound photographs with patient demographics/injury data to predict the need for surgical intervention [64]. The inclusion of patient context boosted accuracy, and their multimodal model was approximately 93.8% accurate, where the CNN using images alone achieved approximately 81.0% on the same task [64]. In this thesis, the difference between the unimodal and multimodal was not as significant (only 5%, RGB only compared to RGB + size, **Figure 8** in chapter 5.2) multimodal still outperforming the unimodal one, making the hypothesis of this thesis correct. However, by adding the fixed tunable weights for the modalities (discussed in chapter 5.3), the overall performance increased approximately 10%, utilizing the multimodality better than the concatenated fusion in the base model. Making the comparison between unimodal and multimodal more complex, because the different modalities are weighted differently in the multimodal model, but the increase in overall performance suggests that the multimodality adds even more value, when utilized correctly.

Interestingly, when thermal images were added as a third modality, the base model accuracy decreased slightly to ~0.73 (**Figure 8**, chapter 5.2). It would be easy to expect that adding more information would improve, or at least maintain, performance. Although one part of the hypothesis in chapter 3 was that thermal images might even confuse the model, due to their low quality and because they provided significant thermal differences only from day 0 to 5, and the last three days were basically identical what comes to thermal images. However, the combination of RGB and thermal data still performed better than the RGB image only version, so the hypothesis was not completely correct. There are a few possible explanations for this dip in performance when using all three modalities, but a slight improvement when adding only thermal images next to RGB images. As covered in chapter 4.1, thermal differences appeared mostly on the early days, which suggests that thermal images likely help the model with the early day predictions but confuse or do not make a difference in the late day predictions. However, size data provides consistent progressive value through all days (**Figure 4**, chapter 4.1), which might explain the slight difference between RGB + size and RGB + thermal combination performance differences, and with all three modalities, the thermal images could lead the model's decision boundary to become less clear on the late days, causing a minor performance drop compared to RGB + size combination. Dynamic weighting through days for thermal input could be a better solution than static weighting,

meaning that the model could give a relatively higher weight to thermal images in the early days and drop the weight in the later days. Overall, this slight degradation in the three-modal case suggests more of an integration challenge rather than a fundamental uselessness of the thermal images.

As used by Busaranuvong et al. with their multimodal AI-application for wound care, the Grad-CAM maps provided a visual interpretation of this thesis model decision-making process [66]. The Grad-CAM maps (**Figure 12**, chapter 5.3**)** revealed that the model's attention often aligned with the medically relevant regions of the wound images, such as the borders and wound bed. In many correctly classified cases, the model focused on the wound borders with the highest activation (yellow). This might indicate that the model makes decisions mostly based on the structure or shape of the wound borders, where new epithelial tissue would form. Usually, in the same mapping, the medium level (blue) activation is focused on the wound center (wound bed). This might indicate that the model assigned a medium-high value to the color, texture, or area of the wound bed. **Figure 12** Grad-CAM images showed great insights on the model's decision making. With poorly responded wounds, the first three examples from true days 2, 5, and 6 were predicted as day 0 wounds. Grad-CAM heatmaps showed that the model was looking at the right areas, but the predictions revealed that it did not look at details such as dry skin on the edges or exudate on the wound bed because none of these did not appear on day 0 in the training data. This might indicate that the model is mostly looking at just the area of the wound and not giving attention to smaller details. Additionally, the model sometimes focused on the highest value on the surrounding skin, and in these cases, the wound itself usually had no significant activation, medium (blue) at most. These activation mappings were more noticeable on worse-quality images, where the model might have focused on the noise rather than the relevant features. However, in some cases, the surrounding skin activation was very sharp around the wound edges; therefore, the model was probably looking for the amount of "healthy skin," which is reversible for the wound area. These activation mappings were not seen as failures; the model still looked at the shape and size of the wound, but through the shape of the surrounding skin.

6.3 Future and Recommendations

The outcomes of this study opened several directions for future work. The HSM architecture was considered as a very promising way for achieving predictive results out of a model for wound care. To utilize this architecture further, a more diverse and clinically representative

dataset is required. Although the current implementation is a valuable POC, it was limited by relatively homogeneous wound data from mice, not from humans, and too simple labeling setup. The HSM architecture combined with the Wagner, UT or more biological, such as ischemic, neuropathic, or neuroischemic classification could be a combination to develop the POC architecture for real-world human data translatable to a clinical context. This would likely open room for significant performance improvements and importantly, make the model's predictions more clinically interpretable. Rather than simply outputting whether a wound has responded well or poorly, a more advanced model might be able to suggest which factors, such as infection signs or poor blood supply are present and guide treatment by real medical factors, as covered in chapter 2.1.

Another significant direction is to investigate how to better utilize transfer learning. With the initial versions of this thesis model, transfer learning failed to improve the performance, making it possible to predict only a single class, and after numerous layers of unfreezing, it was able to distinguish between the maximum and minimum classes. These results were not promising enough to move more towards transfer learning, although many studies covered in chapter 2.4.4 had succeeded in integrating transfer learning to wound care models. The potential causes identified included an "imbalanced dataset," not in terms of class distribution, but in content similarity. Specifically, the six classes were very similar, leading the pre-trained networks to predict them as a single class. This suggests that the pre-trained features were not natively compatible with this task, while the domain gap between everyday images and wound images is very large. Wounds have unique color distributions and textures, unlike everyday photographs, meaning that the features of the early layers may not have been very useful. This also suggests that the fine-tuning process was miscalibrated because, as already covered in chapter 2.4.4, many studies have achieved great results in wound-related ML through transfer learning. Another potential issue was optimization, such as an inappropriate learning rate. Reducing the learning rate occasionally resulted in the model predicting classes 0 and 5 (the first and last), thus distinguishing between fresh and nearly healed wounds, but all intermediate classes were misclassified into these two extremes. However, some studies have noted that transfer learning does not always provide improvements for medical imaging tasks with small datasets, which supports the use of the custom CNN approach [83], [138]. Ultimately, the root cause of this transfer learning failure was not identified, but it would be a highly considerable direction for future work.

The most important aspect of model performance and evaluations in this study was avoiding data leakage, covered in chapter 5.2. Data leakage occurred for more than half of the study timeframe, causing major performance overestimation. This happened because multiple angles of the same wound existed in every case, which was not considered during k-fold CV splits. While exact files were not leaked between training and validation splits, data leakage occurred instead at the file content level, meaning the model did not have to generalize to entirely new wound data. A subject-level split was implemented; images of a particular wound were included in either training or validation sets at a time, with testing from completely separate subjects. Such data leakage can occur in medical imaging, where one subject can contribute multiple data types, such as images and EHR data. When separation between training and test data is compromised, models can achieve unrealistically high accuracies using leaked information. In this study, having the same mouse wound data in both training and testing splits significantly inflated the measured performance.

Finally, a larger and more diverse human-based datasets are needed to get the best out of multimodal AI models for healthcare. Both data quantity and versatile, detailed labeling are essential. However, as discussed in chapter 2.4.4, and noted by Cheplygina et al: a large amount of unlabeled data exists, but without detailed labeling, advanced models remain challenging to build [83]. Complex supervised predictive models require detailed labeling for their tasks. For example, if a model was created to classify wounds based on Wagner or UT classification systems, it would need a dataset labeled with these systems, or for different wound types like ischemic, neuropathic, or neuroischemic wounds, corresponding labels would be needed. In order to predict healing outcomes, labels must include also treatment methods and healing outcomes. Ideally, the data should come from multiple clinics using different devices, enabling the model to generalize worldwide in varying conditions. Unfortunately, such datasets are not currently publicly available.

As Hunter Moseley cautioned in a Nature column that “your results are only as good as your data”, even the most sophisticated AI architectures can be fundamentally ruined by data leakage, small and low-quality datasets, and imprecise labeling [139]. In this study, restrictive labeling, limited size and heterogeneity of the wound data likely constrained model generalization; however, the POC model served its purpose by revealing the possible pitfalls and bad practices, making the study considered successful.

7 Conclusions

In this study, a multimodal AI model was developed to assess wound treatment responses by integrating RGB images, thermal images, and wound area measurements. It was shown that allowing images or measurements from the same wound to appear in all training, validation, and testing sets can substantially inflate the model performance. In medical imaging, models can learn subject-specific cues, such as skin color, wound shape, and camera angle, rather than true pathological features. Therefore, strict wound- or subject-level splitting is necessary to avoid data leakage and to get realistic estimates of generalization.

It was also observed that the POD is a poor label for an AI model in wound healing stage classification. Labeling wounds only by healing days introduced significant confusion into the classification task. Individual wounds progressed through healing stages at different phase. Consequently, using POD as a class label led to many misclassifications. Rather than relying on monitoring time points, future studies should adopt labels based on biological or clinical criteria, such as Wagner or UT grades, the presence of ischemia, infection status, and granulation quality, which more accurately reflect biological healing phases.

Among the three modalities, RGB images contributed the most to the predictions. The color, texture, and morphological features visible in RGB images were found to correlate strongly with the healing status. Wound area measurements offered complementary value to provide context regarding the size of the wound in the RGB images, significantly improving the accuracy of the model compared to the unimodal model. Thermal images also added value, especially in the early days, but also confused decision-making when all three modalities were in use. Therefore, RGB images should be prioritized, ensuring high resolution and consistent imaging angles, with wound size and thermal data integrated to add more context and support the model's decisions.

Overall, this multimodal POC model for wound care demonstrated that a multimodal AI model can capture the relevant indicators of different wound healing stages. However, future improvements will depend on careful data handling during model training to eliminate leakage, labeling based on biological and clinical wound characteristics rather than healing time alone, and focus on high-quality RGB imaging combined with size metrics and thermal profiles. By addressing these factors, AI-driven wound assessment tools can achieve greater reliability and clinical relevance.

References

[1] C. K. Sen, “Human Wound and Its Burden: Updated 2022 Compendium of Estimates,” *Adv Wound Care (New Rochelle)*, vol. 12, no. 12, pp. 657–670, Dec. 2023, doi: 10.1089/wound.2023.0150.

[2] L. Martinengo *et al.*, “Prevalence of chronic wounds in the general population: systematic review and meta-analysis of observational studies,” *Annals of Epidemiology*, vol. 29, pp. 8–15, Jan. 2019, doi: 10.1016/j.annepidem.2018.10.005.

[3] P. Zhang, J. Lu, Y. Jing, S. Tang, D. Zhu, and Y. Bi, “Global epidemiology of diabetic foot ulceration: a systematic review and meta-analysis †,” *Ann Med*, vol. 49, no. 2, pp. 106–116, Mar. 2017, doi: 10.1080/07853890.2016.1231932.

[4] D. G. Armstrong, A. J. M. Boulton, and S. A. Bus, “Diabetic Foot Ulcers and Their Recurrence,” *New England Journal of Medicine*, vol. 376, no. 24, pp. 2367–2375, June 2017, doi: 10.1056/NEJMra1615439.

[5] D. G. Armstrong, T.-W. Tan, A. J. M. Boulton, and S. A. Bus, “Diabetic Foot Ulcers: A Review,” *JAMA*, vol. 330, no. 1, pp. 62–75, July 2023, doi: 10.1001/jama.2023.10578.

[6] S. R. Nussbaum *et al.*, “An Economic Evaluation of the Impact, Cost, and Medicare Policy Implications of Chronic Nonhealing Wounds,” *Value in Health*, vol. 21, no. 1, pp. 27–32, Jan. 2018, doi: 10.1016/j.jval.2017.07.007.

[7] R. Koivuniemi *et al.*, “Clinical Study of Nanofibrillar Cellulose Hydrogel Dressing for Skin Graft Donor Site Treatment,” *Advances in Wound Care*, vol. 9, no. 4, pp. 199–210, Apr. 2020, doi: 10.1089/wound.2019.0982.

[8] R. G. Frykberg and J. Banks, “Challenges in the Treatment of Chronic Wounds,” *Adv Wound Care (New Rochelle)*, vol. 4, no. 9, pp. 560–582, Sept. 2015, doi: 10.1089/wound.2015.0635.

[9] O. Ganesan, M. X. Morris, L. Guo, and D. Orgill, “A review of artificial intelligence in wound care,” *ais*, vol. 4, no. 4, pp. 364–375, Nov. 2024, doi: 10.20517/ais.2024.68.

[10] R. S. Fard *et al.*, “Multimodal AI on Wound Images and Clinical Notes for Home Patient Referral,” Jan. 22, 2025, *arXiv*: arXiv:2501.13247. doi: 10.48550/arXiv.2501.13247.

[11] D. M. Anisuzzaman, Y. Patel, B. Rostami, J. Niezgoda, S. Gopalakrishnan, and Z. Yu, “Multi-modal wound classification using wound image and location by deep neural network,” *Sci Rep*, vol. 12, no. 1, p. 20057, Nov. 2022, doi: 10.1038/s41598-022-21813-0.

[12] J. Zhang, Y. Qiu, L. Peng, Q. Zhou, Z. Wang, and M. Qi, “A comprehensive review of methods based on deep learning for diabetes-related foot ulcers,” *Front Endocrinol (Lausanne)*, vol. 13, p. 945020, 2022, doi: 10.3389/fendo.2022.945020.

[13] M. Berezo, J. Budman, D. Deutscher, C. T. Hess, K. Smith, and D. Hayes, “Predicting Chronic Wound Healing Time Using Machine Learning,” *Advances in Wound Care*, vol. 11, no. 6, pp. 281–296, June 2022, doi: 10.1089/wound.2021.0073.

[14] “AI Academic Writing Tool - Online English Language Check | Paperpal.” Accessed: June 01, 2025. [Online]. Available: https://paperpal.com/

[15] “Introducing OpenAI o3 and o4-mini.” Accessed: June 01, 2025. [Online]. Available: https://openai.com/index/introducing-o3-and-o4-mini/

[16] “Google Colab.” Accessed: May 14, 2025. [Online]. Available: https://colab.research.google.com/

[17] S. M. Nagle, K. A. Stevens, and S. C. Wilbraham, “Wound Assessment,” in *StatPearls*, Treasure Island (FL): StatPearls Publishing, 2025. Accessed: Mar. 17, 2025. [Online]. Available: http://www.ncbi.nlm.nih.gov/books/NBK482198/

[18] S. Bowers and E. Franco, “Chronic Wounds: Evaluation and Management,” *afp*, vol. 101, no. 3, pp. 159–166, Feb. 2020.

[19] H. A. Wallace, B. M. Basehore, and P. M. Zito, “Wound Healing Phases,” in *StatPearls*, Treasure Island (FL): StatPearls Publishing, 2025. Accessed: Mar. 17, 2025. [Online]. Available: http://www.ncbi.nlm.nih.gov/books/NBK470443/

[20] G. I. Broughton, J. E. Janis, and C. E. Attinger, “The Basic Science of Wound Healing,” *Plastic and Reconstructive Surgery*, vol. 117, no. 7S, p. 12S, June 2006, doi: 10.1097/01.prs.0000225430.42531.c2.

[21] B. Guo, R. Dong, Y. Liang, and M. Li, “Haemostatic materials for wound healing applications,” *Nat Rev Chem*, vol. 5, no. 11, pp. 773–791, Nov. 2021, doi: 10.1038/s41570-021-00323-z.

[22] S. Guo and L. A. Dipietro, “Factors affecting wound healing,” *J Dent Res*, vol. 89, no. 3, pp. 219–229, Mar. 2010, doi: 10.1177/0022034509359125.

[23] K. McDermott, M. Fang, A. J. M. Boulton, E. Selvin, and C. W. Hicks, “Etiology, Epidemiology, and Disparities in the Burden of Diabetic Foot Ulcers,” *Diabetes Care*, vol. 46, no. 1, pp. 209–221, Jan. 2023, doi: 10.2337/dci22-0043.

[24] M. R. Zemaitis, J. M. Boll, and M. A. Dreyer, “Peripheral Arterial Disease,” in *StatPearls*, Treasure Island (FL): StatPearls Publishing, 2025. Accessed: Apr. 02, 2025. [Online]. Available: http://www.ncbi.nlm.nih.gov/books/NBK430745/

[25] G. Spentzouris and N. Labropoulos, “The evaluation of lower-extremity ulcers,” *Semin Intervent Radiol*, vol. 26, no. 4, pp. 286–295, Dec. 2009, doi: 10.1055/s-0029-1242204.

[26] B.-J. Jeon, H. J. Choi, J. S. Kang, M. S. Tak, and E. S. Park, “Comparison of five systems of classification of diabetic foot ulcers and predictive factors for amputation,” *Int Wound J*, vol. 14, no. 3, pp. 537–545, June 2017, doi: 10.1111/iwj.12642.

[27] S. O. Oyibo, E. B. Jude, I. Tarawneh, H. C. Nguyen, L. B. Harkless, and A. J. Boulton, “A comparison of two diabetic foot ulcer classification systems: the Wagner and the University of Texas wound classification systems,” *Diabetes Care*, vol. 24, no. 1, pp. 84–88, Jan. 2001, doi: 10.2337/diacare.24.1.84.

[28] L. Yazdanpanah, M. Nasiri, and S. Adarvishi, “Literature review on the management of diabetic foot ulcer,” *World Journal of Diabetes*, vol. 6, no. 1, pp. 37–53, Feb. 2015, doi: 10.4239/wjd.v6.i1.37.

[29] C. Shi *et al.*, “Selection of Appropriate Wound Dressing for Various Wounds,” *Front. Bioeng. Biotechnol.*, vol. 8, Mar. 2020, doi: 10.3389/fbioe.2020.00182.

[30] S. Dhivya, V. V. Padma, and E. Santhini, “Wound dressings – a review,” *BioMedicine*, vol. 5, no. 4, p. 22, Nov. 2015, doi: 10.7603/s40681-015-0022-9.

[31] C. Lindholm and R. Searle, “Wound management for the 21st century: combining effectiveness and efficiency,” *International Wound Journal*, vol. 13, no. S2, pp. 5–15, 2016, doi: 10.1111/iwj.12623.

[32] H. Zhao *et al.*, “Hydrogel dressings for diabetic foot ulcer: A systematic review and meta-analysis,” *Diabetes Obes Metab*, vol. 26, no. 6, pp. 2305–2317, June 2024, doi: 10.1111/dom.15544.

[33] L. Wu, G. Norman, J. C. Dumville, S. O’Meara, and S. E. Bell-Syer, “Dressings for treating foot ulcers in people with diabetes: an overview of systematic reviews,” *Cochrane Database of Systematic Reviews*, no. 7, 2015, doi: 10.1002/14651858.CD010471.pub2.

[34] S. Werner and R. Grose, “Regulation of Wound Healing by Growth Factors and Cytokines,” *Physiological Reviews*, vol. 83, no. 3, pp. 835–870, July 2003, doi: 10.1152/physrev.2003.83.3.835.

[35] S. Yamakawa and K. Hayashida, “Advances in surgical applications of growth factors for wound healing,” *Burns & Trauma*, vol. 7, no. 1, p. 10, Apr. 2019, doi: 10.1186/s41038-019-0148-1.

[36] A. Ullah *et al.*, “Effectiveness of Injected Platelet-Rich Plasma in the Treatment of Diabetic Foot Ulcer Disease,” *Cureus*, vol. 14, no. 8, Aug. 2022, doi: 10.7759/cureus.28292.

[37] T. Hirase, E. Ruff, S. Surani, and I. Ratnani, "Topical application of platelet-rich plasma for diabetic foot ulcers: A systematic review," *World Journal of Diabetes*, vol. 9, no. 10, pp. 172–179, Oct. 2018, doi: 10.4239/wjd.v9.i10.172.

[38] H. OuYang *et al.*, "Platelet-rich plasma for the treatment of diabetic foot ulcer: a systematic review," *Front Endocrinol (Lausanne)*, vol. 14, p. 1256081, Nov. 2023, doi: 10.3389/fendo.2023.1256081.

[39] E. Koivunotko *et al.*, "Cellulase-assisted platelet-rich plasma release from nanofibrillated cellulose hydrogel enhances wound healing," *Journal of Controlled Release*, vol. 368, pp. 397–412, Apr. 2024, doi: 10.1016/j.jconrel.2024.02.041.

[40] R. M. Stoekenbroek, T. B. Santema, D. A. Legemate, D. T. Ubbink, A. van den Brink, and M. J. W. Koelemay, "Hyperbaric Oxygen for the Treatment of Diabetic Foot Ulcers: A Systematic Review," *European Journal of Vascular and Endovascular Surgery*, vol. 47, no. 6, pp. 647–655, June 2014, doi: 10.1016/j.ejvs.2014.03.005.

[41] M. H. Oley *et al.*, "Hyperbaric Oxygen Therapy for Diabetic Foot Ulcers Based on Wagner Grading: A Systematic Review and Meta-analysis," *Plast Reconstr Surg Glob Open*, vol. 12, no. 3, p. e5692, Mar. 2024, doi: 10.1097/GOX.0000000000005692.

[42] Y. Xu *et al.*, "Artificial intelligence: A powerful paradigm for scientific research," *Innovation (Camb)*, vol. 2, no. 4, p. 100179, Nov. 2021, doi: 10.1016/j.xinn.2021.100179.

[43] L. Alzubaidi *et al.*, "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions," *J Big Data*, vol. 8, no. 1, p. 53, 2021, doi: 10.1186/s40537-021-00444-8.

[44] I. H. Sarker, "Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions," *SN COMPUT. SCI.*, vol. 2, no. 6, p. 420, Aug. 2021, doi: 10.1007/s42979-021-00815-1.

[45] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85–117, Jan. 2015, doi: 10.1016/j.neunet.2014.09.003.

[46] M. Pandey *et al.*, "The transformational role of GPU computing and deep learning in drug discovery," *Nat Mach Intell*, vol. 4, no. 3, pp. 211–221, Mar. 2022, doi: 10.1038/s42256-022-00463-x.

[47] A. Esteva *et al.*, "Deep learning-enabled medical computer vision," *npj Digit. Med.*, vol. 4, no. 1, pp. 1–9, Jan. 2021, doi: 10.1038/s41746-020-00376-2.

[48] L. Alzubaidi *et al.*, "A survey on deep learning tools dealing with data scarcity: definitions, challenges, solutions, tips, and applications," *Journal of Big Data*, vol. 10, no. 1, p. 46, Apr. 2023, doi: 10.1186/s40537-023-00727-2.

[49] I. H. Rather, S. Kumar, and A. H. Gandomi, "Breaking the data barrier: a review of deep learning techniques for democratizing AI with small datasets," *Artif Intell Rev*, vol. 57, no. 9, p. 226, Aug. 2024, doi: 10.1007/s10462-024-10859-3.

[50] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.

[51] "keras/keras/src/layers/activations/relu.py at v3.10.0 · keras-team/keras," GitHub. Accessed: May 26, 2025. [Online]. Available: https://github.com/keras-team/keras/blob/v3.10.0/keras/src/layers/activations/relu.py

[52] "keras/keras/src/layers/regularization/dropout.py at v3.10.0 · keras-team/keras," GitHub. Accessed: May 26, 2025. [Online]. Available: https://github.com/keras-team/keras/blob/v3.10.0/keras/src/layers/regularization/dropout.py

[53] A. Khan *et al.*, "A survey of the vision transformers and their CNN-transformer based variants," *Artif Intell Rev*, vol. 56, no. 3, pp. 2917–2970, Dec. 2023, doi: 10.1007/s10462-023-10595-0.

[54] M. Rodrigo, C. Cuevas, and N. García, "Comprehensive comparison between vision transformers and convolutional neural networks for face recognition tasks," *Sci Rep*, vol. 14, no. 1, p. 21392, Sept. 2024, doi: 10.1038/s41598-024-72254-w.

[55] T. Jiao, C. Guo, X. Feng, Y. Chen, and J. Song, "A Comprehensive Survey on Deep Learning Multi-Modal Fusion: Methods, Technologies and Applications," *CMC*, vol. 80, no. 1, pp. 1–35, 2024, doi: 10.32604/cmc.2024.053204.

[56] A. Kline *et al.*, "Multimodal machine learning in precision health: A scoping review," *npj Digit. Med.*, vol. 5, no. 1, pp. 1–14, Nov. 2022, doi: 10.1038/s41746-022-00712-8.

[57] A. Benani *et al.*, "Is Multimodal Better? A Systematic Review of Multimodal versus Unimodal Machine Learning in Clinical Decision-Making," Mar. 13, 2025, *medRxiv*. doi: 10.1101/2025.03.12.25322656.

[58] S.-C. Huang, A. Pareek, R. Zamanian, I. Banerjee, and M. P. Lungren, "Multimodal fusion with deep neural networks for leveraging CT imaging and electronic health record: a case-study in pulmonary embolism detection," *Sci Rep*, vol. 10, no. 1, p. 22147, Dec. 2020, doi: 10.1038/s41598-020-78888-w.

[59] C. Cui *et al.*, "Deep multimodal fusion of image and non-image data in disease diagnosis and prognosis: a review," *Prog. Biomed. Eng.*, vol. 5, no. 2, p. 022001, Apr. 2023, doi: 10.1088/2516-1091/acc2fe.

[60] K. S. Chan and Z. J. Lo, "Wound assessment, imaging and monitoring systems in diabetic foot ulcers: A systematic review," *Int Wound J*, vol. 17, no. 6, pp. 1909–1923, Dec. 2020, doi: 10.1111/iwj.13481.

[61] "Area Determination of Diabetic Foot Ulcer Images Using a Cascaded Two-Stage SVM-Based Classification | IEEE Journals & Magazine | IEEE Xplore." Accessed: May 14, 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/7755785

[62] N. Ohura *et al.*, "Convolutional neural networks for wound detection: the role of artificial intelligence in wound care," *J Wound Care*, vol. 28, no. Sup10, pp. S13–S24, Oct. 2019, doi: 10.12968/jowc.2019.28.Sup10.S13.

[63] M. Goyal, N. D. Reeves, S. Rajbhandari, N. Ahmad, C. Wang, and M. H. Yap, "Recognition of ischaemia and infection in diabetic foot ulcers: Dataset and techniques," *Computers in Biology and Medicine*, vol. 117, p. 103616, Feb. 2020, doi: 10.1016/j.compbiomed.2020.103616.

[64] S. Rambhatla *et al.*, "DL4Burn: Burn Surgical Candidacy Prediction using Multimodal Deep Learning," *AMIA Annu Symp Proc*, vol. 2021, pp. 1039–1048, 2021.

[65] K. A. McLean *et al.*, "Multimodal machine learning to predict surgical site infection with healthcare workload impact assessment," *npj Digit. Med.*, vol. 8, no. 1, pp. 1–10, Feb. 2025, doi: 10.1038/s41746-024-01419-8.

[66] P. Busaranuvong *et al.*, "Explainable, Multi-modal Wound Infection Classification from Images Augmented with Generated Captions," Feb. 27, 2025, *arXiv*: arXiv:2502.20277. doi: 10.48550/arXiv.2502.20277.

[67] R. R. Fletcher *et al.*, "The Use of Mobile Thermal Imaging and Deep Learning for Prediction of Surgical Site Infection," *Annu Int Conf IEEE Eng Med Biol Soc*, vol. 2021, pp. 5059–5062, Nov. 2021, doi: 10.1109/EMBC46164.2021.9630094.

[68] G. Verma, "Leveraging smart image processing techniques for early detection of foot ulcers using a deep learning network," *Pol J Radiol*, vol. 89, pp. 368–377, July 2024, doi: 10.5114/pjr/189412.

[69] M. D. Richard Simman, M. Darren M Gordon, B. Kara Klomparens, and P. T. Frank Aviles Jr, "Use of Multimodal Long-Wave Infrared Thermography Devices in Clinical Practice," *ePlasty*, vol. 23, July 2023, Accessed: Mar. 12, 2025. [Online]. Available: https://www.hmpgloballearningnetwork.com/site/eplasty/case-report/use-multi-modal-long-wave-infrared-thermography-lwit-devices-clinical

[70] K. S. Chan *et al.*, “Clinical validation of an artificial intelligence-enabled wound imaging mobile application in diabetic foot ulcers,” *Int Wound J*, vol. 19, no. 1, pp. 114–124, May 2021, doi: 10.1111/iwj.13603.

[71] D. Ramachandram, J. L. Ramirez-GarciaLuna, R. D. J. Fraser, M. A. Martínez-Jiménez, J. E. Arriaga-Caballero, and J. Allport, “Fully Automated Wound Tissue Segmentation Using Deep Learning on Mobile Devices: Cohort Study,” *JMIR Mhealth Uhealth*, vol. 10, no. 4, p. e36977, Apr. 2022, doi: 10.2196/36977.

[72] B. L. Green, A. Murphy, and E. Robinson, “Accelerating health disparities research with artificial intelligence,” *Front Digit Health*, vol. 6, p. 1330160, 2024, doi: 10.3389/fdgth.2024.1330160.

[73] F. A. Bernardi, D. Alves, N. Crepaldi, D. B. Yamada, V. C. Lima, and R. Rijo, “Data Quality in Health Research: Integrative Literature Review,” *J Med Internet Res*, vol. 25, p. e41446, Oct. 2023, doi: 10.2196/41446.

[74] M. Y. Ng *et al.*, “Perceptions of Data Set Experts on Important Characteristics of Health Data Sets Ready for Machine Learning: A Qualitative Study,” *JAMA Netw Open*, vol. 6, no. 12, p. e2345892, Dec. 2023, doi: 10.1001/jamanetworkopen.2023.45892.

[75] D. Schwabe, K. Becker, M. Seyferth, A. Klaß, and T. Schaeffter, “The METRIC-framework for assessing data quality for trustworthy AI in medicine: a systematic review,” *npj Digit. Med.*, vol. 7, no. 1, pp. 1–30, Aug. 2024, doi: 10.1038/s41746-024-01196-4.

[76] J. Secco, E. Spinazzola, M. Pittarello, E. Ricci, and F. Pareschi, “Clinically validated classification of chronic wounds method with memristor-based cellular neural network,” *Sci Rep*, vol. 14, no. 1, p. 30839, Dec. 2024, doi: 10.1038/s41598-024-81521-9.

[77] M. Kaselimi, E. Protopapadakis, A. Doulamis, and N. Doulamis, “A review of non-invasive sensors and artificial intelligence models for diabetic foot monitoring,” *Front Physiol*, vol. 13, p. 924546, 2022, doi: 10.3389/fphys.2022.924546.

[78] A. Abubakar, H. Ugail, and A. M. Bukar, “Can Machine Learning Be Used to Discriminate Between Burns and Pressure Ulcer?,” in *Intelligent Systems and Applications*, Y. Bi, R. Bhatia, and S. Kapoor, Eds., Cham: Springer International Publishing, 2020, pp. 870–880. doi: 10.1007/978-3-030-29513-4_64.

[79] M. Fridberg *et al.*, “The role of thermography in assessment of wounds. A scoping review,” *Injury*, vol. 55, no. 11, Nov. 2024, doi: 10.1016/j.injury.2024.111833.

[80] J. L. Ramirez-GarciaLuna *et al.*, “Is my wound infected? A study on the use of hyperspectral imaging to assess wound infection,” *Front. Med.*, vol. 10, Aug. 2023, doi: 10.3389/fmed.2023.1165281.

[81] F. Li *et al.*, “Exploration of machine learning models for surgical incision healing assessment based on thermal imaging: A feasibility study,” *International Wound Journal*, vol. 21, no. 2, p. e14677, Feb. 2024, doi: 10.1111/iwj.14677.

[82] F. H. Foomani *et al.*, “Synthesizing time-series wound prognosis factors from electronic medical records using generative adversarial networks,” *Journal of Biomedical Informatics*, vol. 125, p. 103972, Jan. 2022, doi: 10.1016/j.jbi.2021.103972.

[83] V. Cheplygina, M. de Bruijne, and J. P. W. Pluim, “Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis,” *Medical Image Analysis*, vol. 54, pp. 280–296, May 2019, doi: 10.1016/j.media.2019.03.009.

[84] “CIFAR-10 and CIFAR-100 datasets.” Accessed: Apr. 09, 2025. [Online]. Available: https://www.cs.toronto.edu/~kriz/cifar.html

[85] “ImageNet.” Accessed: Apr. 09, 2025. [Online]. Available: https://www.image-net.org/

[86] “diabetic foot ulcer (DFU).” Accessed: Apr. 09, 2025. [Online]. Available: https://www.kaggle.com/datasets/laithjj/diabetic-foot-ulcer-dfu

[87] S. Candemir, X. V. Nguyen, L. R. Folio, and L. M. Prevedello, “Training Strategies for Radiology Deep Learning Models in Data-limited Scenarios,” *Radiol Artif Intell*, vol. 3, no. 6, p. e210014, Nov. 2021, doi: 10.1148/ryai.2021210014.

[88] Z. Hussain, F. Gimenez, D. Yi, and D. Rubin, “Differential Data Augmentation Techniques for Medical Imaging Classification Tasks,” *AMIA Annu Symp Proc*, vol. 2017, pp. 979–984, 2017.

[89] B. Rostami, D. M. Anisuzzaman, C. Wang, S. Gopalakrishnan, J. Niezgoda, and Z. Yu, “Multiclass wound image classification using an ensemble deep CNN-based classifier,” *Computers in Biology and Medicine*, vol. 134, p. 104536, July 2021, doi: 10.1016/j.compbiomed.2021.104536.

[90] *rawpy: RAW image processing for Python, a wrapper for libraw*. Cython, Python. Accessed: May 12, 2025. [MacOS, Microsoft :: Windows, POSIX, Unix]. Available: https://github.com/letmaik/rawpy

[91] *pillow: Python Imaging Library (Fork)*. Python. Accessed: May 12, 2025. [Online]. Available: https://python-pillow.github.io

[92] “Python Release Python 3.11.0,” Python.org. Accessed: May 15, 2025. [Online]. Available: https://www.python.org/downloads/release/python-3110/

[93] “OpenCV: OpenCV modules.” Accessed: May 12, 2025. [Online]. Available: https://docs.opencv.org/4.x/index.html

[94] “os — Miscellaneous operating system interfaces,” Python documentation. Accessed: May 12, 2025. [Online]. Available: https://docs.python.org/3/library/os.html

[95] “shutil — High-level file operations,” Python documentation. Accessed: May 12, 2025. [Online]. Available: https://docs.python.org/3/library/shutil.html

[96] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” Dec. 10, 2015, *arXiv*: arXiv:1512.03385. doi: 10.48550/arXiv.1512.03385.

[97] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” Apr. 10, 2015, *arXiv*: arXiv:1409.1556. doi: 10.48550/arXiv.1409.1556.

[98] M. Tan and Q. V. Le, “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks,” Sept. 11, 2020, *arXiv*: arXiv:1905.11946. doi: 10.48550/arXiv.1905.11946.

[99] “OpenCV: Adding borders to your images.” Accessed: May 12, 2025. [Online]. Available: https://docs.opencv.org/4.x/dc/da3/tutorial_copyMakeBorder.html

[100] E. Yagis *et al.*, “Effect of data leakage in brain MRI classification using 2D convolutional neural networks,” *Sci Rep*, vol. 11, no. 1, p. 22544, Nov. 2021, doi: 10.1038/s41598-021-01681-w.

[101] M. Rosenblatt, L. Tejavibulya, R. Jiang, S. Noble, and D. Scheinost, “The effects of data leakage on connectome-based machine learning models,” *bioRxiv*, p. 2023.06.09.544383, Dec. 2023, doi: 10.1101/2023.06.09.544383.

[102] Z. Liu and H. Rue, “Leave-group-out cross-validation for latent Gaussian models,” Sept. 02, 2024, *arXiv*: arXiv:2210.04482. doi: 10.48550/arXiv.2210.04482.

[103] “LeaveOneGroupOut,” scikit-learn. Accessed: May 13, 2025. [Online]. Available: https://scikit-learn/stable/modules/generated/sklearn.model_selection.LeaveOneGroupOut.html

[104] O. Rainio, J. Teuho, and R. Klén, “Evaluation metrics and statistical tests for machine learning,” *Sci Rep*, vol. 14, no. 1, p. 6086, Mar. 2024, doi: 10.1038/s41598-024-56706-x.

[105] “confusion_matrix,” scikit-learn. Accessed: May 14, 2025. [Online]. Available: https://scikit-learn/stable/modules/generated/sklearn.metrics.confusion_matrix.html

[106] L. Alzubaidi, M. A. Fadhel, S. R. Oleiwi, O. Al-Shamma, and J. Zhang, “DFU_QUTNet: diabetic foot ulcer classification using novel deep convolutional neural network,” *Multimed Tools Appl*, vol. 79, no. 21, pp. 15655–15677, June 2020, doi: 10.1007/s11042-019-07820-w.

[107] M. M. M, M. T. R, V. K. V, and S. Guluwadi, “Enhancing brain tumor detection in MRI images through explainable AI using Grad-CAM with Resnet 50,” *BMC Medical Imaging*, vol. 24, no. 1, p. 107, May 2024, doi: 10.1186/s12880-024-01292-7.

[108] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization,” *Int J Comput Vis*, vol. 128, no. 2, pp. 336–359, Feb. 2020, doi: 10.1007/s11263-019-01228-7.

[109] Y. Zhang, D. Hong, D. McClement, O. Oladosu, G. Pridham, and G. Slaney, “Grad-CAM helps interpret the deep learning models trained to classify multiple sclerosis types using clinical brain magnetic resonance imaging,” *Journal of Neuroscience Methods*, vol. 353, p. 109098, Apr. 2021, doi: 10.1016/j.jneumeth.2021.109098.

[110] M. Abadi *et al.*, “TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems,” Mar. 16, 2016, *arXiv*: arXiv:1603.04467. doi: 10.48550/arXiv.1603.04467.

[111] “Use TPUs | TensorFlow Core,” TensorFlow. Accessed: May 14, 2025. [Online]. Available: https://www.tensorflow.org/guide/tpu

[112] “Module: tf.keras | TensorFlow v2.16.1,” TensorFlow. Accessed: May 14, 2025. [Online]. Available: https://www.tensorflow.org/api_docs/python/tf/keras

[113] C. Barata *et al.*, “A reinforcement learning model for AI-based decision support in skin cancer,” *Nat Med*, vol. 29, no. 8, pp. 1941–1946, Aug. 2023, doi: 10.1038/s41591-023-02475-5.

[114] K. Team, “Keras documentation: Keras 3 API documentation.” Accessed: May 14, 2025. [Online]. Available: https://keras.io/api/

[115] K. Team, “Keras documentation: KerasTuner.” Accessed: May 15, 2025. [Online]. Available: https://keras.io/keras_tuner/

[116] “Tags · keras-team/keras-tuner,” GitHub. Accessed: May 15, 2025. [Online]. Available: https://github.com/keras-team/keras-tuner

[117] S. Saikia *et al.*, “Lesion detection in women breast’s dynamic contrast-enhanced magnetic resonance imaging using deep learning,” *Sci Rep*, vol. 13, no. 1, p. 22555, Dec. 2023, doi: 10.1038/s41598-023-48553-z.

[118] C. R. Harris *et al.*, “Array programming with NumPy,” *Nature*, vol. 585, no. 7825, pp. 357–362, Sept. 2020, doi: 10.1038/s41586-020-2649-2.

[119] “NumPy 2.0.2 Release Notes — NumPy v2.3.dev0 Manual.” Accessed: May 15, 2025. [Online]. Available: https://numpy.org/devdocs/release/2.0.2-notes.html

[120] “What’s new in 2.2.2 (April 10, 2024) — pandas 2.2.3 documentation.” Accessed: May 15, 2025. [Online]. Available: https://pandas.pydata.org/docs/whatsnew/v2.2.2.html

[121] “Release notes — Matplotlib 3.10.3 documentation.” Accessed: May 16, 2025. [Online]. Available: https://matplotlib.org/stable/users/release_notes.html

[122] M. Waskom, “seaborn: statistical data visualization,” *JOSS*, vol. 6, no. 60, p. 3021, Apr. 2021, doi: 10.21105/joss.03021.

[123] Casper da Costa-Luis *et al.*, *tqdm: A fast, Extensible Progress Bar for Python and CLI*. (Nov. 27, 2024). Zenodo. doi: 10.5281/ZENODO.595120.

[124] “scikit-learn: machine learning in Python — scikit-learn 1.6.1 documentation.” Accessed: May 16, 2025. [Online]. Available: https://scikit-learn.org/stable/#

[125] “Releases · tensorflow/tensorflow,” GitHub. Accessed: May 26, 2025. [Online]. Available: https://github.com/tensorflow/tensorflow/releases

[126] “keras/keras/src/layers/convolutional/conv2d.py at v3.9.2 · keras-team/keras,” GitHub. Accessed: May 20, 2025. [Online]. Available: https://github.com/keras-team/keras/blob/v3.9.2/keras/src/layers/convolutional/conv2d.py

[127] "keras/keras/src/layers/pooling/max_pooling2d.py at v3.9.2 · keras-team/keras," GitHub. Accessed: May 20, 2025. [Online]. Available: https://github.com/keras-team/keras/blob/v3.9.2/keras/src/layers/pooling/max_pooling2d.py

[128] "keras/keras/src/layers/reshaping/flatten.py at v3.10.0 · keras-team/keras," GitHub. Accessed: May 26, 2025. [Online]. Available: https://github.com/keras-team/keras/blob/v3.10.0/keras/src/layers/reshaping/flatten.py

[129] "keras/keras/src/layers/core/dense.py at v3.9.2 · keras-team/keras," GitHub. Accessed: May 20, 2025. [Online]. Available: https://github.com/keras-team/keras/blob/v3.9.2/keras/src/layers/core/dense.py

[130] "keras/keras/src/layers/activations/softmax.py at v3.9.2 · keras-team/keras," GitHub. Accessed: May 20, 2025. [Online]. Available: https://github.com/keras-team/keras/blob/v3.9.2/keras/src/layers/activations/softmax.py

[131] "keras-tuner/keras_tuner/engine/hypermodel.py at master · keras-team/keras-tuner." Accessed: May 28, 2025. [Online]. Available: https://github.com/keras-team/keras-tuner/blob/master/keras_tuner/engine/hypermodel.py

[132] "keras/keras/src/optimizers/adam.py at v3.9.2 · keras-team/keras," GitHub. Accessed: May 20, 2025. [Online]. Available: https://github.com/keras-team/keras/blob/v3.9.2/keras/src/optimizers/adam.py

[133] "keras/keras/src/losses/losses.py at v3.9.2 · keras-team/keras," GitHub. Accessed: May 20, 2025. [Online]. Available: https://github.com/keras-team/keras/blob/v3.9.2/keras/src/losses/losses.py

[134] "keras-tuner/keras_tuner/tuners/bayesian.py at master · keras-team/keras-tuner." Accessed: May 28, 2025. [Online]. Available: https://github.com/keras-team/keras-tuner/blob/master/keras_tuner/tuners/bayesian.py

[135] T. J. Bradshaw, Z. Huemann, J. Hu, and A. Rahmim, "A Guide to Cross-Validation for Artificial Intelligence in Medical Imaging," *Radiol Artif Intell*, vol. 5, no. 4, p. e220232, July 2023, doi: 10.1148/ryai.220232.

[136] J. W. Kim, B.-N. Kim, J. I. Kim, C.-M. Yang, and J. Kwon, "Electroencephalogram (EEG) Based Prediction of Attention Deficit Hyperactivity Disorder (ADHD) Using Machine Learning," *Neuropsychiatr Dis Treat*, vol. 21, pp. 271–279, 2025, doi: 10.2147/NDT.S509094.

[137] "keras-tuner/keras_tuner/engine/tuner.py at master · keras-team/keras-tuner." Accessed: May 28, 2025. [Online]. Available: https://github.com/keras-team/keras-tuner/blob/master/keras_tuner/engine/tuner.py

[138] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio, "Transfusion: Understanding Transfer Learning for Medical Imaging," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2019. Accessed: May 30, 2025. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2019/hash/eb1e78328c46506b46a4ac4a1e378b91-Abstract.html

[139] H. Moseley, "In the AI science boom, beware: your results are only as good as your data," *Nature*, Feb. 2024, doi: 10.1038/d41586-024-00306-2.