POLITECNICO DI TORINO

Master's Degree in Biomedical Engineering

Master's Degree Thesis

Methodology for the Diagnosis of Chronic Skin Ulcer Infection through Thermal Imaging and Clinical Trial

Supervisor:
Prof. Jacopo Secco
Co-Supervisors:
Prof. Filippo Molinari
Sara Becchi, M.Sc.

Candidate:
Chiara Franco

Academic Year 2024-2025

Acknowledgments

I am grateful to Jacopo and Sara for their guidance, support, and for always being there to discuss ideas and encourage me throughout this project.

I would like to thank the doctors and nurses at Ospedale Maggiore Niguarda for their collaboration and support.

Abstract

Chronic skin ulcers affect about 2% of the global population and represent a major clinical and economic challenge. Their management involves prolonged treatments, frequent hospital visits, and high costs, while patients often experience pain, reduced mobility, and psychological distress. Infection worsens outcomes, leading to complications such as tissue necrosis, osteomyelitis, necrotizing fasciitis, sepsis, and septic shock, and can transform untreated wounds into life-threatening systemic conditions, emphasizing the need for early detection and intervention.

Although clinical, microbiological, and imaging methods allow pathogen identification and functional assessment, they often have limitations in invasiveness, accessibility, cost, or processing time. In everyday clinical settings, where biopsies or advanced imaging are not always feasible, empirical evaluation frequently guides diagnosis. Thermography emerges as a non-invasive, portable, and cost-effective technique capable of detecting perfusion changes and early inflammatory signs, providing real-time insights into wound conditions. Its ease of use makes it particularly suitable for primary care, outpatient, and home-monitoring contexts.

This thesis combines clinical and computational components developed in collaboration with Niguarda Hospital, Milan. The clinical part involved acquiring RGB and thermal images of chronic wounds under standardized conditions to ensure data quality and reproducibility. Building on this dataset, the research focused on data processing, feature extraction, and the development of classification models for infection assessment. Gradient-based, RGB-derived, and manually annotated parameters, including the Wound Bed Preparation (WBP) scale, were used to train and compare various machine learning and deep learning models, such as CatBoost, XGBoost, Random Forest, SVM, k-NN, logistic regression, and YOLOv11. Model performance was evaluated through balanced accuracy, precision, recall, and F1 score.

Results demonstrate that thermal imaging, combined with AI-based models, enables accurate, rapid, and non-invasive infection assessment in chronic wounds. The proposed methodology supports timely clinical decisions, reduces the need for invasive procedures, and improves patient outcomes, laying the groundwork for integrating non-invasive diagnostics into standard wound care and telemedicine, particularly where traditional methods are limited or delayed.

Table of Contents

Introduction

1.1 Wound Care

Wound care is the branch of medicine focused on the management of skin wounds and ulcers. A wound is defined as a damage or laceration of tissue, typically the skin, caused either by external forces or underlying diseases.

The wound repair process naturally consists of four overlapping phases: hemostasis, inflammation, proliferation, and remodeling [1]. If these phases do not proceed in an orderly and timely manner, the wound may become chronic, leading to increased patient morbidity and imposing a significant economic burden on healthcare systems.

Despite advances in monitoring and treatment, chronic wound management often suffers from fragmented and non-standardized clinical protocols. This lack of standardization results in inconsistent clinical outcomes, prolonged healing times, increased patient discomfort, and a higher workload for healthcare professionals. Effective wound management requires integrated consideration of both local wound characteristics (e.g., tissue composition, exudate) and systemic patient factors (e.g., comorbidities, perfusion, nutritional status).

Epidemiology and Economic Impact on Healthcare Systems

The lifetime incidence of chronic wounds in the global population ranges from 1 % to 2 % [2], varying with age, gender, and geographical location.

In the United States (USA), approximately 6.5 million people are affected by chronic wounds annually, representing about 2 % of the population. The resulting economic burden is conservatively estimated between $28 billion and $31.7 billion per year, considering direct healthcare costs only [3].

In European Union (EU), chronic wounds affect roughly 1 % to 2 % of the population, or about 10 million individuals, with an annual cost to healthcare systems estimated at approximately €40 billion (2 % to 4 % of total healthcare expenditure) [4, 5]. Notably, the primary cost drivers are not dressings or medical devices, but nursing time and hospitalizations. Beyond financial costs, chronic wounds also prolong hospital stays, severely reduce patient quality of life, and increase the risk of severe complications, such

as amputations in the case of diabetic foot ulcers. Data for the United Kingdom (UK) and Australia are shown in figure 1.1.

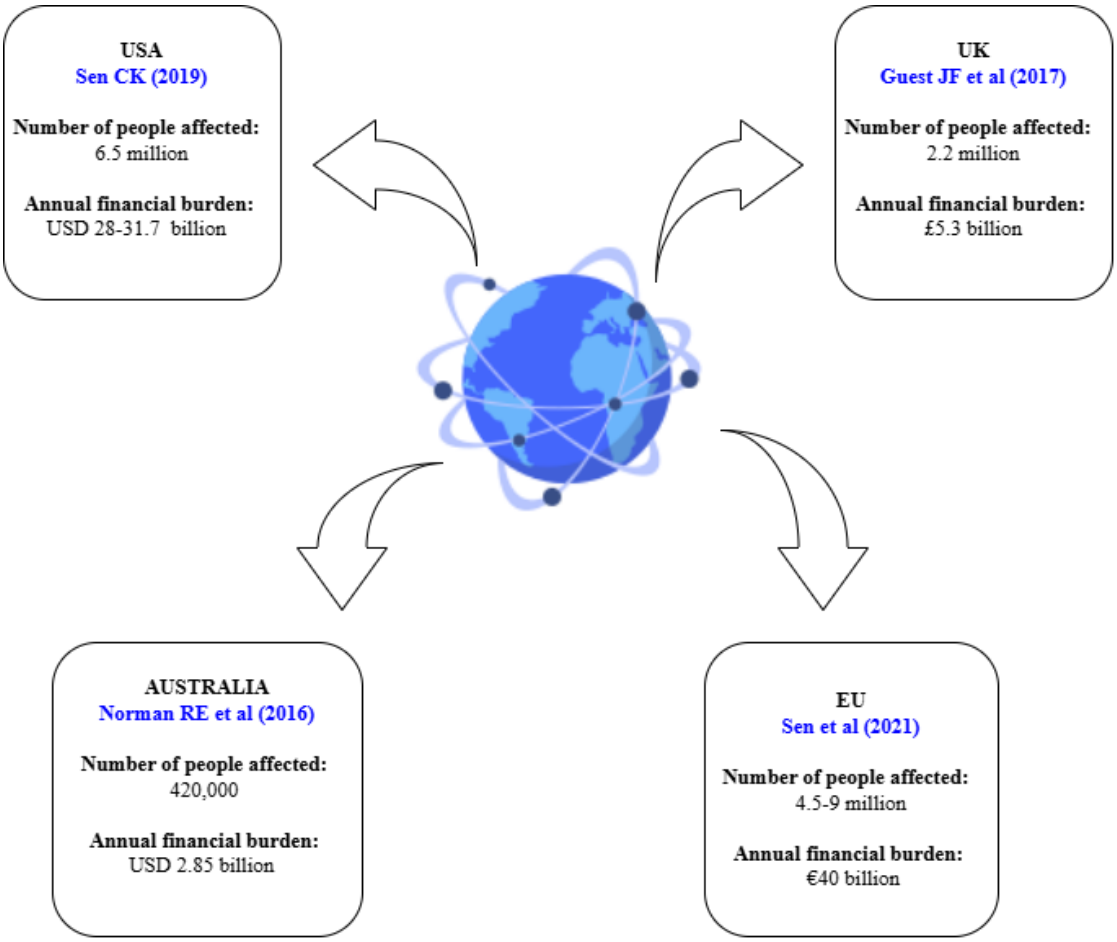


Figure 1.1: Economic and epidemiologic burden of chronic wounds in USA, UK, EU, and Australia [3, 4, 6, 7].

1.1.1 Etiology

Chronic wounds can be classified based on their etiology into venous ulcers, arterial ulcers, pressure ulcers, diabetic foot ulcers, and atypical ulcers [8].

- **Venous ulcers**
  The main cause of venous ulcers is chronic venous disease (CVD), a pathological condition affecting the venous system. CVD afflicts 5% to 8% of the world's population [9] and initially presents with the classic symptoms of venous hypertension,

including edema, hemosiderin staining, and lipodermatosclerosis, eventually leading to the formation of ulcers in the advanced stages [1]. The prevalence of people developing at least one venous ulcer ranges from 0.06% to 2% [10], and they are typically located on the medial supramalleolar aspect of the lower extremity [1]. Due to CVD, which usually remains unresolved, these ulcers frequently become chronic.

- **Arterial ulcers**
  Peripheral arterial disease (PAD), mainly caused by atherosclerosis, affects 14% to 20% of the adult population [11]. This condition leads to the narrowing and obstruction of the arteries, impairing blood flow to the limbs. The resulting poor tissue perfusion prevents the delivery of sufficient oxygen and nutrients to the affected area, ultimately leading to the formation of skin ulcers, which are typically smaller and deeper compared to venous ulcers [8].

- **Diabetic foot ulcers**
  Diabetes mellitus (both type 1 and type 2) prevalence is of 13.9% for women and 14.3% for men [12]. Individuals with diabetes have a diabetic foot ulcers (DFUs) prevalence ranging from 1.2–20% in hospitals and 0.02–10% in the community [12]. Diabetes can lead to several complications, including peripheral neuropathy and peripheral arterial disease (PAD) [13]. DFUs are commonly classified according to the underlying cause: neuropathic ulcers are mainly associated with peripheral neuropathy, ischemic ulcers result from PAD, and neuro-ischemic ulcers arise from a combination of both conditions [13].
  Neuropathic diabetic ulcers typically develop in areas of increased plantar pressure, whereas ischemic ulcers resemble those caused by arterial insufficiency and often present with a necrotic wound bed [8]. Neuro-ischemic diabetic ulcers exhibit features of both types.

- **Pressure ulcers**
  Pressure ulcers (PUs), also known as pressure injuries or bedsores, develop over bony prominences (such as the hips, sacrum, and greater trochanter), where the highest pressure occurs. The combination of pressure, shear, and friction impairs blood flow in both veins and arteries [14], leading to tissue ischemia and, ultimately, ulcer formation. PUs commonly affect individuals with impaired mobility and/or sensation, often due to conditions like spinal cord injuries, sedation, immobilization, or prolonged hospitalization [14]. In hospitals, the prevalence of PUs ranges from 5% to 15% [14], but this significantly increases in intensive care units (ICUs), where the prevalence is estimated to range from 13.1% to 28.7% [15].

- **Atypical ulcers**
  Atypical ulcers constitute approximately 20% of all wounds [16], and are charac-

terized by atypical features in terms of location, etiology, and clinical presentation [8]. They are usually caused by immunological dysregulations and may arise from inflammatory, neoplastic, vasculopathic, hematological, or infectious conditions [17]. These wounds often exhibit atypical clinical characteristics, including an unusual wound bed, borders, and perilesional skin, which directly reflect the underlying etiology of the ulcer.

Some of these wounds are considered comorbidities of other illnesses, such as diabetic foot ulcers (DFUs) in the context of diabetes mellitus. Moreover, diabetes also increases the risk of developing other chronic wounds, including venous ulcers [2].

1.1.2 Wound Appearance and Classification Systems

To analyze and improve a wound's probability of healing, a systematic observation of its appearance is essential. A wound bed typically contains varying amounts of granulation tissue, slough, and necrotic tissue. Additionally, the quantity of exudate and the condition of the perilesional tissue are critical factors in the overall wound assessment.

- **Granulation tissue** appears red, moist, and bumpy due to underlying blood vessel formation (angiogenesis) and the presence of fibroblasts and collagen [18]. Its presence is a positive indicator, as it promotes healing by providing a scaffold for re-epithelialization.
- **Slough** is a white or yellow material, often adherent to the wound bed, composed of fibrin, dead cells, and inflammatory exudate. Persistent slough may lead to infection and delayed healing [18].
- **Necrotic tissue** appears black and thick, typically caused by a lack of blood supply resulting in cell death. Necrotic tissue serves as a nutrient source for bacteria, thereby promoting bacterial colonization and infection [18].

In most cases, the presence of persistent slough or necrotic tissue necessitates their surgical or chemical removal, known as debridement [8], to promote granulation tissue formation.

The amount of exudate produced also plays a crucial role in the healing process. Exudate is a fluid rich in inflammatory cells and biochemical components, continuously produced during the inflammatory response. In chronic wounds, prolonged inflammation leads to sustained vasodilation and vascular permeability, resulting in continuous extracellular fluid formation [19]. While a minimal amount of exudate is necessary for the early stages of healing, excessive or persistent exudate can indicate a sustained inflammatory state that delays closure [20]. Therefore, controlling and reducing exudate levels is essential for promoting tissue regeneration.

WBP score

The Wound Bed Preparation (WBP) score, proposed by Falanga, integrates tissue composition and exudate quantity to estimate the probability of healing [20]. This system provides a standardized method for assessing wound status and predicting outcomes.

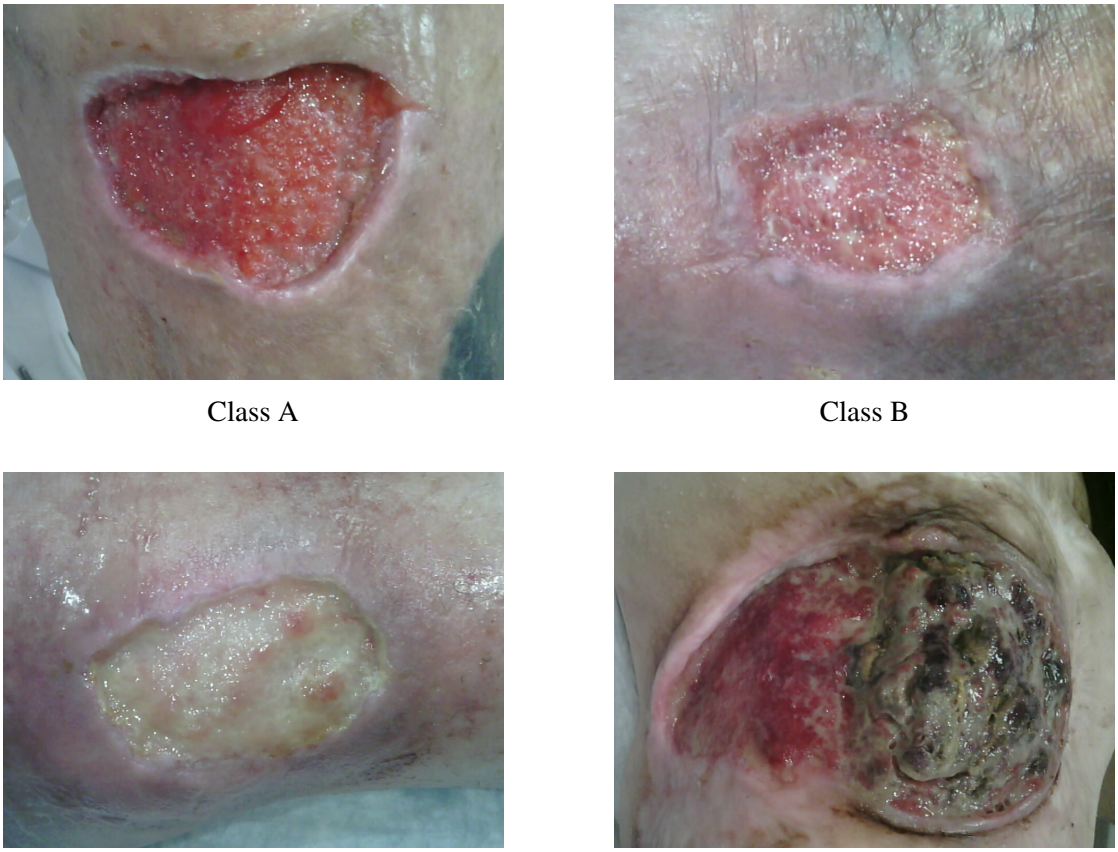
Table 1.1: Wound bed classification based on the proportion of granulation, slough, and eschar

Class	**Granulation**	**Slough**	**Eschar**
A	100%	None	None
B	50–100%	Less than 50%	None
C	Less than 50%	50–100%	None
D	Any amount	Any amount	Present

Table 1.2: Wound exudate score classification

Exudate Score	**Exudate amount**
1	None or minimal
2	Moderate
3	High

Combining the information in Table 1.1 and Table 1.2, it is possible to assign each wound a score that reflects its probability of healing. For example, a wound classified as A1, consisting only of granulation tissue and minimal exudate, has the highest probability of healing. However, a wound with the same tissue composition but a higher amount of exudate, such as A3, would have a lower probability. Overall, the probability of healing decreases progressively from category A to D, and from exudate score 1 to 3. Figure 1.2 shows representative images of granulation tissue for each class, from A to D.



Class A | Class B

Class C | Class D

Figure 1.2: Examples of granulation tissue for classes A–D.

TIME framework

Beyond the WBP score, structured frameworks such as the TIME acronym guide systematic chronic wound management [21]. TIME focuses on four critical components:

- **T (Tissue):** Evaluation of non-viable tissue including necrotic and slough material. As previously discussed, the removal of these tissues via debridement is essential for promoting granulation and healing.
- **I (Infection/Inflammation):** Assessment and control of bacterial colonization, infection, and persistent inflammation, which can impair the healing process if not managed appropriately.
- **M (Moisture):** Maintaining an optimal balance of exudate to ensure a moist wound environment, which facilitates cellular migration and tissue regeneration while preventing maceration.

- **E (Edge):** Evaluation of the wound margins to identify non-advancing or undermined edges, which may indicate stalled healing and require targeted interventions.

The TIME approach encourages a cyclical approach of *assessment, intervention, and reassessment*, allowing clinicians to dynamically adjust treatment strategies according to the wound's evolution.

Texas Classification System

Unlike WBP and TIME, the Texas classification [22] is specific to diabetic foot ulcers (DFUs) and primarily focuses on structural damage and systemic complications, rather than the wound bed contents. It provides a standardized method to grade the wound across two axes:

- **Grade (0–3)**: indicating the depth of the ulcer (Table 1.3).
- **Status (A–D)**: indicating the presence of infection and/or ischemia (Table 1.4).

Each ulcer is classified by a combination of grade and status (e.g., 2B), which is a key predictor of the risk of severe complications, such as amputation.

Table 1.3: Diabetic foot ulcer depth classification according to the University of Texas [22]

Grade	**Description**
0	Pre- or postulcerative site
1	Superficial wound not involving tendon, capsule, or bone
2	Wound penetrating to tendon or capsule
3	Ulcer penetrating to bone or joint

Table 1.4: Diabetic foot ulcer complicating factors classification according to the University of Texas [22]

Status	**Description**
A	Lesions without infection or ischemia
B	Infected lesions without ischemia
C	Ischemic lesions without infection
D	Infected and ischemic lesions

The integration of the WBP score for immediate tissue health and exudate levels, the TIME framework for systematic treatment guidance, and the Texas classification for DFU-specific risk stratification provides a comprehensive, multimodal approach. This combined strategy is essential in modern chronic wound management for accurate prediction of healing outcomes and precise treatment planning.

1.2 Infection

Skin in healthy conditions is colonized by a complex microbiota, whose composition varies depending on local characteristics such as pH, moisture, and skin type [8]. Microorganisms within this microbiota often form biofilms, structured communities embedded in an extracellular matrix, that enhance microbial survival under hostile conditions. In chronic wounds, biofilms are present in the majority of cases, protecting microorganisms from antibiotics and host defenses, and significantly contributing to delayed healing [8].

The progression from colonization to infection hinges not only on microbial factors but also on host characteristics such as immune competence, tissue perfusion, and systemic conditions. Comorbidities like diabetes, vascular insufficiency, or immunosuppression reduce the host's ability to control microbial growth and drastically increase the risk of infection [23].

Chronic wounds often arise from underlying factors that compromise tissue integrity, including impaired blood flow, metabolic disorders, or sustained mechanical pressure [23]. These conditions create an environment favorable to microbial colonization and biofilm formation, which can persist and interfere with the entire healing process [24].

The risk of wound infection can be described through three interacting elements [25]:

- Host resilience: the ability to respond to microbial challenge, affected by age, comorbidities, and immune function.
- Local wound environment: the state of the wound, including removal of devitalized tissue (debridement) and management of exudate.
- Microbial burden: the number, virulence, and interactions of microorganisms, highly amplified by biofilm formation.

Chronic wound infection typically develops as a covert, prolonged condition dominated by biofilm. Biofilm communities resist clearance and trigger sustained local inflammatory responses that damage tissue while persisting, creating a destructive feedback loop that further impairs healing [25].

The International Wound Infection Institute (IWII) proposed a five-stage model to conceptualize infection progression [8]:

1. Contamination: microorganisms are present but not actively proliferating; there is no measurable host response or delay in healing.
2. Colonization: microorganisms begin to multiply, but the host response is minimal; healing is largely unaffected.
3. Local infection: microbial growth triggers a host response, impairing healing. Early signs may be subtle.
4. Spreading infection: microorganisms invade surrounding tissue; effects extend beyond the wound margins.
5. Systemic infection: microorganisms disseminate through the body, provoking a systemic response.

This staged framework aligns with the wound infection continuum [24], which progresses from contamination to colonization, possible critical colonization, and finally overt infection. Critical colonization represents a key transition state in which microbial load and subtle host responses slow healing without obvious clinical signs of infection.

1.2.1 Complications

Infections in chronic wounds can lead to a wide range of severe local and systemic complications if not promptly treated. These complications significantly affect both the healing trajectory of the wound and the overall health of the patient.

- **Localized tissue necrosis:** Due to ischemia and impaired perfusion, infected tissues may progressively die, compromising the structural integrity of the wound bed and surrounding areas.
- **Osteomyelitis:** Infection may spread to the underlying bone, causing osteomyelitis, which can lead to chronic bone infection and prolonged healing [26].
- **Necrotizing fasciitis:** Infections can extend into deeper soft tissues, rapidly destroying fascia and muscle, and representing a life-threatening condition [27].
- **Sepsis:** Microorganisms entering the bloodstream can trigger a systemic inflammatory response, potentially leading to multi-organ dysfunction [28].
- **Septic shock:** In severe cases, sepsis may progress to septic shock, associated with high mortality risk [28].

These complications demonstrate that an untreated infection transforms a chronic wound from a localized healing problem into a serious, potentially life-threatening condition. Early detection and intervention are therefore crucial to prevent both local and systemic consequences.

1.2.2 Infection detection

Accurate detection of wound infection is crucial, as early identification can prevent complications such as delayed healing, chronicity, and increased healthcare costs, as discussed in Section 1.1. Detection relies on a combination of clinical judgement, laboratory testing, and imaging techniques [8].

Clinical and Microbiological Methods

- **Tissue Biopsy:** Biopsy of wound tissue is considered the gold standard for microbiological diagnosis, providing quantitative and qualitative data on pathogens and their susceptibilities. Although highly accurate and able to guide targeted therapy, it is invasive, requires professional expertise, and may cause patient discomfort [8].
- **Swab Cultures:** Superficial specimens can be collected using swabs, such as with the Levine technique. This method is easy, minimally invasive, and inexpensive, but it may be prone to contamination and often does not reflect deeper tissue infection [8].
- **Aspirates and Pus Samples:** Sampling exudate or abscess material allows access to pathogens in deeper tissues. These methods provide more reliable information on causative organisms than swabs, but they are invasive and dependent on wound characteristics [8].
- **Blood Tests and Cultures:** Laboratory tests such as C-reactive protein (CRP), erythrocyte sedimentation rate (ESR), and blood cultures can detect a systemic response. While useful for identifying systemic infection, they are nonspecific for local wound infection and may be influenced by comorbidities [29].
- **Molecular Assays:** Techniques like polymerase chain reaction (PCR) and sequencing can detect unculturable microorganisms. They are highly sensitive and comprehensive but expensive, time-consuming, and unable to distinguish live from dead bacteria [29].
- **Biomarker Assays:** Measurement of biomarkers such as procalcitonin or presepsin reflects the host response. These tests are minimally invasive and rapid but do not specifically indicate local wound infection [29].
- **Biofilm Detection:** Microscopy techniques, like scanning electron microscopy (SEM) and confocal laser scanning microscopy (CLSM), or point-of-care staining methods reveal biofilm presence. These approaches allow direct visualization but are usually laboratory-bound and not routinely available in clinical practice [29].

Imaging-based Methods

- **X-ray:** Radiography can reveal bone involvement in chronic infections. It is inexpensive and widely available, but it has low sensitivity in detecting early osteomyelitis and exposes the patient to ionizing radiation [8].
- **CT:** Computed tomography provides rapid, high-resolution anatomical imaging, useful for identifying abscesses or bone destruction. However, it involves higher doses of ionizing radiation compared to standard radiography and offers limited soft tissue contrast [29].
- **MRI:** Magnetic resonance imaging is sensitive for soft tissue and bone infection, offering excellent contrast and depth. However, it is costly, less accessible, and contraindicated in some patients [8].
- **Ultrasound:** Portable ultrasound identifies fluid collections and can guide procedures. It is non-ionizing, affordable, and bedside-compatible, but results are highly operator-dependent and limited for deep tissues [30].
- **Nuclear Medicine:** PET and scintigraphy highlight metabolic activity. These techniques are highly sensitive for active infection but are expensive and require radiotracers [29].
- **SFDI (Spatial Frequency Domain Imaging):** Optical imaging allows quantitative assessment of tissue chromophore distribution. It is non-contact and quantitative but experimental and limited to research [29].
- **Fluorescence Imaging:** This method visualizes bacterial porphyrins to guide debridement. It provides immediate bedside feedback but mainly detects surface bacteria and may miss species without fluorescence [30].
- **Thermography:** Long-wave infrared thermography detects local perfusion changes and inflammation-related heat [31]. It is non-contact, non-ionizing, inexpensive, and easy to use in bedside or home care settings [32]. Absolute temperature readings can be influenced by environmental conditions, so standardized protocols and comparison with control areas are recommended [31].

Despite the availability of numerous clinical, microbiological, and advanced imaging-based methods, each approach presents inherent limitations in terms of invasiveness, accessibility, cost, or specificity. Consequently, in routine clinical practice, especially when immediate biopsies or advanced imaging are not feasible, clinicians often rely on standardized empirical assessment to guide diagnostic decisions.

Cutting & Harding scale

Among the most widely recognized frameworks, designed to structure clinical judgment, the Cutting & Harding scale [33] (C&H) provides a structured set of observable criteria for identifying wound infection based on expert consensus. Developed through the Delphi method, the C&H scale defines key clinical indicators applicable across different wound types. The main criteria include:

- **Cellulitis:** diffuse redness and swelling of surrounding tissues.
- **Exudate changes:** increase in volume, abnormal viscosity, or purulence.
- **Delayed healing:** slower-than-expected wound closure despite standard care.
- **Discoloration or necrotic changes:** unusual pigmentation or dark tissue areas.
- **Friable granulation tissue:** easily bleeding wound bed tissue.
- **Malodour:** strong, abnormal odor from the wound.
- **Unexpected pain or tenderness:** especially in previously painless wounds.
- **Wound breakdown:** deterioration in size or structure of the wound bed.

Although the criteria are fundamentally based on clinical observation, the Cutting & Harding scale introduces a necessary degree of standardization that helps reduce subjective variability and enhances the reproducibility of wound infection assessment. By translating empirical signs into structured diagnostic parameters, it facilitates consistent evaluation across practitioners and settings, thereby supporting timely and uniform clinical decision-making.

Overall, clinical and microbiological techniques provide definitive microbial identification but are often invasive or slow, whereas imaging methods allow noninvasive functional and structural assessment. Thermography, in particular, stands out as a practical, low-cost adjunct for monitoring perfusion and early signs of infection, making it highly suitable for primary care and home settings [31, 32].

1.2.3 Thermography

Thermography is based on the detection of infrared (IR) radiation naturally emitted by all objects at temperatures above absolute zero. The IR spectrum extends approximately from 0.75 to 1000 µm and is commonly divided into Near Infrared (NIR, 0.75 µm–1.4 µm), Mid Infrared (MIR, 0.75 µm–8 µm), and Far Infrared (FIR, 8 µm–1000 µm), as represented in Figure 1.3. Clinical thermography typically relies on the long-wave infrared (LWIR) band (8 µm–14 µm), which corresponds to the emission

peak of human skin at physiological temperatures (~ 33 °C–36 °C), making it ideal for non-invasive temperature measurements in biomedical applications, including wound assessment.

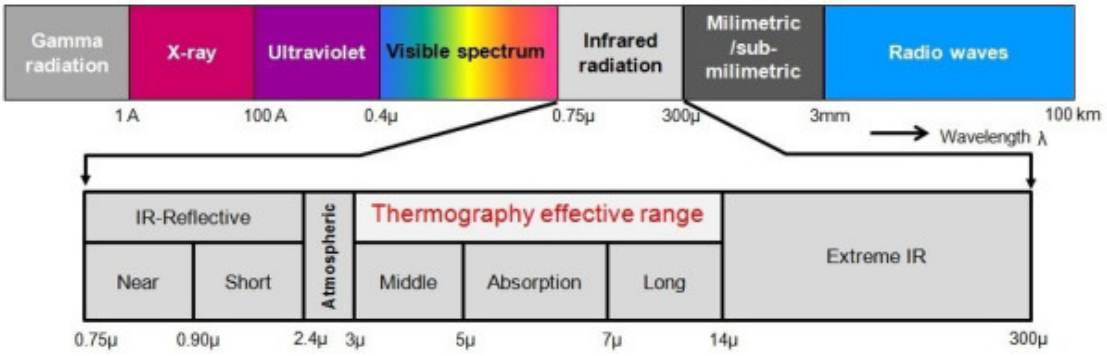


Figure 1.3: Infrared spectrum divided into Near-, Mid-, and Far-Infrared bands. [34].

The thermal radiation emitted by a surface can be described by the Stefan–Boltzmann law (1.1):

$$E = \varepsilon \sigma T^4 \tag{1.1}$$

where E is the emitted radiative power per unit area, ε is the emissivity of the surface, σ is the Stefan–Boltzmann constant, and T is the absolute temperature of the surface.

The peak wavelength of emission $\lambda_{\max}$ can be estimated using Wien's displacement law(1.2):

$$\lambda_{\max} = \frac{b}{T} \tag{1.2}$$

where b is Wien's constant. For human skin, this peak falls within the LWIR range.

In biomedical contexts, the relationship between tissue temperature, blood perfusion, and metabolic activity can be described using the Pennes bio-heat equation (1.3):

$$\rho c \frac{\partial T}{\partial t} = k \nabla^2 T + \rho_b c_b \omega_b (T_a - T) + Q_m \tag{1.3}$$

where ρ and c are the tissue density and specific heat, k is the thermal conductivity, ρ_b, c_b, and ω_b represent blood density, specific heat, and perfusion rate respectively, T_a is the arterial blood temperature, and Q_m represents metabolic heat generation.

Infrared cameras consist of arrays of IR sensors (focal plane arrays), which convert the incoming radiation into electrical signals. Each sensor element corresponds to a pixel in the thermal image, encoding the local surface temperature. The digital signals are then processed and mapped onto color scales (colormaps), producing images where warmer and cooler regions can be easily visualized. This capability enables the non-invasive detection of perfusion abnormalities, inflammation, or necrotic tissue.

Previous studies have already explored the potential of thermal imaging in wound and infection assessment. For example, Liu et al. employed thermal images acquired with a smartphone-based device to predict surgical site infections using convolutional neural networks; in this case, the CNNs were trained not on the visual thermal images themselves, but on the corresponding temperature matrices, achieving high sensitivity and specificity [35]. Other investigations, such as the study by Collins et al., focused on the temperature difference between the wound bed and surrounding healthy skin, reporting that differences of 2–3 °C can indicate infection, although with limited specificity [36].

These contributions demonstrate both the promise and the challenges of thermal imaging for wound assessment.

1.3 Aim of the thesis

As discussed in Section 1.1, chronic ulcers represent an increasingly critical problem in modern healthcare. The aging population, together with the rising incidence of diabetes and other comorbidities, contributes to their growing frequency and impact. This trend translates directly into a higher clinical and economic burden for both patients and the healthcare system.

Consequently, the early identification of infection is crucial to prevent the severe complications described in Section 1.2.1. Among the available diagnostic methods, infrared thermography stands out as a non-invasive, portable, and cost-effective tool.

Therefore, the aim of this thesis is to evaluate the efficacy of infrared thermography in estimating the probability of infection in chronic wounds. By facilitating an earlier diagnosis, this approach could help to initiate timely and appropriate treatment, reduce patient discomfort, improve quality of life, and mitigate the socioeconomic burden of chronic wounds for both individual patients and the healthcare system as a whole.

Materials and Methods

2.1 Data Collection

2.1.1 Instrumentation

In this study, data acquisition was performed using the *Wound Viewer Lite01* (WV Lite01), a compact and portable version of the classical Wound Viewer designed for use with any tablet or smartphone. The device connects via a USB-C cable and operates through a dedicated mobile application, providing the same quantitative assessment capabilities as the standard Wound Viewer [37]. Figure 2.5 shows the classical Wound Viewer device. The WV Lite01 retains essentially the same hardware components and software functionalities as the classical device, while being much more compact and portable. Its small size makes it easier to transport and use in diverse settings, such as home visits or community healthcare, and it additionally integrates a micro thermal camera for infrared imaging.

Hardware architecture

The Wound Viewer integrates a five-megapixel color CMOS camera, sixteen infrared (IR) distance sensors, and four white light-emitting diodes (LEDs) arranged symmetrically around the optical axis (Figure 2.5). The LEDs provide uniform and calibrated illumination, reducing the effects of ambient light and minimizing shadow artifacts during image acquisition. The IR sensors continuously measure the distance between the device and the skin surface at multiple points. These measurements are used to calibrate the focal distance of the camera and to estimate local depth variations, allowing spatial scaling of each image pixel into metric units. The combination of RGB imaging and IR depth sensing enables simultaneous capture of the wound's color, texture, and relative topography, which are essential for an accurate morphological characterization. In addition to these components, the WV Lite01 integrates a micro thermal camera, the FLIR Lepton® 2.5 [38]. This thermal sensor allows acquisition of infrared images of the wound, which are captured asynchronously after the RGB image and the input of relevant wound metadata via the mobile app. The FLIR Lepton® provides low-resolution

thermal maps suitable for relative temperature analysis and gradient extraction across the wound surface. Figure 2.4 illustrates the compact size of the Lite01 device, highlighting the difference in dimensions compared to the original model while retaining the same functionality. Figure 2.5 shows the Wound Viewer Lite01 device, where all hardware components are clearly visible.

Figure 2.4: WoundViewer Lite01 compared to classical WoundViewer device

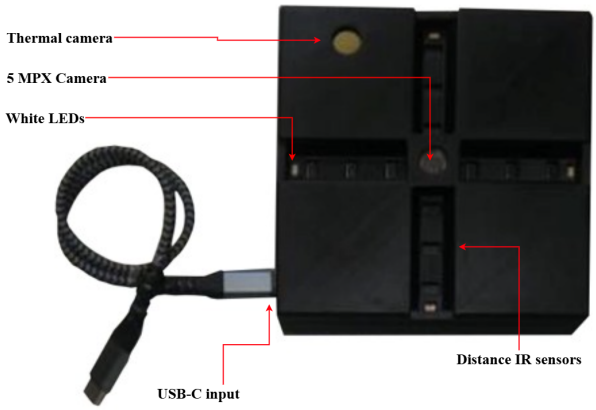


Figure 2.5: WoundViewer Lite01 device with its components [39].

Acquisition workflow

The image acquisition process follows a standardized, operator-independent protocol designed to ensure reproducible imaging conditions across all subjects.

First, the operator positions the device perpendicular to the wound and captures the RGB image using the dedicated mobile application, maintaining a stable distance to minimize geometric distortion. Next, the relevant wound metadata are entered into the app. Finally, the operator manually selects the *Thermal Image* option, located at the bottom of the metadata screen, and triggers the FLIR Lepton® 2.5 sensor to acquire the thermal image asynchronously relative to the RGB acquisition.

Figure 2.6 illustrates the complete acquisition workflow implemented in the mobile application. The diagram summarizes the sequential steps required for standardized wound documentation, from patient registration to RGB and thermal image acquisition. This represents the most comprehensive case, in which a new patient is being added to the system; in other scenarios, only a new wound or a follow-up visit for an already registered patient may be recorded. The final yellow block represents the integration of the thermal image analysis module, which constitutes the novel component developed within the framework of this thesis and will be added to extend the current system capabilities.

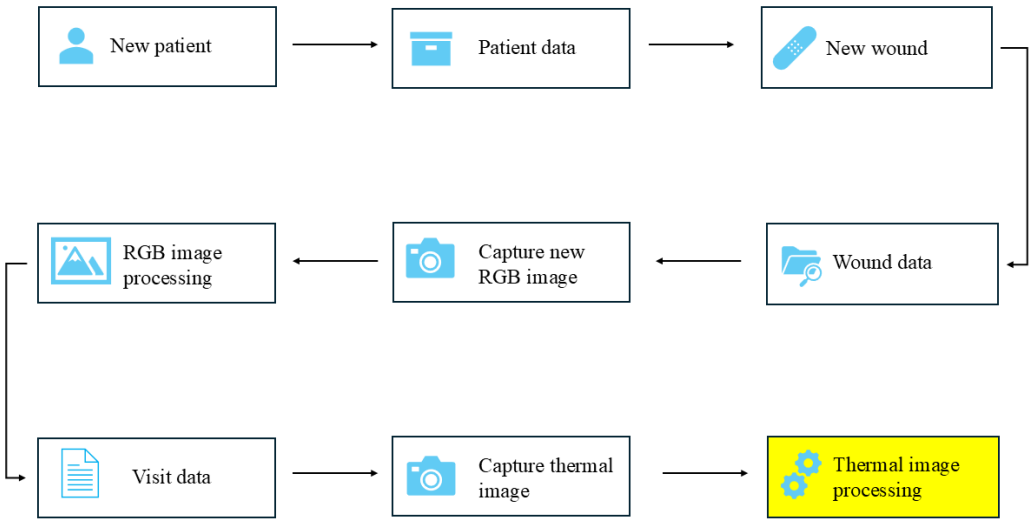


Figure 2.6: Workflow of the acquisition process in the Wound Viewer Lite01 app, from patient registration to thermal image capture. The yellow block indicates the thermal analysis module developed in this thesis.

Data processing and AI algorithm

The acquired RGB images, IR distance data, and thermal images are processed by the same proprietary AI algorithms as the classical device [39]. A convolutional neural network (CNN) first identifies the Region of Interest (ROI) corresponding to the wound and generates a preliminary segmentation mask. A discrete-time cellular neural network (DT-CNN) then refines this mask and assigns each pixel to tissue types such as granulation, slough, or necrosis. From these segmented maps, the system automatically computes quantitative metrics including wound area, perimeter, and relative depth, calibrated with infrared measurements. Finally, the algorithm provides a classification according to the clinically validated four-level Wound Bed Preparation (WBP) scale [40]. Figure 2.7 shows an example of the processing output.

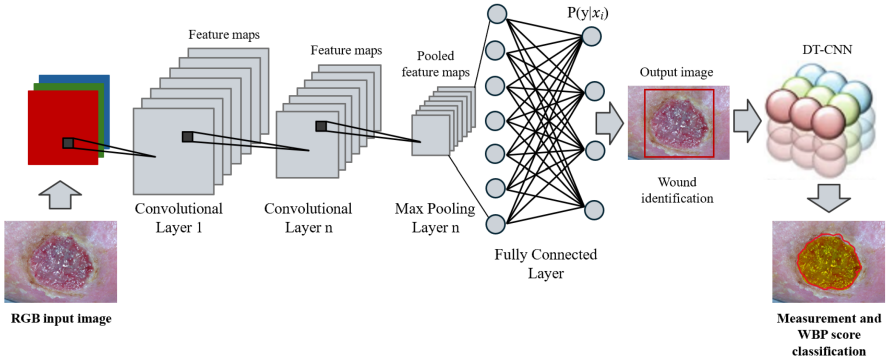


Figure 2.7: Wound Viewer output showing automatic segmentation and tissue classification [41].

Data storage and interoperability

All processed data, including raw images, segmentation masks, and computed parameters, are automatically encrypted and stored in a GDPR-compliant cloud database. Authorized clinicians can securely access the data for longitudinal wound monitoring and teleconsultation. Each measurement is timestamped and associated with patient metadata, allowing trend visualization of wound healing over time. The cloud infrastructure supports integration with electronic health record (EHR) systems through standardized communication protocols.

Rationale for device selection

The Wound Viewer Lite01 was chosen for this study due to its compactness, portability, and ability to provide accurate, reproducible, and operator-independent wound assess-

ments. Its integration of calibrated optical and infrared sensors with the embedded AI segmentation pipeline minimizes subjective interpretation and enhances measurement consistency. The addition of the FLIR Lepton® 2.5 thermal sensor allows relative temperature analysis without affecting the standard RGB acquisition and AI workflow. Overall, its combination of standardized illumination, calibrated depth sensing, and neuromorphic AI analysis ensures objective and repeatable wound characterization, aligning with the methodological needs of this research.

2.1.2 Data Acquisition

Study Site

All patient data were collected at the ASST Grande Ospedale Metropolitano Niguarda, Milan, Italy. The study was conducted as a prospective, observational, monocentric trial. Randomization was not applicable, as comparisons were made against parameters instrumentally measured on the same lesions.

Population and Sample

The study enrolled patients with cutaneous ulcers, divided into six groups according to ulcer type. The target number of lesions per group was planned as follows:

- Lower limb ulcers: minimum 30 infected + 15 non-infected
- Diabetic foot lesions: minimum 30 infected + 15 non-infected
- Pressure ulcers: minimum 30 infected + 15 non-infected
- Acute ulcers: minimum 30 infected + 15 non-infected
- Ulcers from autoimmune diseases: minimum 30 infected + 15 non-infected
- Post-surgical sternal wounds (cardiac, general, vascular surgery): minimum 30 infected + 15 non-infected

In practice, the final distribution of infected lesions was naturally unbalanced, reflecting the real-world incidence of infections in the clinical population. During the training of machine learning and deep learning models, specific techniques for class balancing (such as oversampling or weighting) will be applied to mitigate this inherent imbalance. The final number of patients and lesions per category is summarized in Table 2.5.

Inclusion and Exclusion Criteria

Patients were eligible for inclusion if they met all of the following criteria:

- Age > 18 years
- Receiving standard care
- Provided informed consent
- Lesion size between 2 cm^2 and 100 cm^2

Patients were excluded if any of the following applied:

- Failure to provide informed consent
- Lesion size < 2 cm^2
- Lesion size > 100 cm^2

Data Collection Methods

Data were collected for each patient and lesion during three scheduled visits:

- Non-sternal wounds: Days 1 (T0), 7 (T1), 16 (T2)
- Sternal wounds: Days 1 (T0), 4 (T1), 8 (T2) post-operation

These time points represented the optimal schedule defined by the protocol, with the actual visit dates adjusted according to each patient's individual availability.

All wound parameters were recorded using the Wound Viewer Lite01 system, described in Section 2.1.1.

Collected data included:

- Wound area (cm^2), depth (mm), volume (cm^3)
- Perceived pain (scale 1–10)
- Infection grade (Cutting & Harding scale)
- WBP score
- TEXAS classifications (for diabetic foot)
- Dressing and treatment details

- RGB images and peri-lesional images

Ground Truth for Infection: The ideal confirmation of infection is through biopsy; however, due to regional healthcare policies in Lombardy, routine biopsies were not feasible for all patients. Therefore, infections were confirmed by biopsy only in doubtful cases or when exemption allowed. In the other cases, the presence of infection was determined using the validated Cutting & Harding scale, as described in Section 1.2.2. All collected data were pseudo-anonymized according to the General Data Protection Regulation (EU 2016/679) and stored securely in the Wound Viewer system cloud.

Dataset
Overall, the clinical trial included a total of X patients, from whom Y wound images were collected. Table 2.5 summarizes the breakdown of wound categories, indicating for each type the number of patients and the corresponding number of infected and non-infected images.

Table 2.5: Summary of wound categories included in the clinical dataset, indicating for each ulcer type the number of patients and the corresponding number of infected and non-infected wound images.

Ulcer type		Patients	Ulcer images	Infected ulcers	Non-infected ulcers
Lower limb	Venous	38	208	86	122
	Arterious	6	21	5	16
	Mixed	3	7	4	3
Diabetic		5	10	3	7
Pressure		17	49	6	43
Acute		19	50	12	38
Autoimmune		1	6	0	6
Other		19	61	12	49
Total		108	412	128	284
Total %				31,07	68,93

2.2 Artificial Intelligence

Artificial Intelligence (AI) is a multidisciplinary field of computer science that aims to develop systems performing tasks that traditionally require human intelligence, such as perception, reasoning, learning, and decision-making. To achieve these results, AI uses computational models to mimic how the human mind works. So, machines start to understand tricky data, pick up new things, and figure out problems by themselves.

From its inception in the mid-20th century, AI has been deeply inspired by the structure and functioning of the human brain. Early researchers tried to simulate human thought and learning processes, taking cues from psychology and neuroscience. That's where artificial neural networks came from: the researchers built these basic models, like a stripped-down version of brain cells talking to each other [42]. Nowadays, those early networks may look pretty simple; however, they set the stage for everything that came after.

The historical development of Artificial Intelligence (AI) is generally delineated into distinct major phases, the first one being Symbolic, or Rule-Based, AI [43]. This initial paradigm was based on the fundamental principle that intelligent behavior could be engineered by human experts who manually encoded explicit logical rules and extensive knowledge bases into the machines. While this methodology proved adequate for solving well-defined and logically contained problems, its severe limitations became rapidly apparent. Specifically, the system's inherent rigidity rendered it incapable of operating effectively in the face of real-world phenomena characterized by uncertainty, ambiguity, and complexity that could not be fully captured by predefined rules.

The second approach is the dominant methodology in contemporary Artificial Intelligence and is the data-driven approach. Rather than relying on explicitly programmed rules, these advanced systems derive intelligence by analyzing vast quantities of empirical examples [44]. They operate by autonomously processing massive datasets to discern intricate underlying patterns and establish critical relationships, thus obviating the need for developers to manually specify every operational parameter. This fundamental paradigm shift gave rise to the field of Machine Learning (ML) [44], which empowers models to improve their performance and knowledge iteratively through direct interaction with data. Subsequently, this foundation was advanced by Deep Learning (DL), characterized by the utilization of multi-layered neural networks capable of modeling and handling highly complex, hierarchical features. Consequently, AI has transcended purely niche applications and has become a fundamental force driving progress across scientific and medical disciplines.

Distinguishing Machine Learning from Deep Learning

The hierarchical relationship among these three domains—AI, ML, and DL—is schematically represented in Figure 2.8.

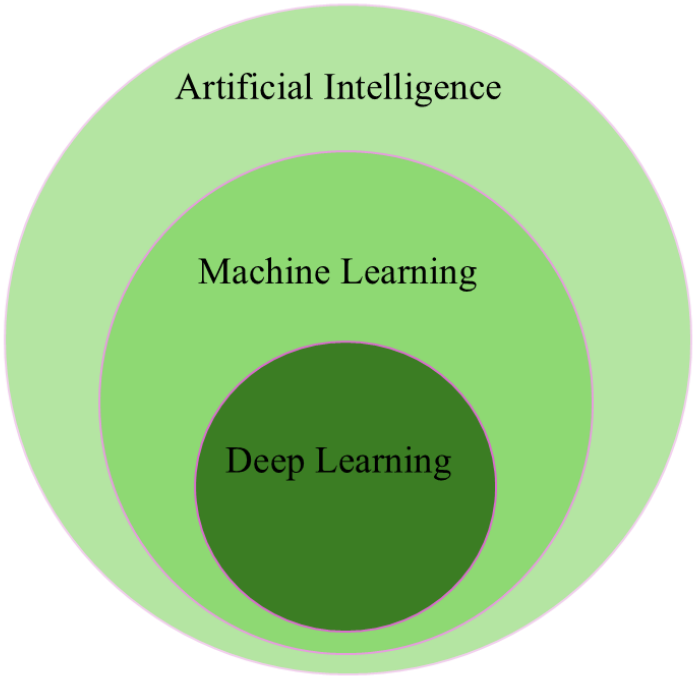


Figure 2.8: Hierarchical relationship among Artificial Intelligence (AI), Machine Learning (ML), and Deep Learning (DL).

While both ML and DL share the objective of enabling machines to learn from data, they differ significantly in the methodological approach to feature extraction.

In traditional ML, the definition of relevant features, such as color moments, texture descriptors, or shape statistics in image analysis, requires manual design guided by domain expertise [45]. This process is time-consuming and relies heavily on human knowledge. In contrast, DL architectures (specifically, Deep Neural Networks) possess the capacity to automatically learn and extract these features directly from the raw data. They build progressively abstract representations of information through multiple computational layers, capturing complex spatial or semantic patterns autonomously [45]. This fundamental distinction is illustrated in Figure 2.9, explaining why DL models often achieve superior accuracy in complex perceptual tasks when sufficient annotated data are available.

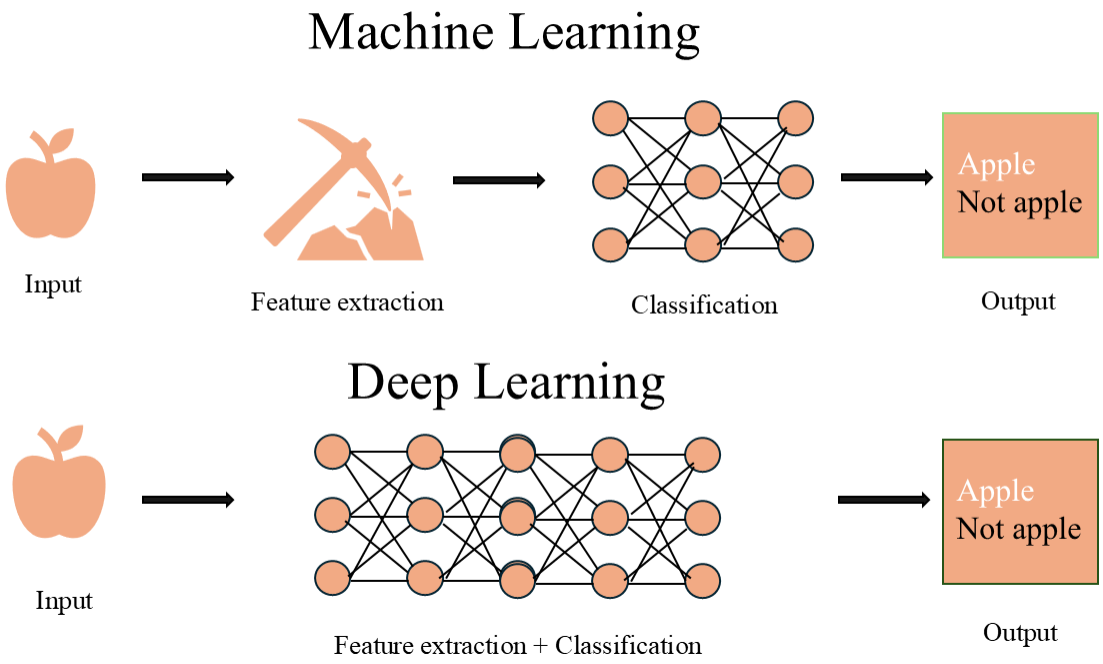


Figure 2.9: Conceptual difference between traditional Machine Learning (manual feature extraction requiring expertise) and Deep Learning (automatic feature learning through abstraction layers).

Another critical difference lies in their performance dependency on data volume. ML algorithms typically perform well with relatively small datasets, where careful feature engineering can partially compensate for data limitations [45]. DL models, instead, require large-scale datasets to effectively train their numerous parameters and generalize successfully. As depicted in Figure 2.10, the performance of ML models tends to plateau sooner, whereas DL performance continues to improve significantly as more data becomes available, fully exploiting its superior representational power for high-level abstraction and precision.

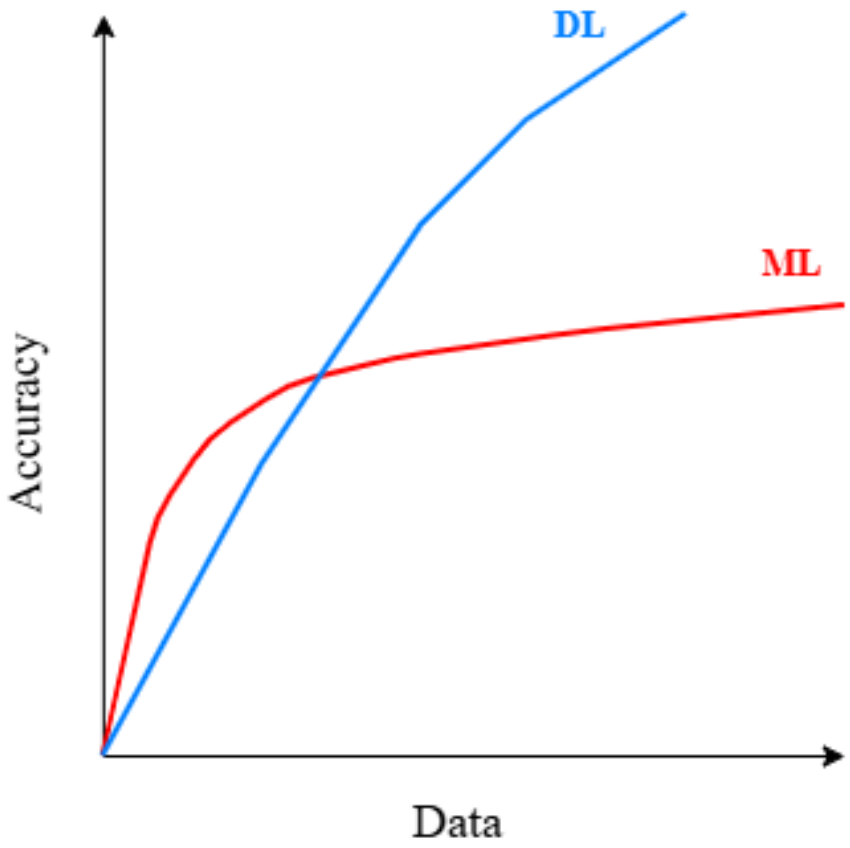


Figure 2.10: Illustrative comparison of data dependency between Machine Learning and Deep Learning models. ML reaches good performance with smaller datasets, while DL requires large data volumes to fully exploit its representational power.

In the following sections, these two paradigms will be explored in greater detail, focusing on their specific methodological principles and their application to wound image analysis within the framework of this thesis.

2.2.1 Machine learning

Machine Learning (ML) represents one of the most mature and impactful areas of Artificial Intelligence. It focuses on the development of models capable of learning patterns and relationships from data without being explicitly programmed. Instead of relying on rigid rules, ML systems "learn" generalizable knowledge from examples, allowing them to make predictions or classifications on new, unseen data.

Traditional ML approaches typically rely on a preliminary feature extraction phase, where meaningful descriptors are manually selected from raw data based on domain

knowledge. These features translate complex information into numerical variables that can be analyzed by learning algorithms. Through this process, ML models establish correlations and decision boundaries that can later be interpreted and validated by human experts.

The selection of Machine Learning (ML) techniques was primarily driven by two critical considerations. Firstly, ML algorithms are well-suited for scenarios involving modest dataset sizes, demonstrating efficient performance provided that the chosen features are relevant and robust. This approach offers superior resource efficiency compared to methodologies that necessitate vast quantities of data.

However, the paramount factor influencing this decision was the need for transparency and interpretability. In clinical and medical applications, a system cannot merely produce diagnostic outcomes; the underlying rationale must be explicitly comprehensible. Clinicians require insight into the logical pathway, not just the final result. ML models facilitate the articulation of connections between the input features and concrete medical concepts, thereby allowing healthcare professionals to trace the reasoning. This inherent clarity is essential for fostering trust and adoption of the analytical tools within the medical domain.

This aspect aligns with the growing emphasis on Explainable Artificial Intelligence (XAI), which aims to make algorithmic decisions understandable to human users. However, as highlighted by Holzinger et al. [46], explainability alone is insufficient in clinical applications. The concept should evolve toward causability, defined as the degree to which an explanation enables an expert to achieve causal understanding. In this sense, causability represents a human-centered property, ensuring that AI-driven insights are not only interpretable but also aligned with causal reasoning and medical logic.

Ultimately, Machine Learning offers an ideal balance between predictive capability, interpretability, and practical applicability. Its ability to provide transparent, traceable, and explainable results makes it particularly suited for healthcare environments, where algorithmic decisions must complement but not replace human expertise.

Given the characteristics of the dataset (its limited size, the structured nature of the extracted features, and the need for model interpretability in a clinical context), the following machine learning algorithms were selected for the classification of thermal wound images: Logistic Regression, K-NN, SVM, Random Forest, XGBoost, and CatBoost. These models offer a balance between classification performance and explainability, ensuring that the extracted features can be meaningfully related to clinically relevant properties. In the following sections, are detailed the operating principles of these models and the rationale for their inclusion in the analysis, focusing on how they address constraints on data size and clinical interpretability.

Logistic Regression

Logistic Regression is a widely adopted statistical and machine learning model primarily employed for binary classification tasks [47]. Unlike linear regression, which predicts a continuous outcome, Logistic Regression models the probability of a sample belonging to a specific class C (e.g., $P(y = 1|\mathbf{x})$) using the logistic function, also known as the sigmoid function (σ).

The logistic regression model estimates the probability P of belonging to the positive class by applying a linear combination of the input features $\mathbf{x}$ to the log-odds (or logit) transformation:

$$\log\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n = \mathbf{x}^T \boldsymbol{\beta}$$

where $\mathbf{x} = [x_1, x_2, \ldots, x_n]^T$ is the feature vector and $\boldsymbol{\beta} = [\beta_0, \beta_1, \ldots, \beta_n]^T$ are the model coefficients. The logit function maps probabilities from the interval $(0, 1)$ to the entire real line, allowing the model to represent linear relationships in this transformed space.

The probability P is then computed by inverting the logit function using the sigmoid function:

$$P(y = 1 \mid \mathbf{x}) = \sigma(\mathbf{x}^T \boldsymbol{\beta}) = \frac{1}{1 + e^{-\mathbf{x}^T \beta}} \tag{2.4}$$

In the general formulation of the binary logistic regression model:

$$P(y = 1 \mid \mathbf{x}) = \sigma(\mathbf{w} \cdot \mathbf{x} + b) \tag{2.5}$$

$$P(y = 0 \mid \mathbf{x}) = 1 - \sigma(\mathbf{w} \cdot \mathbf{x} + b) \tag{2.6}$$

$$\sigma(z) = \frac{1}{1 + e^{-z}} = \frac{1}{1 + \exp(-z)} \tag{2.7}$$

$$z = \left(\sum_{i=1}^{n} w_i x_i\right) + b = \mathbf{w} \cdot \mathbf{x} + b \tag{2.8}$$

The model determines the optimal coefficient vector $\boldsymbol{\beta}$ by maximizing the likelihood function over the training data, a process typically performed through iterative optimization algorithms such as gradient descent. The working principle of the sigmoid function, which transforms a linear output into a probability in the range $(0, 1)$, is illustrated in Figure 2.11.

Logistic Regression assumes a linear relationship between the input features and the log-odds of the outcome. This assumption contributes to its high interpretability and computational efficiency. It is robust, straightforward to implement, and serves as an essential benchmark for evaluating the performance ceiling of classification tasks.

This model was specifically chosen due to its interpretability, its ability to provide stable baseline performance, and its effectiveness in classification tasks. Furthermore, the magnitude and sign of the model's coefficients $\boldsymbol{\beta}$ offer direct insights into feature importance, which is valuable for understanding the factors driving the predictive outcomes.

A summary of the main characteristics and a comparison of all the machine learning models discussed in this study is provided in Table 2.6, highlighting their respective strengths, limitations, and specific applicability to the dataset at hand.

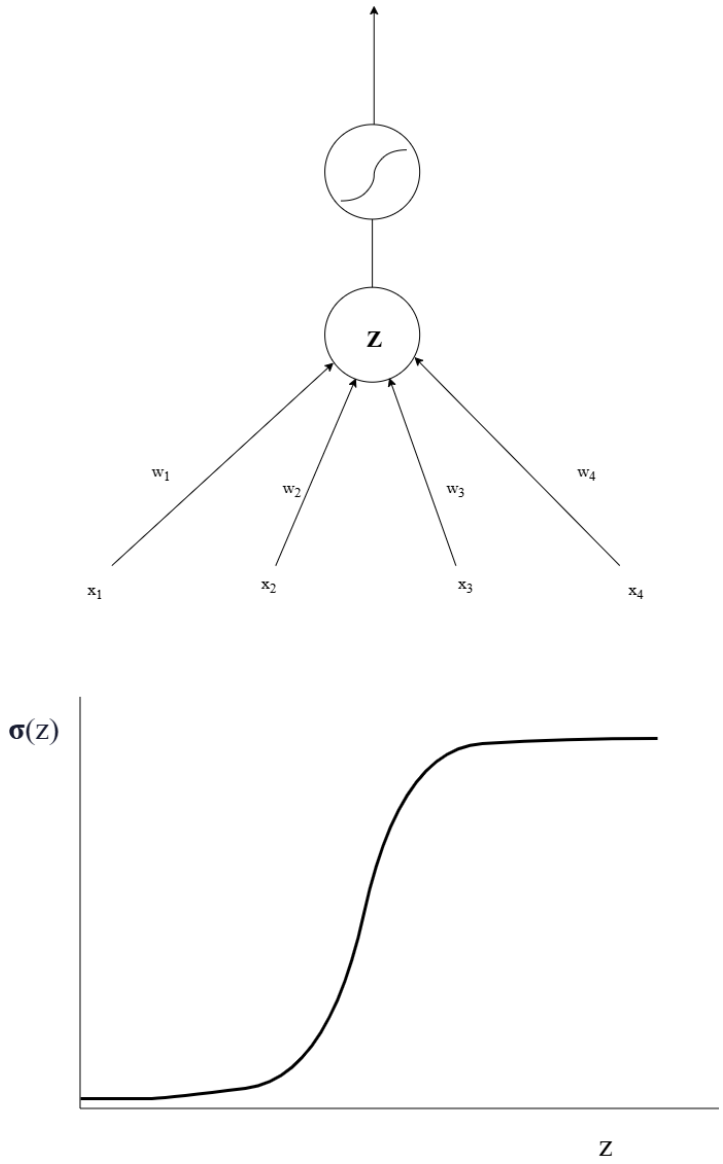


Figure 2.11: Schematic representation of Logistic Regression. The model estimates the probability of belonging to a class using a logistic function applied to a linear combination of input features.

k-Nearest Neighbors (k-NN)

The k-Nearest Neighbors (k-NN) algorithm is a non-parametric, instance-based learning method utilized extensively for both classification and regression tasks [48]. It operates on the fundamental principle of similarity, classifying or predicting a new data point based on the proximity of its features to those in the training dataset.

When presented with a novel instance $\mathbf{x}_{\text{test}}$, k-NN calculates the distance (typically Euclidean distance d) between $\mathbf{x}_{\text{test}}$ and every point in the training set D. It then identifies the k data points, the "neighbors", that exhibit the smallest distance.

- **For Classification:** The algorithm assigns $\mathbf{x}_{\text{test}}$ to the class most frequently represented among its k neighbors (a majority vote mechanism). The predicted class $\hat{y}$ is given by:

$$\hat{y} = \underset{c}{\text{argmax}} \sum_{i=1}^{k} I(y_i = c)$$

  where $I(\cdot)$ is the indicator function and y_i is the label of the i-th nearest neighbor.

- **For Regression:** The algorithm computes the average (or weighted average) of the target values y_i of the k neighbors to determine the predicted output $\hat{y}$:

$$\hat{y} = \frac{1}{k} \sum_{i=1}^{k} y_i$$

A significant advantage of k-NN is its non-parametric nature, meaning it makes no underlying assumptions about the distribution or structure of the data. This flexibility allows it to effectively model complex and non-linear decision boundaries. Furthermore, its conceptual simplicity facilitates easy implementation and interpretation.

The selection of k-NN in this study was motivated by its ability to capture local patterns in the data effectively and efficiently. It serves as a robust and easily interpretable baseline model against which the performance of more computationally intensive and complex algorithms—such as ensemble methods (Random Forest) or gradient boosting techniques (XGBoost, CatBoost), can be systematically compared and benchmarked. The model's efficacy is critically dependent on the optimal selection of the hyperparameter k and the appropriate choice of distance metric, factors that were carefully tuned during the model development phase.

Support Vector Machine (SVM)

The Support Vector Machine (SVM) is a powerful and versatile supervised learning algorithm used for both robust classification and regression tasks [49]. The core principle of the SVM algorithm is to find the optimal separating boundary, known as a hyperplane, that maximally separates the data points of different classes in a high-dimensional feature space. The key to its strength lies in the concept of the maximum margin, defined as the largest distance between the hyperplane and the data points closest to it from each class. These critical closest instances are termed the support vectors, as they uniquely determine the position and orientation of the optimal hyperplane (see Figure 2.12).

SVMs excel when processing high-dimensional data or datasets where the number of features significantly outweighs the number of samples ($p \gg n$). A notable advantage is their ability to perform non-linear classification through the use of the kernel trick. By employing various kernel functions, such as the Polynomial, Radial Basis Function (RBF), or Sigmoid kernels, SVMs implicitly map the original input features into a higher-dimensional space where linear separation becomes feasible. This mechanism grants the model high flexibility and inherent robustness against overfitting, particularly in complex feature spaces.

SVMs were selected for inclusion in this work due to their proven capacity to handle complex, non-linear decision boundaries effectively within potentially large feature spaces. Furthermore, their generalized performance is often achieved with a limited number of hyperparameters, providing a strong balance between predictive power, computational efficiency, and generalization capability for the specific dataset at hand.

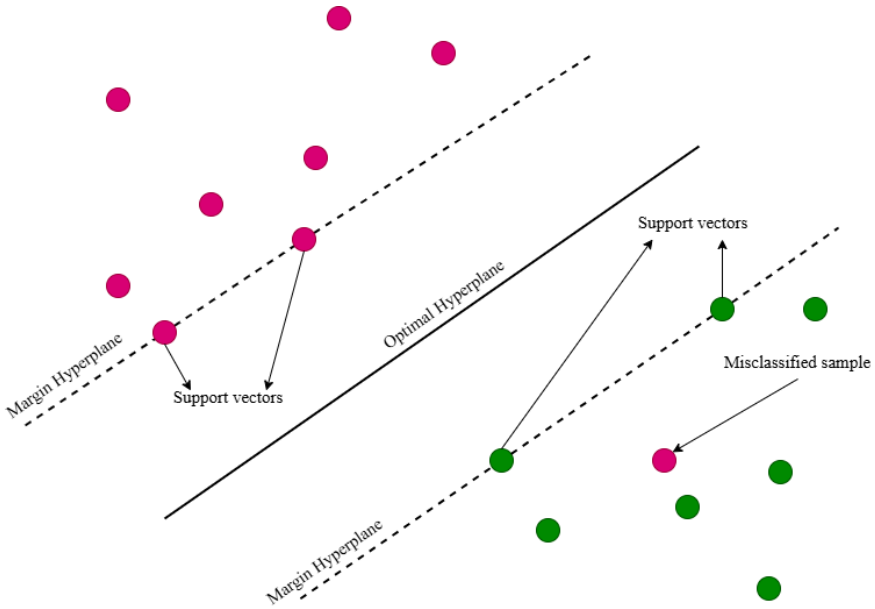


Figure 2.12: Schematic representation of the Support Vector Machine. The algorithm finds the optimal hyperplane that separates classes, maximizing the margin between support vectors.

Random Forest

The Random Forest (RF) is an ensemble learning method that belongs to the family of bagging (bootstrap aggregating) methods, which aim to reduce model variance and overfitting by training several independent estimators on randomly generated subsets of the training data [50]. Random Forest combines the predictions of multiple decision trees

to improve classification accuracy and robustness. Its general structure and principle of operation are shown in Figure 2.13.

Each decision tree in the forest is trained on a bootstrap sample, obtained by randomly sampling the training data with replacement. Furthermore, at each node of a tree, a random subset of features is selected when determining the best split. This double randomness, both in data sampling and feature selection, increases model diversity and prevents individual trees from becoming too correlated, resulting in a more generalizable ensemble.

During the prediction phase, each tree in the forest independently produces an output, and the final prediction is obtained through majority voting (in classification tasks) or averaging (in regression tasks). This aggregation mechanism allows Random Forests to achieve higher accuracy and stability compared to individual decision trees.

These models are particularly suitable for complex datasets with many features and non-linear relationships because they require limited hyperparameter tuning, are relatively robust to noise and outliers, and can provide estimates of feature importance, which contributes to model interpretability. However, their ensemble nature increases computational requirements and reduces transparency compared to single-tree models.

This model was chosen for this study due to its robustness in handling datasets with high-dimensional and potentially correlated features, as well as its ability to manage noise and outliers effectively. Additionally, the feature importance scores provided by Random Forest allow for a deeper understanding of which variables are most influential in the prediction, supporting both model interpretability and performance evaluation.

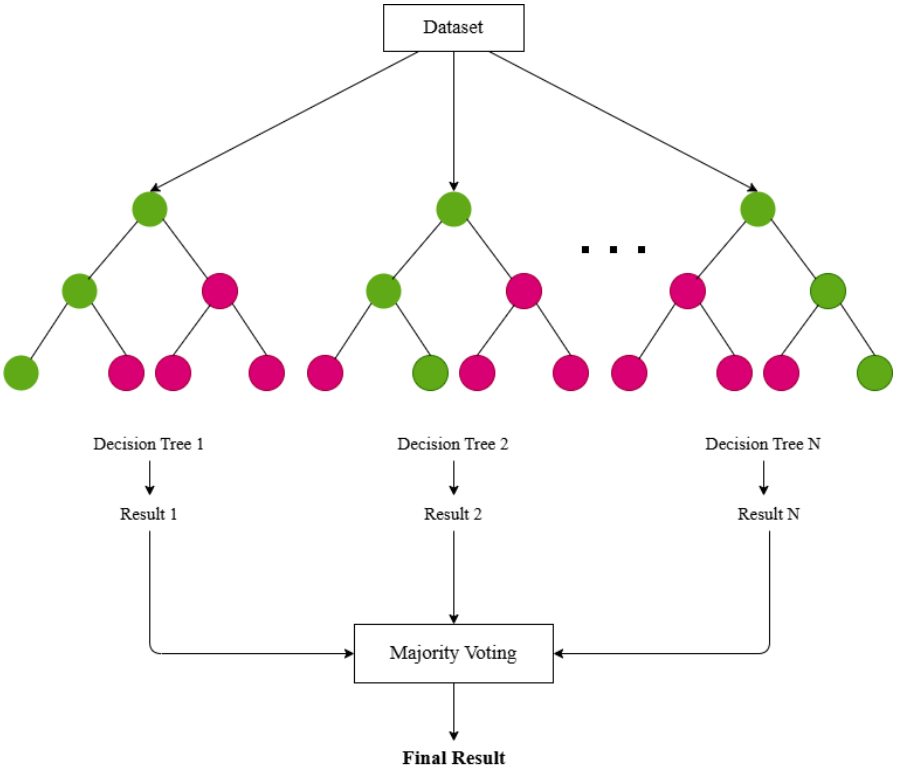


Figure 2.13: Schematic representation of the Random Forest structure. Multiple decision trees are trained on random subsets of the data and features, and their predictions are aggregated through a majority voting process.

XGBoost

XGBoost (Extreme Gradient Boosting) is a powerful ensemble learning algorithm based on gradient boosting, designed to improve prediction accuracy by sequentially building an ensemble of weak learners, typically decision trees [51]. Unlike bagging methods such as Random Forest, gradient boosting methods train each tree to correct the errors made by the previous trees, effectively focusing on difficult-to-predict samples.

The algorithm optimizes an objective function that combines a differentiable loss function, measuring the model's prediction error, and a regularization term, penalizing model complexity. At each iteration, a new tree is fit to the negative gradient of the loss function with respect to the current ensemble's predictions. This allows XGBoost to progressively minimize the training error while controlling overfitting through regularization techniques such as shrinkage, column subsampling, and tree pruning.

During prediction, the outputs of all trees are summed to produce the final prediction. The sequential nature of boosting enables XGBoost to capture complex, non-linear relationships in the data and to emphasize observations that are difficult to predict.

XGBoost is highly efficient and scalable, supporting parallel computation and handling missing values natively. It is particularly effective on structured tabular data and has been widely used in machine learning competitions due to its strong predictive performance. However, the sequential training process makes it less interpretable than simpler models, and careful tuning of hyperparameters is often required to achieve optimal results.

This model was chosen for this study due to its proven ability to handle datasets with complex feature interactions and potential class imbalance, as well as its strong predictive performance in tabular data. Additionally, its feature importance measures allow for insights into which variables contribute most to the model's predictions, supporting interpretability alongside performance.

CatBoost

CatBoost (Categorical Boosting) is a gradient boosting algorithm designed to handle categorical features natively while maintaining high predictive performance and stability [52]. Like other boosting algorithms, CatBoost builds an ensemble of decision trees sequentially, where each tree is trained to correct the errors of the previous ensemble, but it introduces specific techniques to reduce overfitting and improve performance on categorical data.

A key innovation of CatBoost is its method for handling categorical variables. Instead of requiring manual preprocessing or one-hot encoding, CatBoost transforms categorical features using a combination of target statistics and ordered boosting, which preserves the integrity of the training process and reduces target leakage. The algorithm also employs symmetric trees, where each split at a given depth uses the same splitting criterion, allowing for faster computation and better generalization.

During prediction, the outputs of all trees are aggregated to form the final result, similar to other gradient boosting methods. The sequential training enables CatBoost to capture complex non-linear relationships, while its specialized handling of categorical features ensures that information from non-numeric data is fully leveraged.

CatBoost is highly efficient, robust to overfitting, and can handle heterogeneous datasets containing both numerical and categorical features. It is particularly effective for tabular datasets with mixed data types and has been shown to achieve competitive performance with minimal parameter tuning.

This model was chosen for this study because the dataset contains several categorical variables, and CatBoost's native handling of these features eliminates the need for extensive preprocessing. Additionally, its strong predictive performance, robustness to overfitting, and ability to provide feature importance metrics make it a suitable choice for achieving both high accuracy and interpretability.

Table 2.6: Comparison of the main characteristics of the considered machine learning models.

Model	Type	Non-linear	Robust to Noise	Categorical Features	Interpretability	Hyperparameter Tuning
Random Forest	Ensemble (Bagging)	Yes	High	Needs encoding	Medium (feature importance)	Low–Medium
XGBoost	Ensemble (Boosting)	Yes	Medium–High	Needs encoding	Medium (feature importance)	Medium–High
CatBoost	Ensemble (Boosting)	Yes	Medium–High	Native handling	Medium (feature importance)	Medium
SVM	Kernel-based	Yes (with kernel)	Medium	Needs encoding	Low–Medium	Medium
k-NN	Instance-based	Yes	Low–Medium	Needs encoding	Low	Low (choice of k)
Logistic Regression	Linear	No	Low–Medium	Needs encoding	High	Low

2.2.2 Deep learning

Deep Learning (DL) is a subfield of Machine Learning based on multi-layered neural networks that can automatically learn hierarchical representations of data. Unlike traditional ML, which relies on manually engineered features, DL models can extract complex patterns directly from raw input, making them particularly suitable for high-dimensional data such as images, audio, or text.

In the specialized field of medical image analysis, DL has consistently shown an advantage over classical ML approaches in tasks including precise image segmentation, robust classification, and the identification of subtle anomalies. Given a sufficient dataset, DL models demonstrate a superior capacity to detect minute features within spatial and intensity patterns (characteristics often too subtle or non-linear to be formally articulated through conventional feature definition). This capability is especially critical in challenging diagnostic applications, such as the accurate detection of faint thermal shifts in wound images, which may be the earliest signals of developing infection or inflammation.

The operational distinction between DL and traditional ML is often contextualized by data volume and the requirement for model interpretability. While conventional ML retains its value for smaller datasets and offers inherently greater transparency, DL excels when processing large-scale datasets. Under these conditions, DL models can significantly enhance predictive accuracy by thoroughly extracting deep-seated, non-obvious data patterns. This study investigates the strategic deployment of DL methodologies to refine wound characterization, specifically by evaluating their efficacy in detecting minor thermal variations that may serve as sensitive biomarkers for early infectious or inflammatory states.

YOLO11

The You Only Look Once (YOLO) [53] framework is a highly versatile family of single-stage networks renowned for its speed and capability across diverse computer vision tasks, including object detection, segmentation, and motion tracking. For the classification of thermal wound images in this study, the classification-adapted variant,

`YOLOv11n-CLS`, was employed to perform a necessary binary categorization (infected vs. non-infected). YOLO models are released in scalable sizes (e.g., nano (`n`), small (`s`), medium (`m`)), and the nano variant was chosen due to its optimal balance between speed and efficiency, making it highly effective for scenarios involving limited datasets and requiring fast inference times.

The `YOLO11n-CLS` model was selected based on the following key reasons:

- **Efficiency and Speed:** The core YOLO principle of single-pass processing ensures extremely fast inference, making it advantageous for potential real-time integration into clinical settings. Furthermore, the selection of the `n` (nano) variant specifically minimizes the parameter count, which is crucial for reducing the risk of overfitting on smaller medical image datasets, thereby promoting robust generalization.

- **Automatic Feature Extraction and Depth:** Leveraging its deep convolutional architecture, the model autonomously learns complex, hierarchical features, thus eliminating the need for manual engineering. This capacity is enhanced by advanced components in the backbone, such as the optimized C3K2 blocks, which efficiently facilitate feature flow while reducing computational complexity.

- **Robustness to Spatial Variations and Multi-Scale Analysis:** The model's inherent global processing and its Neck structure, which integrates modules like the SPFF (Spatial Pyramid Pooling Fast), allow it to effectively handle considerable variations in wound size, shape, and position. The SPFF aggregates features across multiple scales, ensuring reliable pattern recognition regardless of object granularity.

- **Attention Mechanism for Focus:** The architecture integrates advanced attention mechanisms, such as the C2PSA (Cross Stage Partial with Spatial Attention) block. This feature enables the model to selectively focus on the most relevant, subtle thermal regions of the wound, which is critical for learning discriminative features that accurately signal infection or inflammation.

In the context of this study, YOLOv11n-CLS was used to classify thermal images of ulcers as either *infected* or *non-infected*. Its ability to automatically focus on relevant regions of the wound and learn discriminative features makes it a highly suitable choice, balancing the need for computational efficiency with the requirement for reliable predictive accuracy in medical applications.

Dataset Partitioning

A fundamental step in the preparation of data for artificial intelligence models is the partitioning of the dataset into distinct subsets with specific purposes. As illustrated in Figure 2.14, the data are typically divided into two main components: a construction set and a test set.

The construction set is used during model development and is further divided into a training set and a validation set. The training set provides the examples from which the model learns to recognize patterns and infer relationships between inputs and outputs. The validation set, on the other hand, is employed to monitor the learning process and assess the model's performance during training, supporting hyperparameter tuning and early stopping to prevent overfitting.

Finally, once model optimization is completed, the test set, which remains completely unseen during training, is used to perform a blind evaluation. This ensures an unbiased estimate of the model's generalization ability on new, unseen data.

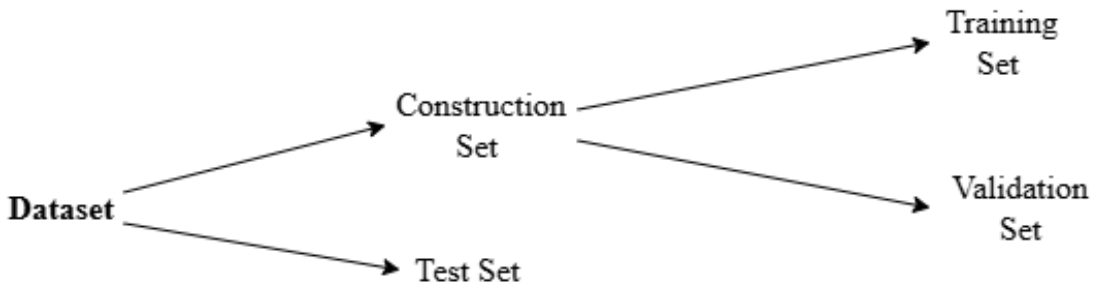


Figure 2.14: Schematic representation of dataset partitioning.

Class Balance and Data Augmentation

An equally crucial aspect in dataset preparation is the balance of class distributions within the training data. When certain categories are underrepresented, models tend to become biased toward the majority class, resulting in degraded predictive accuracy for minority categories and poor generalization.

To mitigate this issue, several complementary strategies are commonly employed:

- **Resampling techniques:** These include the oversampling of minority classes (either by duplicating samples or by generating synthetic examples, e.g., with SMOTE) and the undersampling of majority classes, both aimed at equalizing class proportions within the training set.
- **Data augmentation:** This approach increases the effective size and diversity of the training data by applying controlled transformations such as rotations, flips,

scaling, noise injection, or color perturbations. These transformations enrich the dataset variability without altering the semantic meaning of the samples.

- **Class weighting (cost-sensitive learning):** In many machine learning frameworks, imbalance can also be mitigated by assigning higher weights to minority classes during the loss computation. This weight balancing strategy ensures that errors on underrepresented categories contribute more strongly to the loss function, encouraging the model to learn them more effectively.

Maintaining a balanced training set, either through data-level or algorithm-level approaches, is essential for obtaining fair and robust models, especially in real-world scenarios where data acquisition is inherently unbalanced.

Cross-Validation

When the available dataset is limited, splitting it once into training and validation sets can lead to unstable or misleading performance estimates. To mitigate this issue, k-fold cross-validation is commonly employed.

In k-fold cross-validation, the dataset is divided into k equal parts, or "folds". The model is then trained k times, each time using a different fold as the validation set while the remaining folds are used for training [54]. The final performance metrics are obtained by averaging the results from all k runs.

This approach allows for a more reliable evaluation of model performance, maximizes the use of limited data, and reduces variability due to random partitioning. It is particularly useful when the dataset is too small to set aside a sufficiently large validation set.

2.3 Data processing

Due to the thermal camera's specific setup, two critical methodological constraints were encountered during data acquisition.

Firstly, each acquired thermal image was saved with the *Ironbow* colormap already applied. While this colormap is effective for visualizing the spatial distribution of temperature, ranging intuitively from blue (cold) to white (hot), this image processing step did not preserve the raw temperature data. The raw data represent the only true quantitative reference typically reported in the literature (as discussed in Section 1.2.3). Consequently, it was not possible to obtain absolute temperature values in degrees Celsius or to directly compare the measured values with those reported in previous studies.

Secondly, as described in Section 2.1.1, the device required that RGB and thermal images be captured consecutively rather than simultaneously. This temporal difference, combined with potential slight patient movement between captures, made it impossible to obtain perfectly aligned RGB-thermal pairs. This misalignment complicated wound identification in some cases and rendered any robust image segmentation based on the direct overlay of both modalities unfeasible.

Given these constraints, an alternative analytical strategy was adopted based on the analysis of image gradients. Since the pixel values in the *Ironbow* images are determined by the arbitrary scale of the colormap rather than by physical temperature, they lack direct physical meaning. Nevertheless, image gradients allow the extraction of relative information from the color distribution, effectively capturing variations in thermal intensity within each image. In this way, while the values remain relative, they still provide a meaningful quantification of the spatial patterns of heat across the wound, which can be used for subsequent comparative analysis.

2.3.1 Preprocessing

Prior to feature extraction, all images underwent a preprocessing phase aimed at ensuring consistency across the dataset and enhancing the relevant signal characteristics.
The procedures were tailored to the specific imaging modality, with distinct pipelines applied to thermal images and RGB photographs, as detailed below.

Thermal images

Given the constraints associated with the thermal images, a preprocessing pipeline was developed to standardize the data and facilitate subsequent analysis. The pipeline consists of several sequential steps, summarized in Figure 2.15.

Given the constraints associated with the thermal images, a preprocessing pipeline was developed to standardize the data and facilitate subsequent analysis. The pipeline consists of several sequential steps, summarized in Figure 2.15.

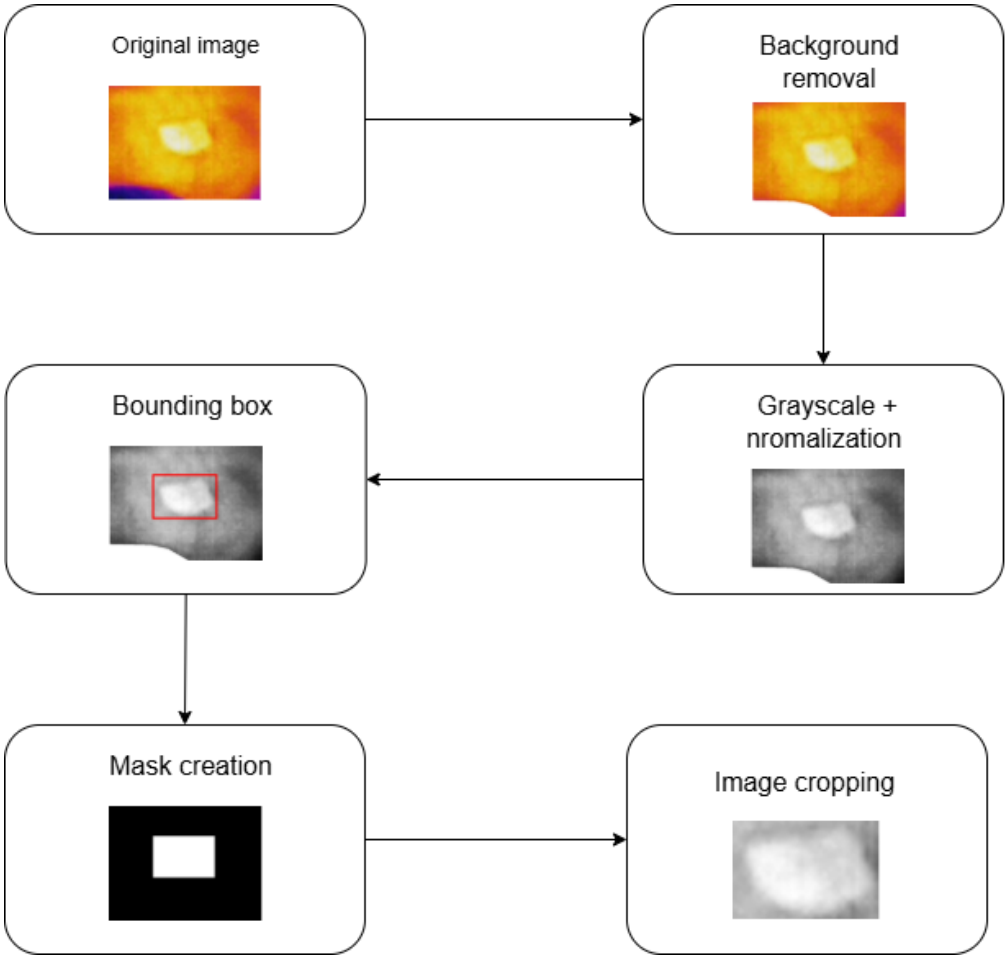
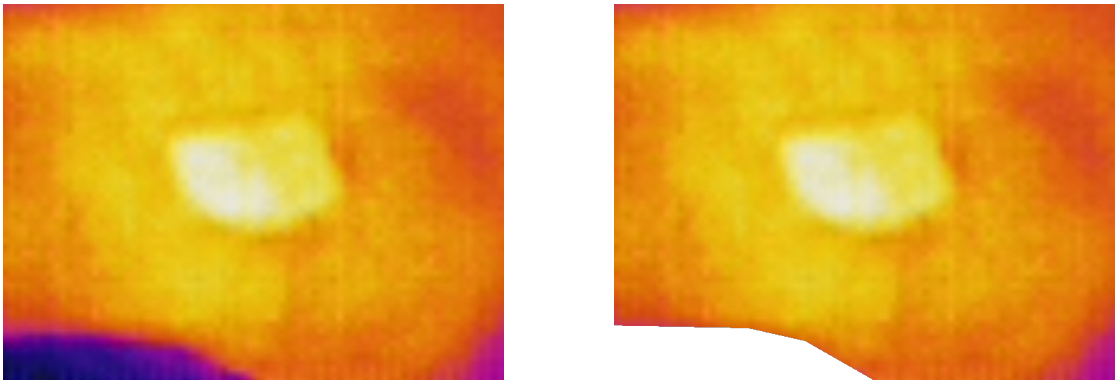


Figure 2.15: Pipeline

Background removal

The acquired thermal dataset included two types of images: (1) those capturing only the skin region and (2) those in which part of the surrounding environment (e.g., the examination table or background) was also visible. In the latter case, the *Ironbow* colormap caused these non-skin regions to appear blue, introducing variability unrelated to the actual temperature distribution. To address this issue, the background was manually removed from all images. The resulting files were saved in PNG format with transparent background, ensuring that only the regions of interest (i.e., skin) were retained for further processing. Figure 2.16 illustrates an example before and after background removal.



(a) Original thermal image. (b) Image after background removal.

Figure 2.16: Comparison of images before and after background removal.

Grayscale conversion and global intensity normalization

The conversion to grayscale was designed to achieve dataset-wide standardization of thermal intensity values. Given that the thermal images were originally saved with the *Ironbow* colormap, the pixel values corresponded to arbitrary color mappings rather than physical temperatures. Therefore, a normalization procedure was required to ensure consistency across all samples.

1. Range determination. A subset of the thermal images—specifically those captured without any background regions and thus containing only skin—was used to determine the global intensity range for normalization. These images were first converted from RGB to grayscale, focusing on the luminance channel that best represents the thermal intensity encoded by the colormap. By analyzing all pixel values across this subset, the global minimum and maximum grayscale intensities were identified as $I_{\text{min}} = 5$ and $I_{\text{max}} = 255$. This range was then adopted as a reference for the normalization of the entire dataset.

2. Global normalization. All thermal images (after background removal) were then converted to grayscale and normalized using the previously defined range $[5, 255]$, ensuring a consistent intensity scale across the dataset. Normalization was applied according to the following transformation:

$$I_{\text{norm}} = 255 \cdot \frac{I_{\text{raw}} - I_{\text{min}}}{I_{\text{max}} - I_{\text{min}}}$$

This step preserves the relative thermal gradients within each image while allowing reliable comparison between samples, despite the lack of absolute temperature calibra-

tion. Figure 2.17 shows a representative example of a thermal image after grayscale conversion and intensity normalization.

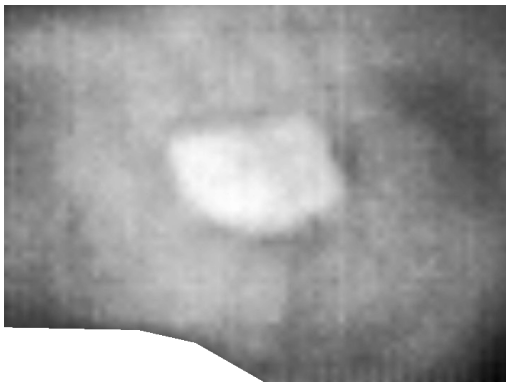


Figure 2.17: Thermal image after grayscale conversion and normalization.

Wound localization

To facilitate accurate wound localization and to constrain subsequent analyses to the relevant anatomical region, bounding boxes were manually drawn around each lesion. Each bounding box was displayed in red (RGB code: $255,0,0$) for optimal visual clarity. This step defines the initial Region of Interest (ROI), ensuring a precise spatial separation between the wound and the surrounding perilesional skin. The defined ROI was then used as a reference for the creation of binary masks and for the extraction of quantitative information from the wound area. Figure 2.18 illustrates an example of a lesion delineated by a red bounding box.

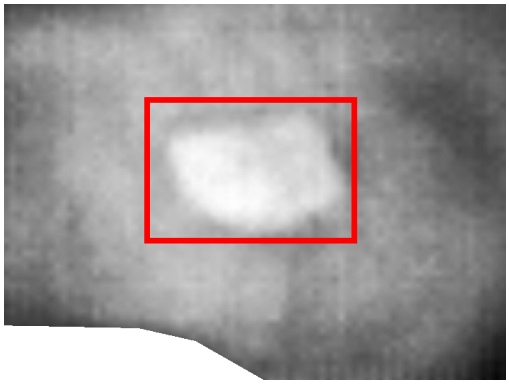
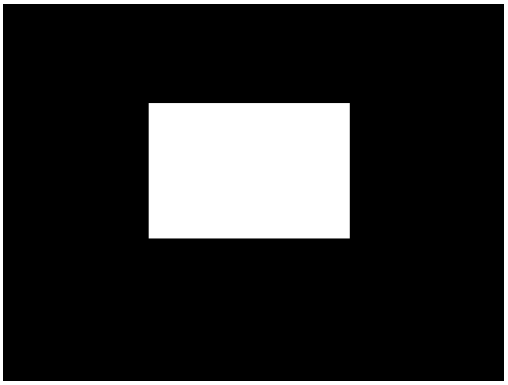


Figure 2.18: Grayscale thermal image with red bounding box sorrounding the wound.

Mask creation and image cropping

From the bounding box annotations, binary masks were generated for each wound. These masks were then applied to the grayscale thermal images, effectively cropping and isolating the Wound Region of Interest (W-ROI). This process ensured that the final extracted regions contained only the wound area while preserving the relevant thermal gradient information. Figure 2.19 illustrates the generated binary mask and the resulting cropped image, which is obtained by overlapping the binary mask onto the grayscale thermal image (see Figure 2.17).



(a) Generated binary mask of the wound area.

(b) Thermal image after applying the mask and cropping.

Figure 2.19: Final preprocessing step: the binary mask (a) and the corresponding thermal image obtained after applying the mask (b).

RGB images

RGB images underwent a dedicated preprocessing pipeline aimed at simplifying subsequent analysis and facilitating the use of artificial intelligence models.
Two main objectives guided this preprocessing:

1. For feature extraction in classical machine learning, it was important to isolate the wound region. This allows extracting features exclusively from the wound, avoiding confounding contributions from surrounding skin or background.

2. For classification using YOLO, which requires rectangular input images, it was necessary to retain a regular image shape while minimizing irrelevant background information, such as surrounding skin or unrelated objects.

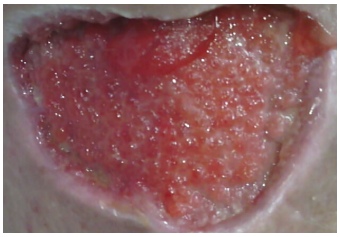
To achieve these goals, all RGB images were segmented using the *Wound Viewer* algorithm , already introduced in Section 2.1.1.
The segmentation outputs were saved in two complementary formats:

- Images with an alpha channel (PNG) containing only the wound region, used for feature extraction in machine learning models.
- Rectangular images including the surrounding skin, suitable as input for YOLO classification, ensuring a standardized rectangular shape while retaining the wound context.

Figure 2.20 illustrates the preprocessing workflow for an example RGB image, showing the original image, the segmented version for machine learning, and the rectangular version used for YOLO.

(a)
Original RGB image

(b)
Segmented for machine learning



(c)
Rectangular for YOLO

Figure 2.20: Preprocessing of RGB images: original, segmented for feature extraction, and rectangular for YOLO detection.

2.3.2 Models implementation

Dataset partitioning

The dataset was divided in order to maintain approximately the same proportion of infected cases between the construction set (training + validation) and the test set. Moreover, all the images belonging to the same patient were included exclusively either in the construction set or in the test set, thus avoiding data leakage and ensuring independence between the two subsets. This approach is crucial to prevent the model from learning patient-specific characteristics that could artificially improve performance, ensuring instead a fair evaluation on unseen patients.

Maintaining a percentage of infected images in the test set that is comparable to the real-world scenario is also essential, since the test phase should reflect the actual prevalence of infection expected in clinical applications. This design allows assessing model performance under realistic operating conditions.

The final dataset composition is summarized in Table 2.7.

Table 2.7: Dataset partitioning by patient and infection status.

Subset	**Number of images**	**Infected (%)**	**Non-infected (%)**
Training + Validation (Construction set)	323	30.0%	70.0%
Test set	89	34.8%	65.2%

This partition guarantees a balanced representation of infected and non-infected cases across the two sets while preserving patient-level separation. The same division was used consistently for both the machine learning and deep learning approaches.

Machine learning

The same general workflow was applied across all machine learning models developed in this thesis. Although different algorithms were tested, the overall methodology for data preparation, training, validation, and testing followed a consistent structure. The following description therefore applies to all six models, while model-specific differences are explicitly indicated.

Features extraction

Feature extraction was performed on both thermal and RGB images, while additional clinical information for each wound was manually annotated. All resulting data were organized into separate CSV tables according to their type and later combined in various configurations for training the machine learning models.

From the thermal images, the ten gradient-based features with the highest values were extracted. Gradients were computed considering multiple spatial distances and directions to capture variations in pixel intensity and texture across the wound surface. Specifically, the following parameters were used:

$$\text{distances} = [1, 3, 5, 10]\,pixels, \quad \text{angles} = [0^\circ, 45^\circ, 90^\circ, 135^\circ].$$

In addition, other statistical and texture-based descriptors were extracted to describe the thermal intensity distribution, including:

- Statistical descriptors: mean, standard deviation, minimum, maximum, median, range, skewness, and kurtosis
- GLCM-based texture descriptors: contrast, dissimilarity, homogeneity, energy, correlation, and ASM

Manual clinical features were associated with each wound rather than with the individual images, since both RGB and thermal images depict the same lesion. These variables included:

- Aetiology, describing the type of wound;
- WBP (Wound Bed Preparation) score, indicating wound status;
- Confidence, representing the reliability of infection labelling. In particular, images whose infection status was confirmed by biopsy were assigned the highest confidence (equal to 1, as this represents the gold standard), while those classified based on clinical assessment (C&H) were assigned lower, case-dependent confidence values.

From RGB images, color-based descriptors were extracted in both RGB and HSV color spaces. The computed variables included:

- RGB space: mean and standard deviation for each channel (R, G, B)
- HSV space: mean and standard deviation for each channel (H, S, V)

These features provided a compact representation of the color distribution and variability potentially associated with infection.

To evaluate the relative contribution of each feature group, different feature set combinations were tested during model development:

1. Gradient-based features only
2. Gradient + RGB features

3. Manual + Gradient features
4. Manual + Gradient + RGB features
5. Manual + Gradient + Thermal statistical features
6. Manual + Gradient + Thermal statistical + RGB features

These configurations allowed assessing how complementary information from thermal, visual, and clinical domains could improve infection classification performance.

K-fold cross-validation
Given the limited dataset size, the construction set was not split once into fixed training and validation subsets. Instead, a k-fold cross-validation strategy was adopted, as described in Section 2.2.2. In this approach, the construction set was divided into $k = 5$ equally sized folds: at each iteration, $k - 1 = 4$ folds were used for training and the remaining one for validation. All images from the same patient were kept within the same fold to prevent overlap between training and validation data, and the approximate proportion of infected wounds was maintained across folds. This ensured that both the distribution of infection cases and the patient-level separation were preserved within each validation step.

Dataset balancing
The dataset presented a moderate class imbalance, with infected wounds being less represented than non-infected ones. Balancing was therefore necessary to prevent the models from being biased toward the majority class and to ensure that the classifier remained sensitive to infected cases, which are of primary clinical relevance.

All balancing procedures were applied exclusively to the construction set (training and validation folds) within the 5-fold cross-validation framework, while the test set always preserved the natural class distribution to allow for an unbiased performance evaluation.

- **Weight-based balancing:** For most classifiers, namely Logistic Regression, Support Vector Machine (SVM), Random Forest, XGBoost, and CatBoost, the imbalance was managed using the `class_weight='balanced'` option in `scikit-learn`. This mechanism automatically assigns weights inversely proportional to class frequencies within each training subset, ensuring that minority classes contribute proportionally to the learning process.
- **Synthetic oversampling:** For the K-Nearest Neighbors (KNN) classifier, which does not natively support class weighting, the Synthetic Minority Oversampling Technique (SMOTE) was applied within each training fold. This algorithm generates synthetic samples of the minority class by interpolating between existing

minority instances, thus achieving a more balanced class distribution prior to model fitting.

This strategy ensured that all models were trained on balanced data while maintaining full separation between patients across folds and avoiding any leakage between training and validation sets.

Threshold optimization
Instead of using the default probability threshold of 0.5 for binary classification, a customized decision threshold was determined for each model to improve classification performance. For each validation fold, the following procedure was applied:

1. **The Receiver Operating Characteristic (ROC)** curve was computed. The ROC curve plots the true positive rate against the false positive rate for different threshold values, providing a visual representation of a classifier's ability to discriminate between positive and negative classes across all possible thresholds.

2. The threshold corresponding to the maximum F1-score was selected. The F1-score represents the harmonic mean of precision and recall, capturing a balance between these two metrics, which is particularly important in the presence of imbalanced datasets.

To ensure that the optimization produces meaningful thresholds, the search was restricted to a reasonable range between 0.4 and 0.9. Thresholds outside this interval would compromise the purpose of having a tunable decision boundary:

- Extremely low thresholds would classify nearly all samples as positive, yielding high recall but very low precision, which defeats the purpose of distinguishing between infected and non-infected cases.

- Extremely high thresholds would classify almost all samples as negative, resulting in poor sensitivity and failing to identify positive cases, again nullifying the objective of threshold optimization.

In addition to ROC analysis, the Precision–Recall (PR) curve was also considered. The PR curve plots precision against recall for different thresholds and is particularly informative in imbalanced datasets, as it focuses directly on the positive class performance rather than being influenced by the number of true negatives.

An example of a ROC curve used to determine the optimal threshold is shown in Figure 2.21. The red marker indicates the threshold corresponding to the highest F1-score.

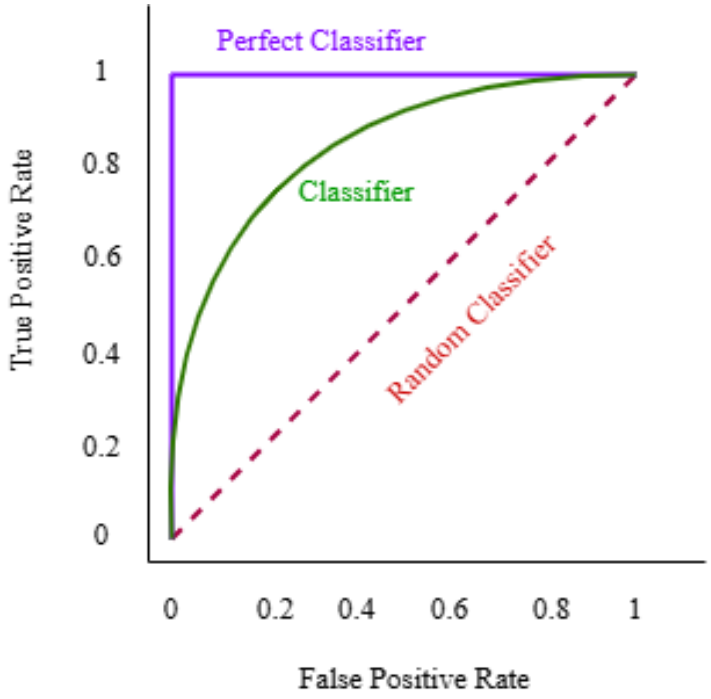


Figure 2.21: Example of ROC curve used for threshold optimization. The red marker indicates the threshold corresponding to the maximum F1-score.

Model saving and final evaluation

After cross-validation, two final models were saved for each algorithm:

1. The best-fold model, corresponding to the fold with the highest F1-score.
2. The mean model, retrained on the entire construction set using the mean of the optimal thresholds identified across folds.

This dual saving strategy was adopted to balance between performance and generalization. While the best-fold model usually achieves the highest validation performance, it might also reflect favorable data partitioning or mild overfitting to specific samples. The mean model, instead, benefits from being trained on all available construction data, providing more stable and generalizable estimates.

Both models were subsequently evaluated on the independent test set, allowing a comparison between fold-specific and overall-trained configurations and providing a robust assessment of model performance under unseen conditions.

Deep Learning

The construction set, previously used for the machine learning experiments, was further divided into training and validation subsets to allow model development and hyper-

parameter tuning. The test set remained unchanged, ensuring a consistent evaluation framework across both machine learning and deep learning experiments.

During this split, all images belonging to the same patient were assigned exclusively to either the training or validation subset to prevent data leakage. Additionally, the split was designed to approximately preserve the proportion of infected images across the two subsets, although exact matching was not possible due to the limited dataset size.

To mitigate class imbalance within the training set, data augmentation was applied to the infected images until the number of infected and non-infected samples was approximately equal. The augmentation was implemented using the `Albumentations` library and included a composition of random geometric transformations such as `HorizontalFlip`, `VerticalFlip`, `Rotate`, and `RandomRotate90`. This approach increased the diversity of the training data and reduced the risk of overfitting, while preserving the original characteristics of the wounds.

The resulting distribution of images across the training, validation, and test subsets is summarized in Table 2.8.

Table 2.8: Distribution of images across training, validation, and test subsets for YOLOv11 experiments.

Subset	**Number of images**	**Infected (%)**	**Non-infected (%)**
Training	262	32.1	67.9
Validation	61	21.3	78.7
Test	89	34.8	65.2

Two separate training runs were performed using the YOLO11 classification model (YOLO11n-cls): one using only thermal images and another combining thermal and RGB images.

YOLO11 Classification with Thermal Images

For the thermal-only run, images were used directly as input to the YOLO11n-cls network. Standard data augmentation techniques were applied to improve model generalization and reduce overfitting. These included horizontal flipping, small rotations, brightness and contrast adjustments, and Gaussian noise. Augmentations were implemented using the `Albumentations` library.

The YOLO11n-cls model was trained for 100 epochs, with early stopping triggered

after 10 epochs without improvement on the validation set. This procedure ensures optimal generalization while preventing overfitting.

YOLO11 Classification with Thermal and RGB Images

For the combined thermal and RGB run, separate YOLO11n-cls models were trained for each modality using the same augmentation and training settings. During inference a logical OR operation was applied between the binary predictions from the thermal model, $\hat{y}_{\text{thermal}}$, and from the RGB model, $\hat{y}_{\text{RGB}}$. Formally, the combined prediction for each image was defined as:

$$\hat{y}_{\text{combined}} = \hat{y}_{\text{thermal}} \text{ OR } \hat{y}_{\text{RGB}} \tag{2.9}$$

Here, $\hat{y}_{\text{thermal}}$ and $\hat{y}_{\text{RGB}}$ represent the binary predictions of infection from the respective models, while $\hat{y}_{\text{combined}}$ is the final classification used for evaluation.

This OR-based combination strategy was chosen to prioritize recall, aiming to identify as many positive cases as possible, which is particularly important in a screening context. By considering an image as positive if either model predicts infection, this approach reduces the risk of false negatives, ensuring that potential infections are less likely to be missed, even at the cost of a slight increase in false positives.

2.3.3 Evaluation Metrics

To evaluate the performance of the machine learning and deep learning models, several standard classification metrics were computed: *accuracy*, *balanced accuracy*, *precision*, *recall*, and *F1-score*. These metrics were derived from the confusion matrix, which summarizes the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), as illustrated in Figure 2.22.

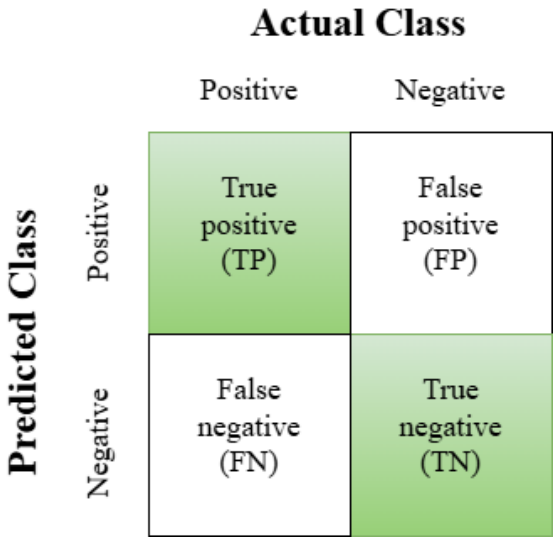


Figure 2.22: Schematic representation of a confusion matrix, illustrating the relationship between predicted and actual class labels. TP = True Positives, TN = True Negatives, FP = False Positives, FN = False Negatives.

Accuracy

Accuracy measures the overall proportion of correctly classified samples among all predictions, providing a general indication of the model's ability to correctly assign samples to their true classes. The metric is defined in Equation (2.10):

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.10}$$

Although accuracy is simple and intuitive, it can be misleading when the dataset is imbalanced, as it may be dominated by the majority class and therefore fail to reflect poor performance on minority classes.

Balanced Accuracy

To overcome the limitations of standard accuracy, balanced accuracy was employed. This metric compensates for class imbalance by averaging the recall obtained for each

class, giving equal weight to both the positive and negative categories, as shown in Equation (2.11):

$$\text{Balanced Accuracy} = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right) \tag{2.11}$$

Balanced accuracy thus provides a fairer evaluation of performance across classes, especially when one class is underrepresented.

Precision

Precision quantifies the proportion of correctly predicted positive samples among all samples predicted as positive (Equation (2.12)):

$$\text{Precision} = \frac{TP}{TP+FP} \tag{2.12}$$

High precision indicates that the model produces few false positives, meaning that when it predicts a positive case, it is likely correct. This metric is particularly relevant in tasks where false positives are costly, such as medical diagnosis or quality control.

Recall

Recall, also referred to as sensitivity or true positive rate, measures the ability of the model to correctly identify positive samples, as reported in Equation (2.13):

$$\text{Recall} = \frac{TP}{TP+FN} \tag{2.13}$$

High recall indicates that the model successfully detects most of the actual positive cases, minimizing false negatives. This metric is particularly important in contexts where missing a positive instance is undesirable, such as disease detection or anomaly identification.

F1-Score

The F1-score combines precision and recall into a single harmonic mean, balancing the trade-off between false positives and false negatives (Equation (2.14)):

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2.14}$$

This metric is especially useful when the dataset is imbalanced, as it simultaneously considers both types of classification errors. A high F1-score indicates that the model achieves a good compromise between precision and recall, making it a robust indicator of overall classification performance.

Results and discussion

3.1 Machine Learning Models Results

The following section presents the results obtained using various machine learning techniques applied to the dataset.
For all methods, performance is reported on both the validation and test sets, considering all combinations of RGB and thermal features, in accordance with the criteria described in Section 2.2.1.
Threshold values used for classification are also included where relevant. All metrics, except for the threshold, are expressed as percentages. For each configuration, two sets of results are reported: those obtained from the average metrics across the cross-validation folds (AF) and those from the best-performing fold (BF).

In the subsections below, the performance of each machine learning method is discussed in detail, highlighting the influence of different feature combinations, the stability of results across folds, and any notable patterns observed in terms of accuracy, balanced accuracy, precision, recall, and F1-score.

Logistic Regression

The logistic regression model was selected as a baseline method due to its simplicity and interpretability. The results obtained using this model are summarized in Table 3.9. Overall, logistic regression achieved satisfactory performance on the validation set across all feature combinations. However, in some cases the F1-score remained below 50%, suggesting that the balance between precision and recall was suboptimal for certain feature configurations.

The first set of experiments, which included only gradient-based features (with and without RGB information), yielded particularly low F1 and balanced accuracy (BA) values on the test set. For this reason, these configurations were excluded from further consideration when selecting the best-performing model.

In contrast, when combining manual and gradient-based features (second group of experiments), the inclusion of RGB-derived features led to a slight yet consistent

improvement in both BA and F1-score for the validation and test sets. However, this improvement was accompanied by a decrease in recall. Recall is a key metric in this context, as a higher recall indicates that more true positive cases are correctly identified, even if this comes at the cost of a moderate increase in false positives. From a screening perspective, it is preferable to prioritize recall over precision, as this reduces the risk of missing positive cases, provided that precision does not decrease excessively.

The addition of statistical features did not result in a clear or consistent improvement in performance.

Within each feature configuration, the performance obtained on the best fold (BF) tended to drop substantially when moving from validation to test data. This degradation suggests the presence of overfitting on specific folds, which justifies the use of the averaged results across folds (AF) as a more reliable performance estimate.

Considering these aspects, and following the principle of Occam's razor (namely, seeking the simplest model capable of achieving competitive performance) [55]the best overall results were obtained using only gradient-based and manual features. This configuration, highlighted in yellow in Table 3.9, represents the most balanced and interpretable solution.

In conclusion, despite its simplicity, logistic regression achieved satisfactory and stable performance across most feature combinations. This can be attributed to its inherent robustness and ability to capture linear relationships between the features and the outcome variable while maintaining low model complexity. These characteristics allow the model to generalize effectively on limited datasets and reduce the risk of overfitting, particularly on the validation data. Moreover, the presence of features with approximately linearly separable patterns further supports the model's ability to achieve competitive results even without the use of more complex architectures.

Table 3.9: Performance of Logistic Regression across feature combination.

Features	Fold	No RGB									With RGB								
		Thre	Val				Test				Thre	Val				Test			
			BA	Rec	Pre	F1	BA	Rec	Pre	F1		BA	Rec	Pre	F1	BA	Rec	Pre	F1
Only gradient	BF	0.56	81.5	87.5	53.6	66.7	47.6	19.4	30.0	23.5	0.69	81.4	75.0	66.7	70.6	55.0	29.0	45.0	35.3
	AF	0.5	67.8	71.6	46.3	54.7	53.1	35.5	39.3	37.3	0.55	69.9	66.3	55.0	59.4	56.2	45.2	42.4	43.8
Gradient + manual	BF	0.47	84.5	92.6	73.5	82	63.0	74.2	45.1	56.1	0.74	87.6	81.2	81.2	81.2	69.8	51.6	69.6	59.3
	AF	0.51	73.8	74.1	53.5	61.3	66.4	74.2	48.9	59	0.53	72.7	68.5	57.1	60.9	72.2	80.6	54.3	64.9
Gradient + manual + statistical	BF	0.45	74.0	76.7	69.7	50.6	57.1	67.7	40.4	50.6	0.76	86.5	81.2	76.5	78.8	66.5	45.2	66.7	53.8
	AF	0.52	70.3	67.1	51.2	57.5	65.0	64.5	50.0	56.3	0.53	73.1	73.3	56.4	61.3	66.5	71.0	50.0	58.7

k-NN

The results obtained using the K-Nearest Neighbors (KNN) model are summarized in Table 3.10. Performance is reported on both the validation and test sets, considering all combinations of RGB, thermal, gradient, and manual features, following the criteria described in Section 2.2.1. All metrics, except for the threshold, are expressed as percentages. For each configuration, results are reported both as averages across cross-validation folds (AF) and for the best-performing fold (BF).

When using only gradient-based features, overall performance was relatively poor. The addition of RGB-derived features further degraded performance, and this configuration was therefore excluded from further consideration. In contrast, the inclusion of manual features led to a substantial improvement in performance. However, when RGB features were added to this combination, no significant improvement was observed; in some cases, performance was even slightly worse than without RGB features.

The addition of statistical features resulted in a noticeable improvement in F1-score and recall on the validation set, both with and without RGB features. Nevertheless, a marked discrepancy between validation and test performance was observed, indicating a high degree of overfitting. This overfitting likely arises because KNN relies on the distances between feature vectors; when many features, particularly correlated or high-dimensional features, are added, the model becomes more sensitive to the specific distribution of the training data, reducing generalization to unseen test data.

Overall, as with the logistic regression model, the combination of manual and gradient-based features consistently produced the most balanced and interpretable results. This configuration is highlighted in yellow in Table 3.10. It is particularly noteworthy that, when considering the averaged results across folds (AF), performance remains relatively stable between validation and test sets. In contrast, for the best-performing fold (BF), a substantial drop in performance from validation to test is observed, further emphasizing the reliability of the AF metrics for model selection.

Table 3.10: Performance of k-NN across feature combination.

Features	Fold	No RGB									With RGB								
		Thre	Val				Test				Thre	Val				Test			
			BA	Rec	Pre	F1	BA	Rec	Pre	F1		BA	Rec	Pre	F1	BA	Rec	Pre	F1
Only gradient	BF	0.44	65.7	81.5	53.7	64.7	54.1	54.8	38.6	45.3	0.50	63.1	63.0	54.8	58.6	55.9	54.8	40.5	46.6
	AF	0.48	61.7	65.2	40.9	49.2	58.5	54.8	43.6	48.6	0.51	55.8	55.4	35.5	41.6	57.3	64.5	40.8	50.0
Gradient + manual	BF	0.45	78.4	87.5	48.3	62.2	56.6	58.1	40.9	48.0	0.56	68.9	66.7	62.1	64.3	58.7	48.4	45.5	46.9
	AF	0.51	64.3	57.1	45.6	48.3	65.0	64.5	50.0	56.3	0.59	59.7	51.9	39.0	43.5	65.1	61.3	51.4	55.9
Gradient + manual + statistical	BF	0.40	73.6	70.0	72.4	71.2	48.6	64.5	33.9	44.4	0.57	73.3	70.4	67.9	69.1	50.3	41.9	35.1	38.2
	AF	0.53	68.2	67.0	48.0	54.4	59.8	64.5	43.5	51.9	0.54	64.2	62.5	44.8	50.3	54.3	51.6	39.0	44.4

SVM

The results obtained using the Support Vector Machine (SVM) model are summarized in Table 3.11.

When using only gradient-based features, both with and without RGB information, the model achieved relatively low balanced accuracy, with particularly poor recall and F1-scores, especially on the test set. These results are therefore considered unsatisfactory.

When manual features were added, performance improved significantly, both in the RGB and non-RGB configurations. Slightly better results were obtained with RGB features on the validation set, whereas the opposite trend was observed on the test set, where the model without RGB performed better.

The inclusion of statistical features led to a further improvement in performance, with comparable results between the RGB and non-RGB configurations. Consequently, the configuration without RGB features was selected as the optimal one, as it represents a simpler and more efficient engineering solution. This choice follows the *principle of model parsimony*, which states that among models achieving similar results, the simplest one, in terms of structure and number of features, should be preferred, as it tends to generalize better and is easier to interpret.

A general trend observed in the SVM results is that precision tends to be consistently higher than recall, which is not ideal for the target clinical application. In this context, a higher recall is preferable, as it reduces the likelihood of missing positive cases, even at the cost of a moderate decrease in precision.

The observed behavior of the SVM model may be explained by several factors. First, SVMs are particularly sensitive to the distribution and scaling of features; the combination of heterogeneous feature types (e.g., gradient, manual, and statistical) may have affected the optimal separation of classes in the feature space. Moreover, the limited dataset size likely constrained the model's ability to define a robust decision boundary,

especially when using nonlinear kernels. These factors, combined with potential class imbalance, may explain the variability observed between validation and test performance.

As for the comparison between the two evaluation schemes, results obtained from the averaged folds (AF) show more stable and consistent performance between validation and test sets. In contrast, the best-performing fold (BF) tends to exhibit a more pronounced drop in performance on the test data, further confirming that the averaged approach provides a more reliable and less overfitted estimate of the model's generalization ability.

Table 3.11: Performance of SVM across feature combination.

Features	Fold	No RGB									With RGB								
		Thre	Val				Test				Thre	Val				Test			
			BA	Rec	Pre	F1	BA	Rec	Pre	F1		BA	Rec	Pre	F1	BA	Rec	Pre	F1
Only gradient	BF	0.40	63.6	31.2	71.4	43.5	50.5	9.70	37.5	15.4	0.40	71.0	56.2	56.2	56.2	56.1	22.6	53.8	31.8
	AF	0.40	54.5	16.1	36.7	19.7	51.3	12.9	40.0	19.5	0.40	58.7	30.0	47.1	34.4	52.8	19.4	42.9	26.7
Gradient + manual	BF	0.40	76.1	62.5	66.7	64.5	69.5	58.1	62.1	60.0	0.53	85.5	81.2	72.2	76.5	67.5	41.9	76.5	54.2
	AF	0.42	61.1	35.2	52.7	38.4	64.7	48.4	57.7	52.6	0.43	65.0	45.1	61.4	47.5	68.0	51.6	64.0	57.1
Gradient + manual + statistical	BF	0.46	85.5	75.0	85.7	80.0	69.8	51.6	69.6	59.3	0.42	83.5	81.2	65.0	72.2	68.2	48.4	68.2	56.6
	AF	0.43	69.8	51.3	64.2	55.6	62.3	41.9	56.5	48.1	0.40	64.5	46.0	54.0	44.9	65.5	51.6	57.1	54.2

Random Forest

The results obtained using the Random Forest model are summarized in Table 3.12. When using only gradient-based features, both with and without RGB information, the model achieved low F1-scores and recall values on both the validation and test sets, indicating limited discriminative ability in this configuration.

The inclusion of manual features led to a clear improvement in performance, with similar metrics observed between the RGB and non-RGB configurations. However, when statistical features were also added, the overall performance degraded, with lower F1-scores and balanced accuracy values across both validation and test sets, regardless of the inclusion of RGB features. This decline may be attributed to the introduction of redundant or noisy information from the statistical features, which can lead to reduced model generalization. Random Forest models, while robust to feature variability, can still be affected by irrelevant or weakly correlated variables, especially when the dataset is relatively small, as these may obscure the most informative feature patterns during tree construction.

Based on these results, the configuration using gradient-based and manual features without RGB information was selected as the most suitable one. This configuration

achieved stable and consistent results between the validation and test sets when using averaged metrics across folds (AF), suggesting good generalization capability and robustness. In contrast, results from the best-performing fold (BF) showed a more pronounced drop in performance on the test data, further confirming that the averaged approach provides a more reliable estimate of the model's actual predictive ability.

Overall, Random Forest demonstrated satisfactory performance and stability when appropriately constrained to informative feature subsets. Its ensemble structure and inherent capacity to handle non-linear relationships make it a suitable choice for heterogeneous feature sets, provided that the feature space remains compact and well-curated.

Table 3.12: Performance of Random Forest across feature combination.

Features	Fold	No RGB									With RGB								
		Thre	Val				Test				Thre	Val				Test			
			BA	Rec	Pre	F1	BA	Rec	Pre	F1		BA	Rec	Pre	F1	BA	Rec	Pre	F1
Only gradient	BF	0.40	66.8	50.0	50.0	50.0	54.5	45.2	40.0	42.4	0.45	71.0	62.5	50.0	55.6	61.6	38.7	57.1	46.2
	AF	0.40	57.0	41.5	39.2	38.8	51.1	41.9	36.1	38.8	0.41	57.5	36.6	40.9	35.8	58.9	41.9	48.1	44.8
Gradient + manual	BF	0.40	77.3	75.0	54.5	63.2	61.8	58.1	47.4	52.2	0.40	76.3	75.0	52.2	61.5	66.0	61.3	52.8	56.7
	AF	0.40	63.0	47.9	49.4	46.8	62.8	54.8	50.0	52.3	0.41	61.7	43.6	47.2	41.8	68.6	61.3	57.6	59.4
Gradient + manual + statistical	BF	0.40	74.2	68.8	52.4	59.5	63.6	54.8	51.5	53.1	0.46	79.3	68.8	68.8	68.8	60.0	35.5	55.0	43.1
	AF	0.40	56.1	36.0	44.1	35.0	63.5	58.1	50.0	53.7	0.41	59.5	38.1	48.1	37.8	63.0	48.4	53.6	50.8

XGBoost

The results obtained using the XGBoost model are summarized in Table 3.13. When considering only gradient-based features, the overall performance remained rather low across both validation and test sets. In particular, in the configuration without RGB features, a notable drop in performance was observed on the test set, indicating poor generalization. Conversely, when RGB features were included, the validation performance slightly decreased, but the test performance improved. This behavior might be explained by the fact that RGB-derived features capture complementary information that enhances generalization on unseen data, even if they increase the model's complexity and reduce training stability.

When manual features were added, validation performance slightly decreased for the configuration without RGB, but test performance improved. This trend may suggest that manual features, despite introducing some degree of redundancy during training, contribute to better generalization by providing more semantically meaningful information. In contrast, the inclusion of RGB features led to relatively stable validation results but a substantial improvement on the test set, supporting the idea that RGB-based information

helps XGBoost capture more generalizable patterns when combined with handcrafted features.

The addition of statistical features did not significantly affect validation performance in either configuration. However, test performance improved consistently, suggesting that the statistical descriptors, while not directly enhancing cross-validation metrics, contributed to a better overall model robustness. This effect may arise from the ensemble nature of XGBoost, which benefits from additional, moderately informative features by leveraging its regularization mechanisms and boosting strategy to mitigate overfitting.

Overall, XGBoost demonstrated a stable and interpretable behavior across the different feature combinations, with gradual improvements in generalization as the feature set became more comprehensive. The results highlight the model's ability to balance bias and variance effectively through regularization, even in the presence of heterogeneous and partially redundant feature sets.

Table 3.13: Performance of XGBoost across feature combination.

Features	Fold	No RGB									With RGB								
		Thre	Val				Test				Thre	Val				Test			
			BA	Rec	Pre	F1	BA	Rec	Pre	F1		BA	Rec	Pre	F1	BA	Rec	Pre	F1
Only gradient	BF	0.42	76.2	68.8	57.9	62.9	48.4	22.6	31.8	26.4	0.48	70.1	68.8	44.0	53.7	59.0	38.7	50.0	43.6
	AF	0.43	61.3	47.6	43.3	44.0	52.3	32.3	38.5	35.1	0.42	55.3	34.5	37.3	33.1	61.4	45.2	51.9	48.3
Gradient + manual	BF	0.40	64.6	55.6	60.0	57.7	50.6	32.3	35.7	33.9	0.40	72.1	62.5	52.6	57.1	66.7	64.5	52.6	58.0
	AF	0.40	61.1	42.1	49.6	43.8	62.0	51.6	50.0	50.8	0.40	57.9	35.2	41.7	35.7	71.2	61.3	63.3	62.3
Gradient + manual + statistical	BF	0.40	79.3	68.8	68.8	68.8	65.4	54.8	54.8	54.8	0.40	77.2	68.8	61.1	64.7	69.5	58.1	62.1	60.0
	AF	0.40	60.4	35.6	51.6	37.8	68.5	64.5	55.6	59.7	0.40	60.6	36.5	49.9	38.6	67.9	54.8	60.7	57.6

CatBoost

The results obtained using the CatBoost model are summarized in Table 3.14. When only gradient-based features were used, the overall performance remained relatively low. However, in the configuration without RGB features, the results were consistent between validation and test sets, indicating stable generalization. In contrast, when RGB features were included, the performance improved on the test set, suggesting that the model was able to exploit the additional visual information to some extent.

By adding manual features, the overall performance increased substantially, particularly in the configuration without RGB data. In this case, the average-fold (AF) results showed strong consistency between validation and test sets, reflecting a robust model behavior. For the configuration with RGB features, however, the best-fold (BF) performance dropped significantly on the test set, while the average-fold results improved.

This discrepancy indicates that the model may have overfitted specific folds when RGB features were present. For this reason, the configuration combining gradient-based and manual features without RGB information was considered the most stable and was thus selected as the best-performing setup.

When statistical features were added, the performance for the configuration without RGB remained approximately stable, with a slight decrease in recall. In contrast, in the configuration with RGB features, a performance improvement was observed on the validation set, but this gain did not translate to the test data, suggesting potential overfitting to the validation samples.

Overall, the combination of manual and gradient-based features without RGB data offered the most reliable and consistent results. In particular, the average-fold configuration maintained balanced and stable metrics across validation and test sets, indicating that the CatBoost model achieved a good compromise between fitting capacity and generalization ability.

Table 3.14: Performance of CatBoost across feature combination.

Features	Fold	No RGB									With RGB								
		Thre	Val				Test				Thre	Val				Test			
			BA	Rec	Pre	F1	BA	Rec	Pre	F1		BA	Rec	Pre	F1	BA	Rec	Pre	F1
Only gradient	BF	0.40	66.9	56.2	45.0	50.0	52.8	45.2	37.8	41.2	0.40	63.8	56.2	39.1	46.2	64.4	58.1	51.4	54.5
	AF	0.40	54.7	45.2	35.2	38.8	51.1	41.9	36.1	38.8	0.40	52.6	31.6	32.5	29.2	64.5	54.8	53.1	54.0
Gradient + manual	BF	0.40	80.4	75.0	63.2	68.6	65.1	61.3	51.4	55.9	0.45	71.2	55.6	75.0	63.8	62.2	45.2	53.8	49.1
	AF	0.40	65.2	51.3	51.1	49.7	62.5	61.3	47.5	53.5	0.44	61.1	44.4	44.0	41.6	70.9	67.7	58.3	62.7
Gradient + manual + statistical	BF	0.40	80.4	75.0	63.2	68.6	64.4	58.1	51.4	54.5	0.67	84.4	75.0	80.0	77.4	64.2	38.7	66.7	49.0
	AF	0.40	62.8	45.1	51.0	44.4	64.4	58.1	51.4	54.5	0.49	64.1	42.4	55.5	44.8	67.8	58.1	58.1	58.1

3.2 Deep Learning Results

The results obtained using the YOLO11cls model are summarized in Table 3.15.
Overall, the performance of this deep learning approach is unsatisfactory across multiple metrics.

Significant discrepancies are observed between validation and test results, with the test set showing better performance. This difference may be partially explained by the distribution of infected cases: the validation set contained a lower percentage of positive cases compared to the test set, which likely affected model evaluation.

Several factors likely contributed to the poor performance. First, the dataset remains relatively small, which limits the ability of deep neural networks to learn robust representations and generalize to unseen data. Second, the high variability among images, including differences in quality, illumination, and acquisition conditions, as well as substantial heterogeneity between classes, may have hindered the model's ability to learn consistent patterns during training.

It is important to note that the primary goal of this experiment was to serve as a proof of concept for applying deep learning methods to the problem, rather than to achieve optimal performance.
Despite the limited results, these preliminary experiments highlight the need for a larger and more balanced dataset to fully exploit the potential of deep learning approaches in this clinical context.

Table 3.15: Performance of the YOLO11cls model on the validation and test sets.

Features	Validation				Test			
	BA	Rec	Pre	F1	BA	Rec	Pre	F1
Thermal images	49.0	23.1	20.0	21.4	49.9	51.6	34.8	41.6
Thermal images + RGB	47.7	30.8	19.0	23.5	56.1	74.2	39.0	51.1

3.3 Discussion

Overall, the classical machine learning models provided more satisfactory results than the only deep learning model employed, YOLO11cls. This outcome can be explained by the relatively small number of available images, which is insufficient for training a deep neural network to achieve robust generalization.

Among the machine learning models, performance was generally comparable. In all cases, configurations without RGB features were preferred, both with and without statistical features. This suggests that the inclusion of RGB features may not yet contribute effectively to model performance, and that further investigation is needed to identify which RGB-derived features could provide meaningful information rather than introducing noise and increasing model confusion. This indicates that, for a future practical application, manually inserted features by the operator combined with features extracted solely from thermal images may be sufficient, reducing computational cost while maintaining reliable performance.

Of all the models tested, logistic regression achieved the best performance, reaching 66% balanced accuracy and 59% F1-score on the test set, while maintaining a high recall of 74.2%. This strong recall is particularly important in a clinical screening context, as it ensures that most positive cases are correctly identified. The effectiveness of logistic regression in this scenario may be partially attributed to the simplicity of the dataset and the presence of features that are approximately linearly separable, which allows the model to capture meaningful patterns despite the limited number of samples. Nevertheless, the choice of the optimal technique should be reevaluated in the future once a larger dataset becomes available.

For the purposes of this thesis, model outputs were expressed as binary decisions to enable the calculation of standard performance metrics. However, in a real clinical scenario, providing the clinician with a probability of infection rather than a strict yes/no output would be far more informative. A probabilistic prediction does not simply classify a wound as infected or non-infected; it conveys the model's degree of confidence relative to the ground truth, offering a quantifiable measure of uncertainty. This information is particularly valuable in borderline or ambiguous cases, where both the algorithm and the clinician may reasonably question the diagnosis.

Such an approach promotes a more synergistic interaction between human expertise and artificial intelligence. Instead of replacing the clinician's judgment, the system would serve as a decision-support tool, highlighting cases that warrant closer attention and strengthening confidence when model and clinical impressions align. By presenting predictions as probabilities, potential conflicts between the device and the medical professional are minimized, as the clinician remains the ultimate decision-maker, interpreting the AI output within the broader clinical context.

From a practical standpoint, probabilistic outputs could also help reduce the number

of biopsies performed. When both the device and clinician indicate a clearly high or low likelihood of infection, invasive procedures may be avoided, reserving biopsies for truly uncertain cases or for persistent infections requiring an antibiogram to identify the causative bacteria. This selective strategy would benefit both patients and the healthcare system. While biopsies are often covered by public healthcare, in regions such as Lombardy, where the clinical trial was conducted, they may be charged to the patient in the absence of exemptions, adding financial strain to individuals already coping with chronic and painful conditions.

Furthermore, biopsies require long processing times, during which no intervention can be performed, potentially allowing the wound to worsen in case of infection. Providing an earlier probabilistic assessment would allow timely decision-making, even in home settings or facilities where a doctor is not immediately available, including for patients with limited mobility. Thus, a rapid, probability-based decision support system could improve both clinical outcomes and patient experience while optimizing resource allocation within the healthcare system.

Future Developments

The clinical trial is still ongoing, and consequently, one of the primary objectives is to collect as many images as possible to improve model performance. An expanded dataset would not only increase the statistical power of the analyses but also introduce greater variability in patient conditions, wound types, and acquisition settings, which is essential for developing models that generalize well across different clinical scenarios. In the current study, all analyses were performed using gradient-based features, as the thermal camera did not provide direct temperature values for each pixel.

With a larger dataset, it will be possible to apply AI explainability techniques, which are crucial for clinical applications where model decisions must be interpretable and trustworthy. Feature importance analysis can be conducted to identify which features contribute most significantly to the model's decisions. These insights can then be discussed with one or more medical experts, providing clinical validation and ensuring that the model's reasoning aligns with domain knowledge.

Another important future development is the collection of data that includes direct temperature values for each pixel. Access to precise temperature measurements would enable the exploration of even simpler and more interpretable methods, such as rule-based or threshold-based algorithms, provided that high performance can still be maintained. This approach could lead to models that are not only accurate but also highly transparent, facilitating their acceptance in clinical practice and reducing computational demands compared to more complex models.

Furthermore, there is considerable interest in exploring deep neural network (DNN) architectures for this task. However, this approach requires substantially larger datasets to avoid overfitting and to fully exploit the capabilities of such models. Future work could investigate convolutional neural networks or hybrid architectures that integrate both thermal and RGB information, potentially improving feature extraction and capturing more complex patterns. The combination of larger datasets, explainable features, and advanced deep learning models has the potential to significantly enhance predictive performance while maintaining interpretability.

Finally, these developments could have a direct impact on clinical workflows. By providing clinicians with reliable probabilistic assessments of wound infection, these AI systems could reduce the number of unnecessary biopsies, support timely intervention,

and allow monitoring in remote or home-based settings.

Overall, future work aims to develop AI models that are accurate, interpretable, and clinically actionable, contributing to both improved patient care and more efficient healthcare resource utilization.

Conclusions

This study investigated the use of machine learning and deep learning approaches for the analysis of wound images to support infection detection. Overall, classical machine learning models consistently outperformed the deep learning approach, which was constrained by the limited size of the dataset. Across different models, configurations relying on features derived from thermal images and operator-provided annotations were sufficient to achieve reliable and stable performance, while the inclusion of RGB features or additional statistical descriptors did not consistently improve results.

Among the models tested, even relatively simple algorithms, such as logistic regression, demonstrated satisfactory performance, highlighting that interpretable and well-curated features can provide meaningful predictions, especially when the available dataset is small. A general observation is that, despite differences in underlying algorithms, most machine learning models provided comparable performance, with averaged cross-validation folds offering a robust estimate of generalization. In contrast, deep learning methods, such as YOLO11cls, did not perform effectively due to the small number of available images, illustrating the importance of dataset size for complex models.

From a practical perspective, the findings of this work support the development of AI-based tools that can assist clinicians in evaluating wounds and estimating the likelihood of infection. Such systems have the potential to reduce unnecessary biopsies, provide faster assessments, and guide treatment decisions, particularly in settings where direct access to specialized medical personnel is limited or patients have reduced mobility. The approach adopted in this study, focusing on interpretable and well-curated features, ensures that models remain computationally efficient and applicable in real-world clinical environments.

Limitations of the current study include the relatively small dataset and the absence of direct temperature measurements for each pixel, which restricted the types of features that could be extracted and analyzed. Consequently, future work will focus on collecting larger and more diverse datasets, including direct thermal measurements, and exploring the integration of more advanced machine learning and deep learning architectures. Combining these developments with explainability techniques will facilitate the creation of models that are both accurate and interpretable, ultimately supporting safer and more

effective clinical decision-making.

In summary, this work demonstrates the feasibility of using data-driven approaches to assist wound assessment, highlights the importance of carefully selected features, and lays the groundwork for future research aimed at developing clinically relevant, robust, and efficient AI tools for infection detection.

Bibliography

[1] Steven Bowers and Eginia Franco. Chronic wounds: Evaluation and management. *Am Fam Physician*, 101(3):159–166, February 2020.

[2] Vincent Falanga, Roslyn Rivkah Isseroff, Athena M Soulika, Marco Romanelli, David Margolis, Suzanne Kapp, Mark Granick, and Keith Harding. Chronic wounds. *Nature Reviews Disease Primers*, 8(1):50, July 2022.

[3] Chandan K. Sen. Human wounds and its burden: An updated compendium of estimates. *Advances in Wound Care*, 8(2):39–48, 2019. PMID: 30809421.

[4] Chandan K Sen. Human wound and its burden: Updated 2020 compendium of estimates. *Adv Wound Care (New Rochelle)*, 10(5):281–292, May 2021.

[5] Zena Moore and Sebastian Probst. Building the business case for shared wound care: a cost-benefit case for service providers. *Wounds Int*, 13(4):62–66, 2022.

[6] J. F. Guest, K. Vowden, and P. Vowden. The health economic burden that acute and chronic wounds impose on an average clinical commissioning group/health board in the uk. *Journal of Wound Care*, 26(6):292–303, 2017. PMID: 28598761.

[7] Rosana E Norman, Michelle Gibb, Anthony Dyer, Jennifer Prentice, Stephen Yelland, Qinglu Cheng, Peter A Lazzarini, Keryln Carville, Karen Innes-Walker, Kathleen Finlayson, Helen Edwards, Edward Burn, and Nicholas Graves. Improved wound management at lower cost: a sensible goal for australia. *International Wound Journal*, 13(3):303–316, 2016.

[8] Elia Ricci and Monica Pittarello. *Vulnology (Also Known as Wound Care): History and Myths of Chronic Wounds*, pages 3–9. Springer International Publishing, Cham, 2023.

[9] Chukwuemeka N. Etufugh and Tania J. Phillips. Venous ulcers. *Clinics in Dermatology*, 25(1):121–130, 2007.

[10] Luciana P. F. Abbade, Sidnei Lastória, and Hamilton de Almeida Rollo. Venous ulcer: clinical characteristics and risk factors. *International Journal of Dermatology*, 50(4):405–411, 2011.

[11] Gregory Ralph Weir, Hiske Smart, Jacobus van Marle, and Frans Johannes Cronje. Arterial disease ulcers, part 1: Clinical diagnosis and investigation. *Advances in Skin & Wound Care*, 27(9), 2014.

[12] Worldwide trends in diabetes prevalence and treatment from 1990 to 2022: a pooled analysis of 1108 population-representative studies with 141 million participants. *The Lancet*, 404(10467):2077–2093, November 2024.

[13] Rie Roselyne Yotsu, Ngoc Minh Pham, Makoto Oe, Takeshi Nagase, Hiromi Sanada, Hisao Hara, Shoji Fukuda, Junko Fujitani, Ritsuko Yamamoto-Honda, Hiroshi Kajio, Mitsuhiko Noda, and Takeshi Tamaki. Comparison of characteristics and healing course of diabetic foot ulcers by etiological classification: Neuropathic, ischemic, and neuro-ischemic type. *Journal of Diabetes and its Complications*, 28(4):528–535, 2014.

[14] Joshua S. Mervis and Tania J. Phillips. Pressure ulcers: Pathophysiology, epidemiology, risk factors, and presentation. *Journal of the American Academy of Dermatology*, 81(4):881–890, 2019.

[15] Nicholas Graves and Henry Zheng. *Wound Practice Research: Journal of the Australian Wound Management Association*, 22(1):4–12, 14–19, 2014.

[16] Kirsi Isoherranen, Julie Jordan O'Brien, Judith Barker, Joachim Dissemond, Jürg Hafner, Gregor B. E. Jemec, Jivko Kamarachev, Severin Läuchli, Elena Conde Montero, Stephan Nobbe, Cord Sunderkötter, and Mar Llamas Velasco. Atypical wounds. best clinical practice and challenges. *Journal of Wound Care*, 28(Sup6):S1–S92, 2019. PMID: 31169055.

[17] Agata Janowska, Valentina Dini, Teresa Oranges, Michela Iannone, Barbara Loggini, and Marco Romanelli and. Atypical ulcers: Diagnosis and management. *Clinical Interventions in Aging*, 14:2137–2143, 2019.

[18] Elizabeth Nichols. Describing a wound: from presentation to healing. *Wound Essentials*, 10(1):56–61, 2015.

[19] Kathryn Vowden and Peter Vowden. Understanding exudate management and the role of exudate in the healing process. *British journal of community nursing*, 8(Sup5):S4–S13, 2003.

[20] Vincent Palanga. Classifications for wound bed preparation and stimulation of chronic wounds. *Wound Repair & Regeneration*, 8(5), 2000.

[21] Gregory S Schultz, David J Barillo, David W Mozingo, and Gloria A Chin. Wound bed preparation and a brief history of time. *International Wound Journal*, 1(1):19–32, 2004.

[22] Xuan Wang, Chong-Xi Yuan, Bin Xu, and Zhi Yu. Diabetic foot ulcers: Classification, risk factors and management. *World J Diabetes*, 13(12):1049–1065, December 2022.

[23] P. G. Bowler, B. I. Duerden, and D. G. Armstrong. Wound microbiology and associated approaches to wound management. *Clinical Microbiology Reviews*, 14(2):244–269, 2001.

[24] Abdul R. Siddiqui and Jack M. Bernstein. Chronic wound infection: Facts and controversies. *Clinics in Dermatology*, 28(5):519–526, 2010. Controversies in Dermatology: Part III.

[25] Jenny Hurlow and Philip G. Bowler. Acute and chronic wound infections: microbiological, immunological, clinical and therapeutic distinctions. *Journal of Wound Care*, 31(5):436–447, 2022.

[26] D.C. Bury et al. Osteomyelitis: Diagnosis and treatment. *American Family Physician*, 104:395–400, 2021.

[27] Z. Roje et al. Necrotizing fasciitis: literature review of contemporary management. *World Journal of Emergency Surgery*, 6:46, 2011.

[28] D.C. Angus et al. Severe sepsis and septic shock. *New England Journal of Medicine*, 369:840–851, 2013.

[29] Shuxin Li, Paul Renick, Jon Senkowsky, Ashwin Nair, and Liping Tang. Diagnostics for wound infections. *Advances in Wound Care*, 10(6):317–327, 2021. PMID: 32496977.

[30] Douglas Queen and Keith Gordon Harding. Importance of imaging to wound care practice. *Int Wound J*, 20(2):235–237, February 2023.

[31] Diane K. Langemo and James G. Spahn. A reliability study using a long-wave infrared thermography device to identify relative tissue temperature variations of the body surface and underlying tissue. *Advances in Skin & Wound Care*, 30(3), 2017.

[32] Mercè Iruela Sánchez, Rosa García-Sierra, Rafael Medrano-Jiménez, Diana Bonachela-Mompart, Natalia Maella-Rius, Esther Soria-Martín, Mar Isnard-Blanchar, and Pere Torán-Monserrat. Use of infrared thermometry to observe temperature variation associated with the healing process in wounds and ulcers: Tihuap cohort study protocol. *Healthcare*, 11(12), 2023.

[33] K F Cutting and K G Harding. Criteria for identifying wound infection. *J Wound Care*, 3(4):198–201, June 1994.

[34] Abbas K. Abbas, Konrad Heimann, Katrin Jergus, Thorsten Orlikowsky, and Steffen Leonhardt. Neonatal non-contact respiratory monitoring based on real-time infrared thermography. *Biomedical engineering online*, 10:93, 10 2011.

[35] Richard Ribón Fletcher, Gabriel Schneider, Laban Bikorimana, Gilbert Rukundo, Anne Niyigena, Elizabeth Miranda, Robert Riviello, Fredrick Kateera, and Bethany Hedt-Gauthier. The use of mobile thermal imaging and deep learning for prediction of surgical site infection. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, pages 5059–5062, 2021.

[36] Adam R. Collins, Gerard M. O'Connor, Darragh A. Ryan, Molly Parmeter, Sean Dinneen, and Georgina Gethin. Wound bed temperature has potential to indicate infection status: A cross-sectional study. *Wound Repair and Regeneration*, 33(4):e70072, 2025. e70072 WRR-25-04-0223.R1.

[37] Omnidermal. Wound viewer. `https://www.omnidermal.it/wound-viewer/`.

[38] Flir. Flir lepton 2.5. `https://oem.flir.com/it-it/products/lepton/?model=500-0763-01&segment=oem&vertical=microcam`.

[39] Jacopo Secco, Elisabetta Spinazzola, Monica Pittarello, Elia Ricci, and Fabio Pareschi. Clinically validated classification of chronic wounds method with memristor-based cellular neural network. *Scientific Reports*, 14(1):30839, December 2024.

[40] Rosanna Cavazzana, Angelo Faccia, Aurora Cavallaro, Marco Giuranno, Sara Becchi, Chiara Innocente, Giorgia Marullo, Elia Ricci, Jacopo Secco, Enrico Vezzetti, and Luca Ulrich. Enhancing clinical assessment of skin ulcers with automated and objective convolutional neural network-based segmentation and 3d analysis. *Applied Sciences*, 15(2), 2025.

[41] Elisabetta Spinazzola, Guillaume Picaud, Sara Becchi, Monica Pittarello, Elia Ricci, Marc Chaumont, Gérard Subsol, Fabio Pareschi, Luc Teot, and Jacopo Secco. Chronic ulcers healing prediction through machine learning approaches:

Preliminary results on diabetic foot ulcers case study. *Journal of Clinical Medicine*, 14(9), 2025.

[42] Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biol.*, 52(1-2):99–115, January 1990.

[43] Baoyu Liang, Yuchen Wang, and Chao Tong. Ai reasoning in deep learning era: From symbolic ai to neural–symbolic ai. *Mathematics*, 13(11), 2025.

[44] Chunwei Tian, Tongtong Cheng, Zhe Peng, Wangmeng Zuo, Yonglin Tian, Qingfu Zhang, Fei-Yue Wang, and David Zhang. A survey on deep learning fundamentals. *Artif. Intell. Rev.*, 58(12), October 2025.

[45] Emmanuel Chris, Anita Johnson, and Grace Phonix. Deep learning vs. traditional machine learning: Key differences. 11 2024.

[46] Andreas Holzinger, Georg Langs, Helmut Denk, Kurt Zatloukal, and Heimo Müller. Causability and explainability of artificial intelligence in medicine. *WIREs Data Mining and Knowledge Discovery*, 9(4):e1312.

[47] Park Hyeoun-Ae. An introduction to logistic regression: From basic concepts to interpretation with particular attention to nursing domain. *jkan*, 43(2):154–164, 2013.

[48] Bo Sun and Haiyan Chen. A survey of k nearest neighbor algorithms for solving the class imbalanced problem. *Wirel. Commun. Mob. Comput.*, 2021:5520990:1–5520990:12, 2021.

[49] Richard G Brereton and Gavin R Lloyd. Support vector machines for classification and regression. *Analyst*, 135(2):230–267, February 2010.

[50] Gérard Biau. Analysis of a random forests model. *J. Mach. Learn. Res.*, 13(1):1063–1095, April 2012.

[51] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA, 2016. Association for Computing Machinery.

[52] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. Catboost: unbiased boosting with categorical features. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[53] Ultralytics. YOLO11. https://docs.ultralytics.com/it/models/yolo11/.

[54] Machine Learning Mastery. k-fold cross validation. https://machinelearningmastery.com/k-fold-cross-validation/.

[55] Tom F Sterkenburg. Statistical learning theory and occam's razor: the core argument. *Minds and Machines*, 35(1):3, 2024.