Alma Mater Studiorum · University of Bologna

School of Science
Department of Physics and Astronomy
Master Degree in Physics

An Innovative Machine Learning Approach to Wound Edge Assessment using Depth Maps and Wound Border Rectification

Supervisor:
Dr. Nico Curti

Co-supervisor:
Dr. Alessandra Merlotti

Submitted by:
Michele Fruci

Academic Year 2023/2024

"Men at some time are masters of their fates. The fault, dear Brutus, is not in our stars but in ourselves, that we are underlings."

W. Shakespeare

"Non è nelle stelle che è conservato il nostro destino, ma in noi stessi.
Uomini forti, destini forti.
Uomini deboli, destini deboli."

L. Spalletti

Abstract

Chronic wounds represent a widespread and costly healthcare crisis, undermining patients' quality of life and challenging clinicians with subjective and inconsistent assessment methods. This thesis aims to enhance wound edge assessment by introducing a innovative method based both on advanced Artificial Intelligence models and cutting-edge image processing techniques. A unique feature of the method is the integration of wound depth maps with a novel wound edge rectification technique.

A dataset of 159 annotated wound images, provided by IRCCS Sant'Orsola-Malpighi University Hospital in Bologna, was analyzed. Depth maps were used to capture the three-dimensional structural details of wound edges. The wound border rectification method was applied to standardize the periwound area geometries, enabling consistent depth feature extraction for subsequent analysis.

Clustering algorithms identified natural groupings of wound edge features in agreement with 5 labels of BWAT scoring. Moreover, it identified 8 clusters providing insights into underlying structural patterns to be investigated for a finer wound edge stratification. Additionally, a Random Forest model based on extracted depth features was employed. The model encountered similar challenges as clinicians in differentiating between ambiguous wound edge types; however, by merging these categories, it achieved a strong accordance with clinical assessments (ACC$= 0.72 \pm 0.17$).

The results validate the potential of Machine Learning approaches to improve diagnostic accuracy and assist clinicians in wound management. Moreover, the clustering analysis highlights new avenues for exploring the micro-structural organization of wound edges, paving the way for more detailed investigations.

Finally, the proposed method offers a fully automated solution for image characterization of wound edges, aiming to overcome the limitations of traditional manual assessments and to support clinicians with objective, consistent, and reproducible insights. This work lays the foundation for future advancements in automated wound care systems, with the goal of optimizing patient outcomes and healthcare resource utilization.

Contents

Introduction 1

1 Wound Healing Assessment 3
- 1.1 Assessment and Documentation of Wound status . . . 4
  - 1.1.1 Key parameters for wound assessment . . . 4
- 1.2 BWAT - Bates-Jensen Wound Assessment Tool . . . 10
  - 1.2.1 BWAT scoring system . . . 10
- 1.3 PWAT - Photographic Wound Assessment Tool . . . 13
  - 1.3.1 PWAT scoring system . . . 13

2 Machine Learning Algorithms for Wound Management 15
- 2.1 Linear Discriminant Analysis - LDA . . . 16
- 2.2 Random Forest - RF . . . 18
- 2.3 Clustering Algorithms . . . 19
  - 2.3.1 Distance and similarity functions . . . 20
  - 2.3.2 Types of clustering . . . 21
  - 2.3.3 Dimensionality Reduction for Clustering . . . 23
- 2.4 Artificial Neural Networks - ANN . . . 24
  - 2.4.1 Perceptron . . . 24
  - 2.4.2 Feedforward Neural Networks - FNN . . . 25
  - 2.4.3 Deep Neural Networks - DNN . . . 26
  - 2.4.4 CNN for Medical Images Analysis . . . 32
- 2.5 A Review of Machine Learning for Wound Care . . . 34

3 Materials and Methods 37
- 3.1 Description of Dataset . . . 37
- 3.2 Image Segmentation . . . 38
  - 3.2.1 Deepskin . . . 38
  - 3.2.2 Selection of the Largest Mask Component . . . 39
- 3.3 Depth Maps Computation . . . 39
  - 3.3.1 ZoeDepth . . . 39
- 3.4 Wound Border Rectification . . . 40
- 3.5 Depth Map Correction . . . 41
- 3.6 Mean Depth Profile . . . 43
- 3.7 Features extraction . . . 44

4 Results **47**
4.1 Clinician Agreement in Edges Classification 47
4.2 Wound Edge Clustering 51
4.2.1 Features embedding 51
4.2.2 Cluster Identification 52
4.2.3 Cluster Composition 56
4.3 Linear Discriminant Analysis for Wound Edge Classification 59
4.4 Random Forest for Wound Edge Classification 62

5 Discussions **69**
5.1 Interpretation of the results 69
5.1.1 Clinician Agreement 69
5.1.2 Clustering Results 69
5.1.3 Results of Linear Discriminant Analysis 70
5.1.4 Classification Results 70
5.2 Limitations of the Research 70
5.3 Future Developments 71

Conclusions **73**

List of Figures

1.1 Body diagram of anatomic locations and bony prominences. [7] 6
1.2 The figure shows a practical visual example of how measure the maximum dimensions of the wound along two perpendicular directions. 7
1.3 Illustration of tunneling in a wound, where the wound extends underneath the surface, creating a tunnel-like structure. 8
1.4 Example of undermining in a wound, where tissue loss occurs underneath the wound edges, creating a larger surface area 8

2.1 Example of Linear Discriminant Analysis (LDA) applied to a two-class dataset, showing the separation of classes in the transformed feature space. 17
2.2 Example of a Random Forest model: tree are built with a subset of random features. Outputs of trees are combined to generate the finale results. . . 19
2.3 Overview of the key steps in clustering process. 20
2.4 Visual comparison of different clustering algorithms, illustrating the variation in results depending on the model. 21
2.5 Comparison between the application of t-SNE and PaCMAP for dimensionality reduction of the Mammoth dataset [33]. 24
2.6 Architecture of Perceptron 25
2.7 Architecture of a Feedforward Neural Network. [34] 26
2.8 Example of architecture of a Deep Neural Network (specifically a Multi-Layer Perceptron) [34] 26
2.9 Illustration highlighting deep learning ability to automatically learn hierarchical features from raw data, in contrast to traditional machine learning models. Adapted from [35]. 27
2.10 Scheme of backpropagation algorithm. [36] 30
2.11 Skip connection simple scheme. 31
2.12 For enhanced visual representation, it is common practice to depict convolutional layers arranged in a two-dimensional grid. Each neuron in the hidden layer is only connected to the correspondent local receptive field. . 32
2.13 Basic architecture of a CNN [42]. 33
2.14 Original U-Net architecture for input image 572×572 34

3.1 Example of Deepskin segmentation. On the left, the original image; on the right, the segmented ROIs: wound (red), patient body (green), and background (blue). 38
3.2 Example of multi-components wound. On the left is the original image, in the middle is the wound mask extracted by Deepskin, and on the right is the largest component, which is used as the mask for the analysis. 39

3.3 On the left is the original image, and on the right is the depth map computed by the ZOED_NK model, presented using the a colormap. It can be observed that the wound depth is accurately represented. 40
3.4 Conceptual idea of border rectification: identification of the wound border (green); extraction of the periwound region as a ”strip” to be ”cut”; unrolling of the strip. 41
3.5 (a) Original image; (b) Depth map computed by ZoeDepth; (c) Quadratic fit; (d) Difference between the fit and the ZoeDepth map. 42
3.6 3D representation of the wound data: (top left) original wound image; (top right) ZoeDepth map; (bottom left) quadratic fit; (bottom right) difference between the quadratic fit and the ZoeDepth map. 43
3.7 Mean depth profiles with standard deviation (STD) for the three different trends. 44
3.8 On the left, a rectified depth border. In the middle, the original depth profiles. On the right, the profiles the baselining procedure and high-frequency filtering 44

4.1 Normalized co-occurrence matrix showing the alignment between DERM and DERM-1M classifications. 48
4.2 Normalized co-occurrence matrix showing the alignment between DERM and CO classifications. 48
4.3 Normalized co-occurrence matrix showing the alignment between CO and DERM-1M classifications. 49
4.4 Normalized co-occurrence matrix showing the alignment between DERM and DERM-1M classifications merging labels. 50
4.5 Normalized co-occurrence matrix showing the alignment between DERM and CO classifications merging labels. 50
4.6 Normalized co-occurrence matrix showing the alignment between CO and DERM-1M classifications merging labels. 51
4.7 PaCMAP embedding of depth profiles features into a two-dimensional space. 52
4.8 Visualization of the clusters identified by HDBSCAN in the embedding space. 53
4.9 Visualizations of 30 individual feature distributions within the 2D embedded space. The color variation represents the change in the feature value across the dataset. 54
4.10 Visualizations of 25 individual feature distributions within the 2D embedded space. The color variation represents the change in the feature value across the dataset. 55
4.11 DERM labeled data in the embedding space distinguished by color. 56
4.12 Cluster composition matrix that illustrate how DERM classified data are distributed across the clusters. The matrix is normalized across the rows. . 56
4.13 DERM-1M labeled data in the embedding space distinguished by color. . . 57
4.14 Cluster composition matrix that illustrate how DERM-1M classified data are distributed across the clusters. The matrix is normalized across the rows. 57
4.15 CO labeled data in the embedding space distinguished by color. 58
4.16 Cluster composition matrix that illustrate how CO classified data are distributed across the clusters. The matrix is normalized across the rows. . . 58

4.17 On the left, LDA scatterplot in the transformed space for DERM classification. On the right, plot of KDE overlaid to visualize the density of points across the space. . . . 59
4.18 On the left, LDA scatterplot in the transformed space for DERM-1M classification. On the right, plot of KDE overlaid to visualize the density of points across the space. . . . 60
4.19 On the left, LDA scatterplot in the transformed space for CO classification. On the right, plot of KDE overlaid to visualize the density of points across the space. . . . 60
4.20 Normalized dissimilarity matrix for DERM classification. The colored squares highlight the distances between points with the same label. . . . 61
4.21 Normalized dissimilarity matrix for DERM-1M classification. The colored squares highlight the distances between points with the same label. . . . 61
4.22 Normalized dissimilarity matrix for CO classification. The colored squares highlight the distances between points with the same label. . . . 62
4.23 Confusion matrices for the ensemble of binary classifiers for DERM dataset, showing both Accuracy (ACC) and Matthews Correlation Coefficient (MCC) for each class pair. . . . 64
4.24 Confusion matrices for the ensemble of binary classifiers for DERM-1M dataset, showing both Accuracy (ACC) and Matthews Correlation Coefficient (MCC) for each class pair. . . . 64
4.25 Confusion matrices for the ensemble of binary classifiers for CO dataset, showing both Accuracy (ACC) and Matthews Correlation Coefficient (MCC) for each class pair. . . . 65
4.26 Confusion matrix of the model for predicting the original five labels, using only the 53 images where all clinicians agree. . . . 66
4.27 Confusion matrix of the model for predicting the grouped labels, using the 79 images where all clinicians agree. . . . 66
4.28 Wound with edges classified as Well-defined by clinicians but misclassified as Distinct by the model, likely because the wound bed is filled with slough tissue, which alters the depth change perceived by the model. . . . 67
4.29 Wound image with edges classified as Indistinct by clinicians but misclassified as Distinct by the model, likely because of the scarring around the wound, which alters the depth change measured by the model. . . . 67
4.30 Wound image with edges classified as Fibrotic by clinicians but misclassified as Rolled under by the model. . . . 68
4.31 Wound image with edges classified as Well-defined by clinicians but misclassified as Rolled under by the model, likely due to jagged edges that hinder an accurate depth evaluation within the wound border. . . . 68

List of Tables

1.1 Table displaying the parameters of the BWAT and their corresponding classification values for wound assessment.[7] 12
1.2 PWAT parameters and their scoring system [20]. 14

3.1 Summary of wound edges classification. **DERM** and **DERM-1M** refer to the first and one-month post-classification by the dermatologist. **CO** refers to the classification made by the clinical operator. 38
3.2 Summary of the extracted features from the depth profiles. 45

4.1 Summary of concordance and ICC for each classification comparison. . . . 47
4.2 Summary of concordance and ICC for each classification comparison merging the Distinct to the Indistinct labels and the Rolled under to the Fibrotic labels 49
4.3 Count of points across the identified clusters. Cluster -1 corresponds to the noise points. 53
4.4 Label mapping across different class structures analyzed. 63
4.5 Accuracy(ACC) and Matthews Correlation Coefficient (MCC) for different classification structures across three datasets DERM, DERM-1M, and CO. 63

Introduction

Chronic wounds represent a silent yet pervasive crisis in clinical medicine, affecting millions of people worldwide. These wounds, which result from a variety of underlying conditions such as diabetes, vascular diseases, or prolonged pressure, deviate from the normal healing trajectory, persisting for months or even years. The high prevalence of these wounds and the associated physical complications such as pain, as well as the severe psychological and social burden, demonstrate the substantial impact they have on patients.

Furthermore, the management of chronic wounds places a considerable financial strain on healthcare systems worldwide, accounting for roughly 3% of total healthcare expenditure. A significant portion of these costs is related to professional care, which includes the time spent by wound care operators, as well as hospitalizations [1].

Despite the importance of precise wound management, traditional wound assessment methods are predominantly subjective, relying heavily on clinician expertise and visual inspection. Indeed, it has been estimated that variability in wound assessment can reach 30% also even when experienced clinicians are involved[2].

The accurate characterization of wound characteristics, particularly the assessment of wound edges and tissue type, poses a significant challenge due to the potential for short-term variations in wound appearance and the influence of environmental factors. These limitations underscore the necessity for objective and reproducible methods to enhance wound assessment, improve patient outcomes, and optimize healthcare resource allocation.

This thesis aims to explore and validate an innovative machine learning-based approach to wound edge assessment and classification, aiming to reduce the subjectivity inherent in traditional clinical evaluations.

The analyzed dataset of images was provided by the IRCCS Sant'Orsola-Malpighi University Hospital in Bologna and was acquired during clinical practice.

Central to the proposed approach is the use of depth maps, i.e. images which represent the distance between the surfaces of the illustrated objects and a specific viewpoint, allowing for the capture of the three-dimensional structure of wound edges.

A distinctive feature of the proposed method is the "wound border rectification", i.e. the computational transformation of the wound border into a standardized rectangular format, regardless of its original shape. This approach ensures uniformity in the analysis across wounds of varying geometries, enhancing consistency and improving diagnostic precision.

Chapter 1 will present a comprehensive account of the wound problem and the fundamental aspects of wound healing and assessment. It will also discuss available tools, such as the Bates-Jensen Wound Assessment Tool (BWAT), for evaluating wound status and monitoring healing progresses. Chapter 2 provides an overview of machine learning algorithms, including Clustering, Linear Discriminant Analysis (LDA), Random Forest

models, and Artificial Neural Networks (ANN). Particular emphasis is placed on their application in medical image analysis, with a focus in the context wound healing estimation. Chapter 3 presents a comprehensive account of the materials and methods employed in this study. This chapter includes a description of the analyzed dataset, the image segmentation techniques used, the extraction of depth maps from wound images and the computational workflow for wound border rectification and feature extraction. Chapter 4 presents the results, focusing on clustering patterns observed in wound edge classification and the performance of machine learning models in comparison to clinician assessments. Chapter 5 discusses the results interpretation, highlighting their potential for implementation in clinical practice, and outlining future research directions. Finally, the Conclusions summarize the findings of the study.

Chapter 1

Wound Healing Assessment

A **wound** is defined as a disruption of the skin and underlying tissue, which may result from various factors including surgical procedures, trauma, pressure, or underlying medical conditions such as diabetes and vascular disease [1].

Wounds are primarily classified into two categories: **acute wounds**, including surgical incisions and burns, and **chronic wounds**, which group together diabetic foot ulcers (DFUs), leg ulcers, and pressure ulcers [2].

Chronic wounds present a significant challenge as they deviate from the conventional healing process, potentially persisting for extended periods, even years. These wounds have a profound impact on patients, affecting them physically, mentally, and socially. This can result in a range of complications, including pain, anxiety, social isolation, and prolonged hospital stays. Chronic wounds have been described as a "***silent epidemic***" due to their pervasive yet frequently unacknowledged impact [3].

The management of chronic wounds is a significant financial challenge for healthcare systems around the world. In Europe, an estimated 1.5 to 2 million people are affected by chronic wounds, while in the United States the figure is around 6.5 million. The economic burden associated with these conditions is astonishing, with costs reaching $11 billion/year in the US and estimates of €4-6 billion/year in Europe. [4, 5, 6]

It is estimated that chronic wounds account for approximately 3% of total healthcare expenditure. A common misconception is that the costs associated with wound care are primarily driven by the costs of materials, such as dressings or bandages [1]. However, it is the **time** spent by healthcare professionals and hospital care that constitutes the majority of these costs, at 80-85%. Furthermore, complications such as infections or amputations can significantly escalate treatment costs and prolong healing times [1].

Given the complexity and high costs associated with chronic wound management, it is of paramount importance to accurately assess the healing process in order to ensure the optimal provision of care and the optimal allocation of resources. Unfortunately, the assessment healing methods are largely subjective and rely heavily on visual inspections conducted by healthcare professionals. Research indicates that even **experienced clinicians** demonstrate up to **30% variability** when assessing the same wound [2]. This inconsistency arises from the challenges associated with distinguishing between different tissue types, such as granulation tissue, fibrin, and necrosis, as well as difficulties in measuring wound depth and edges. Furthermore, factors such as edema and alterations in exudate can cause short-term variations in the appearance of wounds, complicating the assessment process and increasing the likelihood of misjudgment.

The introduction of wound assessment methods for objective and reproducible mea-

surement of wound healing has the potential to improve the accuracy of wound evaluations and to produce more consistent treatment plans. These advances could lead to a reduction in the variability of care, which in turn could lead to improved patient outcomes and a reduction in healthcare costs.

1.1 Assessment and Documentation of Wound status

A standardized approach for the assessment of wound status is critical for ensuring consistency in evaluations, optimizing treatment strategies, and effectively monitoring the healing process.

The *Wound, Ostomy and Continence Nurses Society* delineates comprehensive guidelines for wound assessment, specifying the key parameters that must be measured and documented. These include wound dimensions, depth, the presence of undermining or tunneling, and a thorough description of the wound base, edges, surrounding tissues and exudate [7].

The monitoring of the wound healing process is achieved through periodic and serial evaluations, which involve the use of scoring systems where each specific wound characteristic is assessed using a Likert scale, i.e. rating its severity with ordinal values [8].

A physical examination remains the cornerstone for the assessment of wound status. Nevertheless, novel approaches to wound evaluation via telemedicine, particularly photo-based wound analysis, have been developed and investigated. In both assessments, the need for well-trained personnel remains essential to minimize operator dependence and ensure accurate evaluations.

1.1.1 Key parameters for wound assessment

The accurate documentation of the following parameters, as outlined in clinical guidelines [7], is of critical importance, as all wound assessment tools are based on them. The use of these tools in clinical practice enhances communication between care providers and is strongly endorsed by many experts [9].

Wound history

A detailed wound history is vital for understanding its severity. This should include inquiries about the onset, duration, and previous treatments.

The duration of the wound is a crucial prognostic indicator. Wounds that persist for more than 12 months are significantly less likely to heal within a 12- to 20-week time frame[10, 11]. As the duration of the wound increases, the likelihood of effective healing decreases.

Etiology

Identify the wound etiology is essential for the development of effective treatment plans. Wounds can be classified into several categories based on their underlying etiology.
The following section lists the principal categories of wound etiology.

- **Arterial Ulcers:** the results of conditions such as atherosclerosis and they are most frequently observed in the lower extremities.

- **Venous Ulcers:** these are a consequence of venous insufficiency and are commonly observed in the lower leg, in the area surrounding the gaiter.
- **Diabetic Ulcers:** associated with diabetic neuropathy and peripheral arterial disease, these wounds often manifest on the feet.
- **Pressure Ulcers:** resulting from prolonged pressure on bony prominences.
- **Traumatic Wounds:** result from injuries and exhibit a wide range of appearances and depths.
- **Pyoderma Gangrenosum:** an inflammatory dermatological condition characterized by painful ulcerative lesions. It is often associated with systemic diseases.
- **Vasculitis ulcers:** inflammation of blood vessels which can progress to the formation of cutaneous ulcers.
- **Scleroderma:** autoimmune condition which can lead to thickening and ulceration of the skin.
- **Tumoral wounds:** certain malignancies can lead to the development of chronic wounds.
- **Lymphatic Disorders:** issues affecting lymphatic drainage have the potential to result in chronic wounds.
- **Radiation wounds:** radiation exposure may result in tissue damage leading to non-healing ulcers.

Location

The use of precise anatomical terminology to document wound locations is fundamental in clinical practice.

To guarantee accurate recording,the wound location is selected from a pre-defined list of anatomical sites or, alternatively, body diagrams are employed (example in Figure 1.1). The high level of specificity in documentation is required to allow consistent assessment of wound progression, avoiding any ambiguities on the status of the wound.

The anatomical location of a wound is frequently determined by its etiology, therefore this information can provide insights to facilitate the diagnosis and to establish treatment planning.

Shape

The wound shape is needed in estimating its overall dimensions and it is determined by evaluating the wound's perimeter.

Wound morphology often reflects processes occurring beneath the surface, such as wound contraction, whereby the wound assumes a more regular circular or oval shape as healing progresses.

Also the wound shape is frequently associated with specific etiologies. For instance, arterial ulcers usually exhibit a round appearance, while venous ulcers typically present with irregular or oval contours.

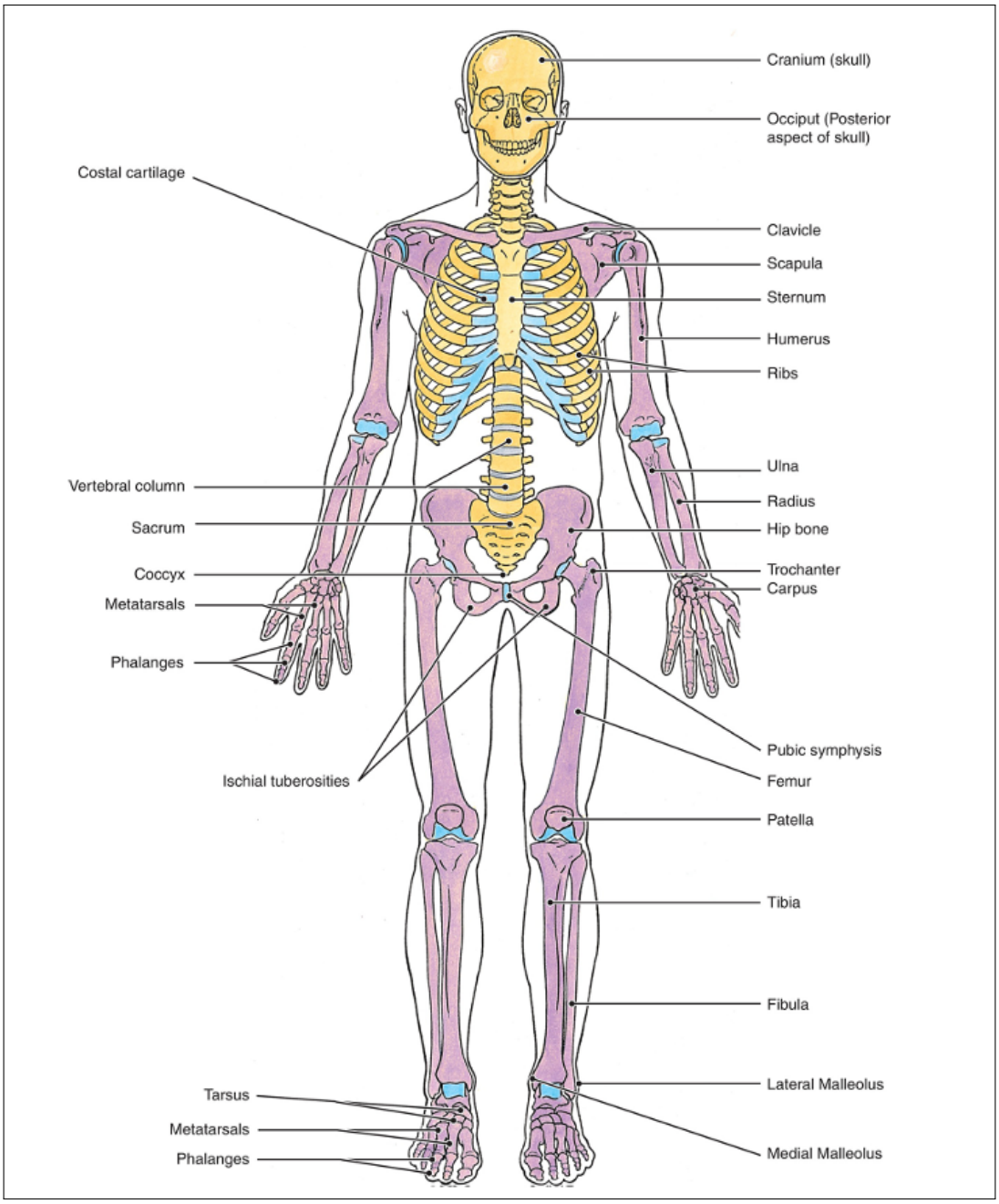

Figure 1.1: Body diagram of anatomic locations and bony prominences. [7]

Size

The evaluation of wound size is equivalent to calculate its surface area in cm^2.

The most straightforward and commonly employed approach is to measure the maximum dimensions of the wound along two perpendicular axes (usually vertical and horizontal) and then multiply them (Figure 1.2).

This method has been demonstrated and validated [12, 13] to show a high correlation with more complex planimetric techniques used for calculating wound area.

The wound area must be measured periodically because its percentage reduction during the first four weeks of clinical treatment is a good predictor of the healing process[14].

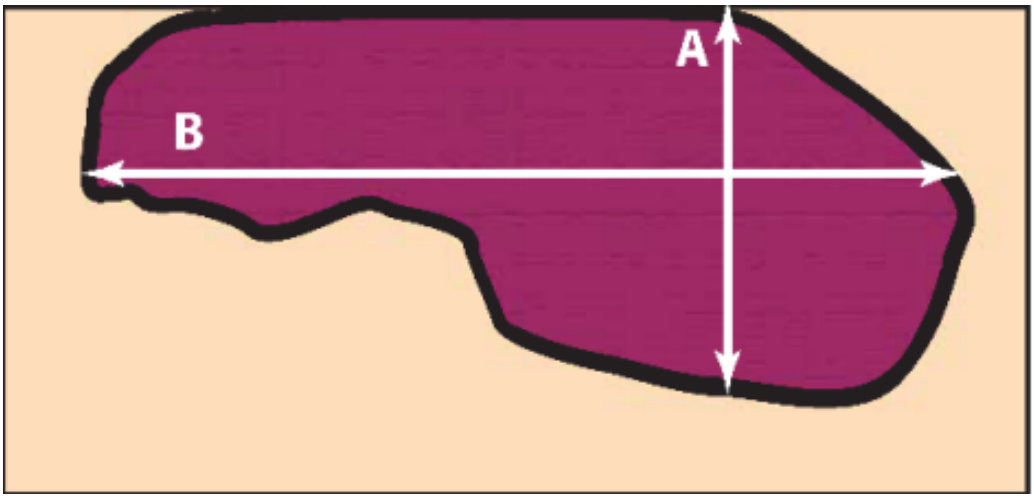

Figure 1.2: The figure shows a practical visual example of how measure the maximum dimensions of the wound along two perpendicular directions.

Depth

The wound depth is measured by inserting a cotton-tipped applicator vertically into the deepest part of the wound and marking it with a pen at the plane of the skin. The mark is then compared with a measuring device (a ruler).

Edges

The characterization of wounds edges is a pivotal yet challenging aspect of wound assessment. In order to obtain a correct evaluation and valuable insight into the healing process, it is essential to employ both observation and palpation techniques.

The initial step is to ascertain whether the edges are clear and distinct, or indistinct and diffuse, i.e. if there is or not a sharp and distinguishable demarcation line between the wound and the surrounding healthy tissues.

Subsequently, it must be established if they are attached to the surrounding skin and the wound base, or they are unattached, which means the presence of crater or bowl-shaped wounds.

In addition, the assessment must ascertain whether the wound edges are open and proliferative or closed and rolled under. In the latter case, a thickening and hardening of the edges may occur, possibly degenerating into hyperkeratosis. This may impede the migration of epithelial cells and delay the healing process.

Tunneling/Undermining

Tunneling and undermining, or pocketing, represent the loss of tissue underneath an intact skin surface [15].

Tunneling is caused by the destruction of tissues in the wound bed, which results in a narrow tract of dead space that may develop into an abscess. To measure tunneling, a cotton-tipped applicator is inserted into the passageway until resistance is felt (Figure 1.3). The applicator is marked at the skin plane and then compared with a measuring device to obtain the depth of the tunnel.

Undermining is caused by erosion under the wound edges, which results in a large wound with a small opening (*"iceberg effect"*). Undermining is measured directly under

the wound edge, with the applicator held almost parallel to the wound surface, stopping when resistance is felt (Figure 1.4). The applicator is marked at the wound edge to measure the undermining extension.

The use of clock terms is recommended for an accurate description of the location of tunneling and undermining. Tunneling generally includes a narrower area of tissue than undermining and typically occurs in one direction, whereas undermining may occur in one or more directions. The occurrence of undermining is most prevalent in patients with pressure wounds, whereas tunneling is most commonly observed in surgical wounds.

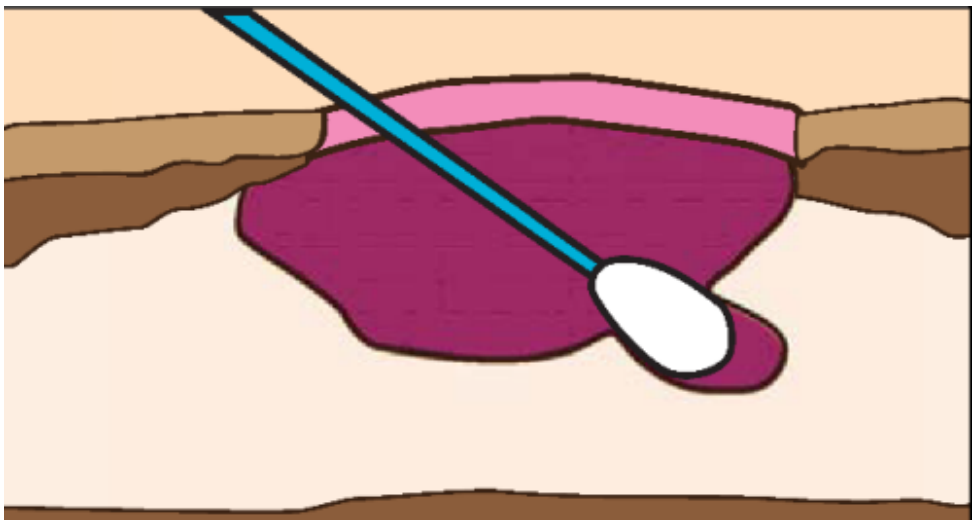

Figure 1.3: Illustration of tunneling in a wound, where the wound extends underneath the surface, creating a tunnel-like structure.

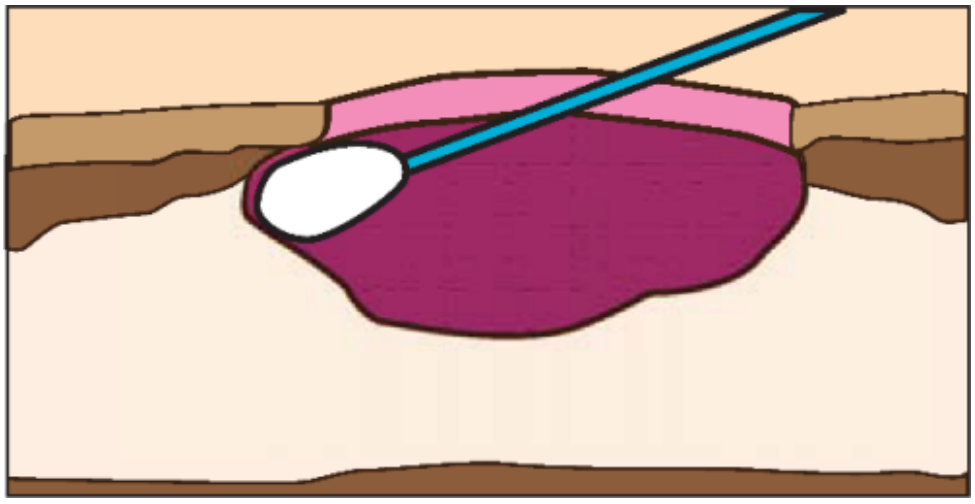

Figure 1.4: Example of undermining in a wound, where tissue loss occurs underneath the wound edges, creating a larger surface area

Necrosis

Necrosis is defined as the presence of dead cells leading to the disintegration of tissues and it is clinically characterized by its quantity, color, consistency, and adherence to the wound bed.

The evaluation of necrotic tissue may be conducted through clinical judgment, to visually estimate the percentage of necrosis covering the wound, or through linear measurements of the dimensions of the necrotic tissue, which involve measuring its length and width to calculate the total surface area affected (same procedure used to estimate the wound size). A review of the literature reveals that visual estimation remains a reliable method in clinical practice [16].

The consistency of necrotic tissue can be described as the cohesiveness of the debris, while adherence refers to the degree to which the necrotic material is attached to the wound bed and the ease of separation from it.

Two principal forms of necrosis are distinguished: slough and eschar. Slough is typically characterized by a yellow color and may appear as thin, viscous, or gelatinous fibrin-like material dispersed throughout the wound bed. Eschar indicates more profound tissue damage and is usually black, gray or brown, presenting various textures from soggy to hard and crusty. A hard, crusty eschar firmly attached to both the base and edges of the wound, occasionally leading to confusion with scab formation.

Exudate

The volume and characteristics of exudate serve as indicators of infection and wound healing progress, as infections contribute to prolonged healing times due to ongoing tissue degradation. In non-infected wounds, exudate is generally serous or serosanguinous. However, when infection is present, the exudate may increase in thickness and volume, becoming purulent.

It is difficult to accurately estimate the amount and characteristics of exudate due to the inherent variations in wound size and the absorptive capacity of dressings. An assessment of the exudate should be conducted following the removal and cleansing of the wound dressing with either normal saline or water.

The estimation of the volume should take into account both the wound and the removed dressing. It is necessary to analyze the wound's moisture level and the amount of fluids absorbed by the dressing, taking into account the absorbency and state of saturation of the bandaging, as well as the time in contact with the wound.

The assessment of drainage volume necessitates the involvement of personnel with a high level of expertise in the evaluation of exudate levels in relation to different wound types and healing phases.

Surrounding Skin

Erythema or induration in the tissue surrounding a wound often provides the earliest indication of potential tissue damage.

Non-blanchable erythema or discoloration may signal impending wound extension due to ischemic or hypoxic injury. While slight firmness is common at the wound edge itself, the surrounding tissue should feel soft and spongy; hardness or firmness in these areas indicates induration. The presence of both erythema and induration around the wound is typically a sign of invasive infection.

Granulation Tissue and Epithelialization

The appearance of granulation and epithelialization indicates the commencement of the proliferative phase of wound healing, which typically precedes the process of wound clo-

sure.

Granulation tissue is characterised by the growth of small blood vessels and connective tissue within the wound cavity. In a healthy state, granulation tissue presents as a bright red, shiny substance with a textured appearance and a tendency to bleed easily. In contrast, the granulation tissue that is unhealthy is observed to be pale pink or dull red. A smooth pink-red base is frequently observed in wounds that have reached a plateau or in wounds involving muscle, indicating that the healing process has ceased.

Epithelialization is the process of restoring and reconstructing the epithelium. This process can occur throughout the wound bed in superficial wounds, affecting only the outermost layers of the epidermis. However, in deeper wounds, it only occurs from the edges of the wound. The newly formed epithelium is observed to be pink and dry in individuals of all racial groups.

It is essential to evaluate and document the percentage of the wound filled with granulation tissue and/or the extent of epithelialization during each wound assessment.

1.2 BWAT - Bates-Jensen Wound Assessment Tool

The Beta-Jensen Wound Assessment Tool (BWAT) was first proposed by Beta-Jensen in 2001 as an improved version of the previous Pressure Sore Status Tool (PSST). Of the various assessment tools currently available, the BWAT is the most widely used internationally as it is considered the gold standard in the field.

Originally validated for the assessment of wound from Pressure Injuries, the BWAT is now used to assess any type of wound, both acute and chronic [17] .

The strength of the BWAT lies exactly in its reliability across a wide range of wounds, to the extent that it has been used as a benchmark to assess the validity of other assessment tools.

1.2.1 BWAT scoring system

The BWAT consists of 15 wound characteristics: location, shape, size, depth, wound edges, undermining or tunneling, type and amount of necrotic tissue, type and amount of exudate, surrounding skin discoloration, peripheral tissue edema and induration, gratulation tissue and epithelialization [18].

Of these characteristics, the first two (location and shape) are only recorded and do not contribute to the score, but are needed by clinicians to monitor gross changes in the visual characteristics of the wound.

The remaining 13 characteristics are scored on a scale of 1 to 5, with 1 being the healthiest condition and 5 being the most severe. For four of these characteristics (size, edges, depth, and undermining), a score of 0 is also possible, indicating *none present* or *healed wound*.

The scores for each item are then simply summed to give a total score between 9 and 65. In clinical practice, a score of 13 or less indicates a healing wound, while higher scores suggest a worsening condition. The scores near the maximum of 65 are associated with severe tissue degeneration.

A brief description of the scoring of each characteristic can be found in Table 1.1.

Item	Score
1.Size	0 = Healed wound 1 = Length x Width $\leq 4cm^2$ 2 = $4cm^2$ < Length x Width $\leq 16cm^2$ 3 = $16cm^2$ < Length x Width $\leq 36cm^2$ 4 = $36cm^2$ < Length x Width $\leq 80cm^2$ 5 = Length x Width > $80cm^2$
2.Edges	0 = Healed wound 1 = Indistinct, not easily visible 2 = Distinct, attached to the wound base 3 = Well-defined, not attached to the wound base 4 = Well-defined, rolled under, thickened 5 = Fibrotic, hyperkeratotic
3.Depth	0 = Healed wound 1 = Intact skin 2 = Partial thickness skin loss involving dermis/epidermis 3 = Full thickness skin loss in subcutaneous tissue 4 = Obscured by necrosis 5 = Full thickness skin loss with tissue necrosis, damage to muscle or bone
4.Undermining	0 = Healed wound 1 = None present 2 = Undermining < $2cm$ in any area 3 = Undermining $2-4cm$ in < 50% wound margins 4 = Undermining $2-4cm$ in > 50% wound margins 5 = Undermining > $4cm$ or Tunneling
5.Necrotic Tissue Type	1 = None visible 2 = White/gray tissue or non-adherent yellow slough 3 = Loosely adherent yellow slough 4 = Adherent, soft, black eschar 5 = Firmly adherent, hard, black eschar
6.Necrotic Tissue Amount	1 = None visible 2 = Wound covered $\leq 25\%$ 3 = $25 <$ % wound covered $\leq 50\%$ 4 = $50 <$ % wound covered $\leq 75\%$ 5 = Wound covered > 75%
7.Exudate Type	1 = None 2 = Bloody 3 = Serasanguineous: thin, watery, pink 4 = Serous: thin, watery, clear 5 = Purulent: opaque, tan/yellow

Item	Score
8.Exudate Amount	1 = Dry wound 2 = Scant, wound moist but no exudate 3 = Small 4 = Moderate 5 = Large
9.Surrounding Skin Color	1 = Pink or normal for ethnicity 2 = Bright red and/or blanches to touch 3 = White/gray pallor 4 = Dark red or purple 5 = Black or hyperpigmented
10.Peripheral Edema	1 = No edema 2 = Non-pitting edema extends $< 4cm$ around wound 3 = Non-pitting edema extends $> 4cm$ around wound 4 = Pitting edema extends $< 4cm$ around wound 5 = Crepitus and/or pitting edema extends $> 4cm$ around wound
11.Peripheral Induration	1 = No induration 2 = Induration, $< 2cm$ around wound 3 = Induration, $2-4cm$ extending $< 50\%$ around wound 4 = Induration, $2-4cm$ extending $> 50\%$ around wound 5 = Induration, $> 4cm$ around wound
12.Granulation Tissue	1 = Skin intact 2 = Bright red, 75 to $< 100\%$ of wound filled 3 = Bright red, 25 to $< 75\%$ of wound filled 4 = Pink or dusky red and/or $< 25\%$ of wound filled 5 = No granulation tissue
13.Epithelialization	1 = 100% wound covered 2 = 75 to $< 100\%$ wound covered and/or epithelial tissue extends $> 0.5cm$ into wound bed 3 = 50 to $< 75\%$ wound covered and/or epithelial tissue extends $< 0.5cm$ into wound bed 4 = 25 to $< 50\%$ wound covered 5 = Wound covered $< 25\%$

Table 1.1: Table displaying the parameters of the BWAT and their corresponding classification values for wound assessment.[7]

1.3 PWAT - Photographic Wound Assessment Tool

The Photographic Wound Assessment Tool (PWAT) is a specifically developed instrument for the evaluation of wound status based on wound photographs. Unlike subjective descriptions, which may be prone to misinterpretation, a medical photograph can convey more precise and reliable information.

Advancements in healthcare technology have made digital imaging an integral part of telemedicine. The ability to transmit high-quality images in real time, even with small file sizes, allows for effective remote consultations, especially for patients in isolated or underserved areas. This advancement is possible as the reliability of decisions based on digital images needs to be comparable to bedside evaluations.

In wound care, photography enhances both assessment and documentation, creating a consistent visual record to monitor changes in wound appearance over time. Furthermore, by using image analysis software, sequential wound images can be evaluated to measure wound size and depth, such as other characteristics.

For accurate photographic documentation of wounds, specific guidelines should be followed. This includes consistently positioning the patient for each photo, maintaining the same angle and distance for every image, and capturing both detailed close-ups and broader contextual views of the wound area. Additionally, a centimeter ruler and a color guide should be included in each photograph for reference. The images should also be labeled with the date, anatomical site, and an anonymous patient ID to protect patient privacy.

1.3.1 PWAT scoring system

Originally, the PWAT included a subset of six BWAT parameters that could be assessed from photographs alone and did not require bedside examination.

These domains are wound edges, necrotic tissue type and amount, surrounding skin, granulation tissue type and epithelialization. Each domain was scored on a scale from 0 to 4, adjusted from the original BWAT scale (1 to 5) to allow a score of zero to indicate a completely healed wound [19].

The PWAT has been validated, showing a good interrater consistency (intraclass correlation coefficient $\simeq 0.70$). It was also demonstrated to be sensitive to changes over time, making it a valuable tool for repeated photographic assessments. The updated version of the PWAT expanded the parameters by adding two more domains: wound size and depth, bringing the total to eight domains[20]. The scoring remains consistent across all domains, with a total score ranging from 0 (fully healed) to 32, representing the most severe wound condition.

A brief description of scoring instruction for the PWAT can be found in Table 1.2 .

Item	**Score**
1.Size	0 = Healed wound or nearly closed ($< 0.3cm^2$) 1 = $0.5cm^2$ < Surface area $\leq 2cm^2$ 2 = $2cm^2$ < Surface area $\leq 10cm^2$ 3 = $10cm^2$ < Surface area $\leq 20cm^2$ 4 = Surface area > $20cm^2$

Item	Score	
2.Depth	0 = Healed wound 1 = Full thickness 2 = Unable to judge, wound base covered by yellow/black eschar 3 = Full thickness of underlying tissue layers 4 = Tendon joint capsule visible or bone in wound base	
3.Edges	0 = Healed wound or edge indistinct 1 = > 50% of edges attached with an advancing border of epithelium 2 = > 50% of edges attached even with wound base, not advancing 3 = > 50% of edges unattached and/or undermined 4 = > 50% of edges rolled, thickened or fibrotic	
4.Necrotic Tissue Type	0 = None visible 1 = White/gray tissue or yellow slough 2 = > 50% of tissue is thick, adherent white yellow slough 3 = > 50% of tissue is white/gray and devitalized or eschar 4 = > 50% of tissue is hard gray/black eschar	
5.Necrotic Tissue Amount	0 = None visible 1 = Wound covered < 25% 2 = 25 < % wound covered ≤ 50% 3 = 50 < % wound covered ≤ 75% 4 = Wound covered > 75%	
6.Granulation tissue	0 = Healed wound or nearly closed ($< 0.3cm^2$) 1 = > 50% of tissue appears bright red 2 = > 50% of granulation tissue is unhealthy (pale, dusk, hypergranulation) 3 = > 50% of tissue is damaged, friable, degrading 4 = No granulation tissue (all necrotic)	
7.Epithelialization	0 = Healed wound or nearly closed ($< 0.3cm^2$) 1 = Wound covered with granulation tissue > 75% 2 = 50 < % wound covered ≤ 75% 3 = 25 < % wound covered ≤ 50% 4 = 25 < % wound covered ≤ 50%	
8.Skin Surrounding Wound	**# of factors** 0 = None 1 = One only 2 = 2 or 3 3 = 4 or 5 4 = 6 or more	**Factors** Maceration Edema Excoriation Bright red erythema Others:_____

Table 1.2: PWAT parameters and their scoring system [20].

Chapter 2

Machine Learning Algorithms for Wound Management

The concept of **Artificial Intelligence (AI)** is attributed to Alan Turing, who in 1950 laid its foundations with the *Turing Test* [21]. Artificial Intelligence is a broad field that encompasses all systems and techniques aimed at mimicking human intelligence, such as reasoning, problem-solving, and decision-making.

In 1959, Arthur Samuel introduced **Machine Learning (ML)** , a subset of AI, as *"the ability of computers to automatically learn from data and improve their performance over time without being explicitly programmed"* [22].

Hence, **Machine Learning** focuses on algorithms and statistical models that allow computers to learn from data and make predictions or decisions based on it.

One of the firsts ML algorithm was the **Perceptron**, the simplest neural network model developed in 1957 by Frank Rosenblatt [23].
Nevertheless, the period of significant advancement of neural networks commenced only in 1986 with the development of the back-propagation algorithm by Hinton et al. [24].
In 2006, Hinton et. al also introduced the concept of **deep learning**, based on deep neural networks, which have been demonstrated to be highly effective in image recognition and/or classification [25].

Over time, others ML algorithms have been developed in parallel, including **Linear Discriminant Analysis (LDA)**, **Random Forests** and various **Clustering Algorithms**. These new methods have expanded the possibilities for data analysis and improved the flexibility of predictive models.

Artificial Intelligence has identified a range of potential applications in the medical domain, with significant implications for the speed and precision of diagnoses, as well as the delivery of patient-centric, personalised care. Specifically, the utilisation of ML in the assessment and management of wounds has allowed the analysis of extensive datasets, comprising wound images, patient clinical information and treatment histories, thereby offering substantial assistance to medical personnel.

The ML algorithms can be broadly classified as **supervised** or **unsupervised** learning[26], on the basis of the learning paradigm they follow.

- **Supervised learning** algorithms are provided with a dataset that includes both features and corresponding labels or targets. The goal is to learn how to predict the label for previously unseen examples based on their features. Essentially, supervised learning entails the construction of a mapping function from input x to output y, typically by estimating the conditional probability $p(y|x)$.

- **Unsupervised learning** algorithms are exposed to a dataset containing many features but no labels. Their aim is to uncover the hidden structure or to find patterns within the data without any explicit guideline.
  One such example is **clustering**, whereby the algorithm groups data items with similar characteristics together.

It is important highlight that **supervised** and **unsupervised** learning are not inherently fixed in their definition and many algorithms can be adapted to handle both types of tasks.

For instance, **semi-supervised** learning algorithms represent a hybrid of the two aforementioned paradigms. They employ a restricted quantity of labeled data and a considerably larger volume of unlabeled data, with the objective of capitalizing on the strengths of both supervised and unsupervised approaches.

2.1 Linear Discriminant Analysis - LDA

Linear Discriminant Analysis (LDA) is a supervised machine learning technique for **classification** and **dimensionality reduction**, particularly effective when the classes are linearly separable and follow a normal distribution.

The original algorithm was developed by R. Fisher in 1936 [27], which introduced the concept of LDA to separate classes of data based on measurable features, creating a statistical method that maximizes the separation between classes through a linear combination of these features.

The first step of LDA is the computation of two matrices from the dataset: the **Between-Class Scatter Matrix** and the **Within-Class Scatter Matrix**.

- **Within-Class Scatter Matrix ($\mathbf{S}_W$)** quantifies the spread of data points within each class. It is computed as follows:

$$\mathbf{S}_W = \sum_{i=1}^{C} \sum_{\mathbf{x} \in C_i} (\mathbf{x} - \boldsymbol{\mu}_i)(\mathbf{x} - \boldsymbol{\mu}_i)^T \tag{2.1}$$

  where:

  - C is the total number of classes,
  - C_i are the data points belonging to class i,
  - $\mathbf{x}$ is the feature vector for a data point,
  - $\boldsymbol{\mu}_i$ is the mean feature vector for class i.

- **Between-Class Scatter Matrix ($\mathbf{S}_B$)** measures the separation between the means of different classes. It is calculated as follows:

$$\mathbf{S}_B = \sum_{i=1}^{C} N_i (\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})^T \tag{2.2}$$

  where:

– N_i is the number of data points in class i,

– $\boldsymbol{\mu}$ is the overall mean feature vector of all data points in the dataset,

– $\boldsymbol{\mu}_i$ is the mean feature vector for class i.

Once these matrices are computed, LDA finds optimal projection directions addressing the generalized eigenvalue problem involving $\mathbf{S}_B$ and $\mathbf{S}_W$:

$$\mathbf{S}_B\mathbf{w} = \lambda\mathbf{S}_W\mathbf{w} \tag{2.3}$$

where:

- λ represents the eigenvalues,
- $\mathbf{w}$ denotes the eigenvectors.

The largest eigenvalues correspond to directions that capture the most variance related to class differences, i.e., they correspond to the non-zero components of the eigenvectors, which represent the features providing the most information for the classification.
The top k eigenvectors define a new feature space where the data is projected for classification. The value of k is a integer chosen a priori, with the condition $k \leq C - 1$, where C is the number of classes in the dataset.

Finally, LDA identifies the optimal projection vector that best separates the classes in the lower-dimensional space (example in Figure 2.1), by maximizing the following ratio:

$$J(\mathbf{w}) = \frac{\mathbf{w}^T\mathbf{S}_B\mathbf{w}}{\mathbf{w}^T\mathbf{S}_W\mathbf{w}} \tag{2.4}$$

where $\mathbf{w}$ is the projection vector onto which the data points are projected.

This approach aims to maximize the separation between different classes while minimizing variance within each class, ultimately leading to improved classification accuracy.

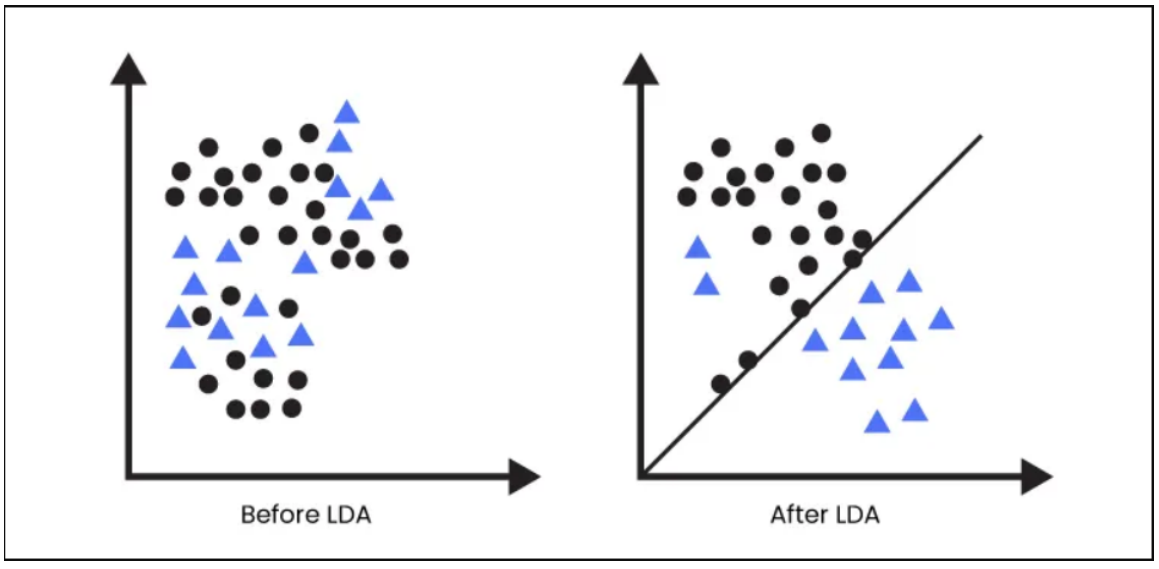

Figure 2.1: Example of Linear Discriminant Analysis (LDA) applied to a two-class dataset, showing the separation of classes in the transformed feature space.

2.2 Random Forest - RF

Random Forest (RF) is a supervised machine learning algorithm proposed by Breiman in 2001 [28] and designed for both **classification** and **regression** tasks. It applies the concept of *bagging* to decision trees while also introducing additional elements of randomness.

Bagging (short for **Bootstrap Aggregating**) is an ensemble technique that aims to enhance the stability and accuracy of inherently unstable models [29].
The concept is to generate multiple versions of the same model through the training of each on a distinct subset of the training data. These subsets are generated via **bootstrap method** [30], which entails the random sampling from the original dataset with the possibility of multiple selections of the same element. Subsequently, the predictions from each individual model are aggregated. In the case of regression tasks, the average of predictions is taken; in cases of classification tasks, a majority vote (plurality vote) is employed. The combination of multiple outputs reduces model variance and prevents overfitting.

A **Decision Tree (DT)** is a predictive model that represents a sequence of decisions and their potential consequences in a hierarchical tree structure. The architecture of a decision tree can be described as follows:

- **Root Node**: The starting point of the tree that contains the entire dataset, where the data is initially split.
- **Internal Nodes**: Each node represents a decision based on a specific feature of the dataset. The branches of the node correspond to the possible outcomes of that decision. The splits are chosen to optimize the separation of the data by minimizing some loss function which depends on the task being solved.
- **Leaf Nodes**: They represent the final predictions of the model. In classification problems, each leaf contains a class label, while in regression problems it holds a continuous value.

The Decision Tree recursively divides the dataset until a leaf node is reached, thereby determining the final prediction.

The process behind this partitioning is based on the **CART** (Classification and Regression Trees) algorithm which was introduced by Breiman in 1986 [31] and remains widely used today. The algorithm splits the data in each node minimizing an impurity, or loss, function. The choice of the impurity depends on the task being solved.

However, **Decision Trees** are recognized as inherently **unstable models**, meaning that small fluctuations in the training data can result in significant alterations in the tree structure. This instability renders them particularly well-suited for bagging.

Indeed, the **Random Forest** algorithm exploits it by constructing an ensemble of Decision Trees, each trained on a different bootstrap sample of the dataset.
Random Forest improves the bagging approach by incorporating an additional layer of randomness, i.e. constructing each decision tree with random subset of features (see Figure 2.2).

The additional layer of feature selection results in:

- An improvement in **robustness** and in model ability to generalize to unseen data, due to the increase in diversity among trees and the reduction in correlation between them.
- An increase in model efficiency in high-dimensional datasets, reducing the risk of overfitting to specific features.
- Decrease of generalization error, as it depends on the **Strength of Individual Trees** (accuracy of each tree), and the **Correlation Between Trees**, which is significantly reduced by the random selection, as previously stated.

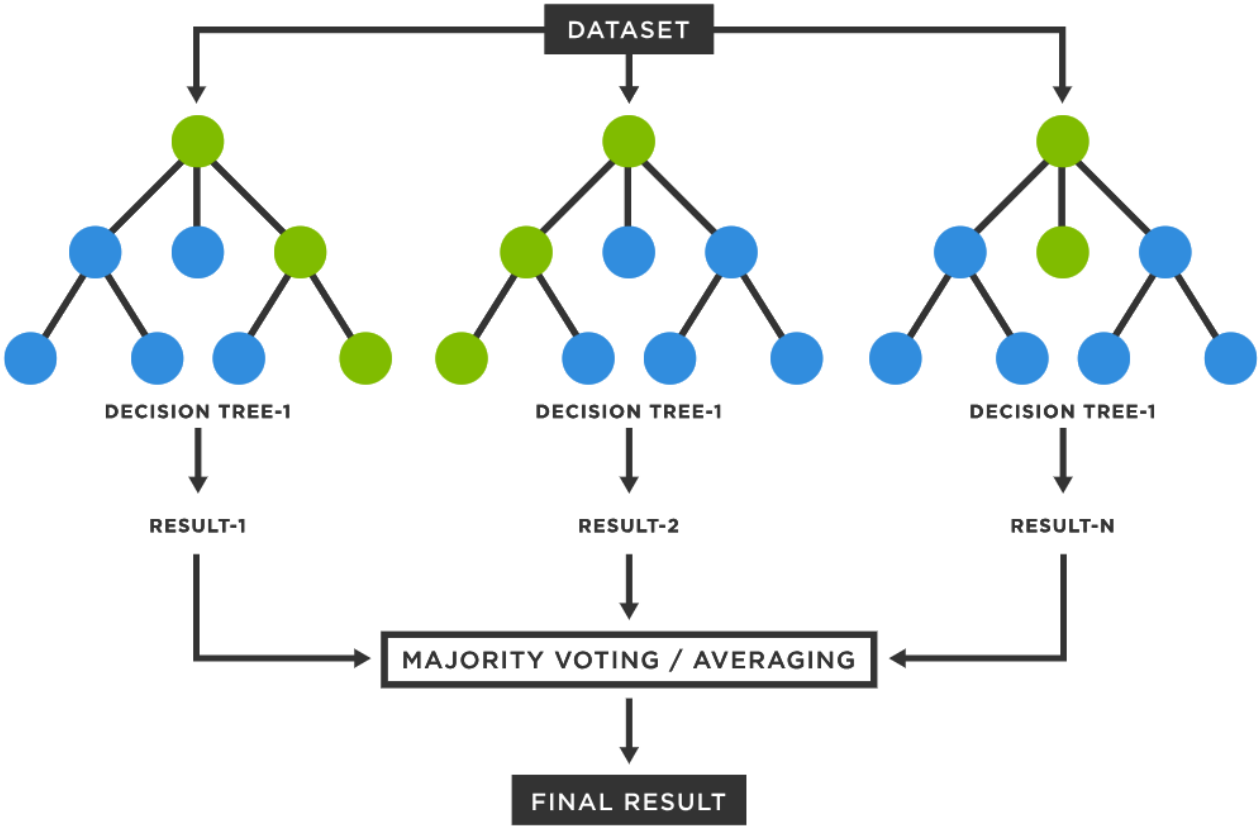

Figure 2.2: Example of a Random Forest model: tree are built with a subset of random features. Outputs of trees are combined to generate the finale results.

2.3 Clustering Algorithms

Clustering is an **unsupervised** machine learning technique that is employed to group a set of data into clusters based on their similarity. The model's task is to automatically identify the underlying structure or groups in the data without any provided label. The posterior cluster validation can assist in the interpretation of the results (see Figure 2.3).

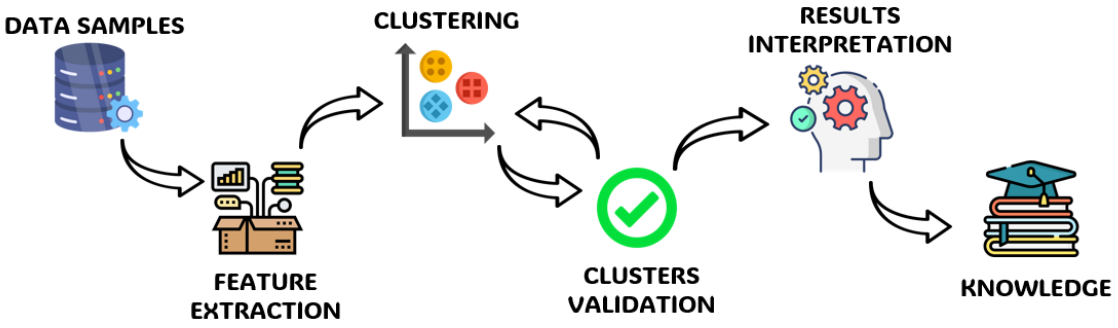

Figure 2.3: Overview of the key steps in clustering process.

2.3.1 Distance and similarity functions

The foundational concepts of Clustering concern to the notions of distance (dissimilarity) and similarity between the data points. Clustering algorithms assess one of these two properties using specific mathematical functions [32]. In the case of quantitative data, dissimilarity is typically measured, whereas in the case of qualitative data similarity is computed.

Below are the most commonly used distance metrics. In all the following equations x_i represents the feature vector on the point i in the data space.

- **Minkowski Distance**

$$d_M = \left(\sum_{i=1}^{d} |x_{il} - x_{jl}|^n \right)^{1/n} \tag{2.5}$$

  It is equivalent to Euclidean distance when $n = 2$.

- **Standardized Euclidean Distance**

$$d_E = \left(\sum_{i=1}^{d} \left| \frac{x_{il} - x_{jl}}{s_l} \right|^2 \right)^{1/2} \tag{2.6}$$

  s_l stands for the standard deviation.

- **Cosine Distance**

$$1 - \cos(\alpha) = \frac{x_i^T x_j}{\|x_i\| \|x_j\|} \tag{2.7}$$

  It is invariant to a rotation change in the data.

- **Pearson Correlation Distance**

$$1 - \frac{\mathrm{Cov}(x_i, x_j)}{\sqrt{D(x_i)} \cdot \sqrt{D(x_j)}} \tag{2.8}$$

  Where Cov stands for the covariance and D stand for the variance. It measures the distance based on the linear correlation.

- **Mahalanobis Distance**

$$(x_i - x_j)^T S^{-1} (x_i - x_j) \tag{2.9}$$

  Where$S$ is the covariance matrix inside the cluster.

For completeness, below are explained the two most common similarity functions.

- **Jaccard Similarity** is defined as:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{2.10}$$

  Where $|X|$ represents the number of elements in set X.
  The function measures the similarity between two sets. It is also possible to define the Jaccard distance as $J_d = 1 - J(A, B)$.

- **Hamming Similarity** is defined as the minimum number of substitutions required to transform one data point into another. The smaller the number, the higher the similarity. The Hamming distance is defined as the opposite of Hamming similarity.

2.3.2 Types of clustering

Despite sharing the same basic concept, Clustering Algorithms differ significantly in their practical approaches, leading to variations in results depending on the methods and metrics used (see Figure 2.4). Below, we explore the main types of Clustering Algorithms and their unique characteristics.

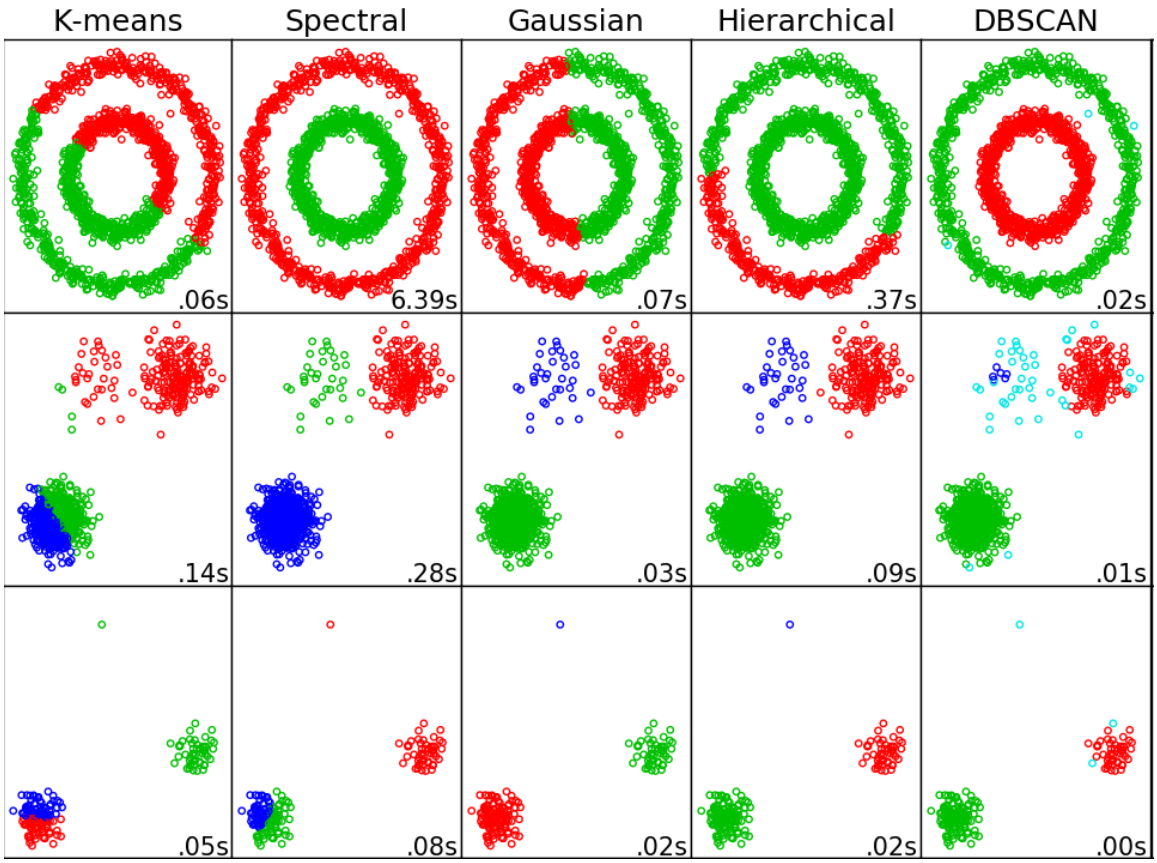

Figure 2.4: Visual comparison of different clustering algorithms, illustrating the variation in results depending on the model.

Centroid-Based Clustering

Centroid-based algorithms belong to the category of hard clustering, where each data point is assigned exclusively to one cluster. These methods partition the data into k clusters, where the number k is decided a-priori intuitively or estimating it with mathematical methods (e.g. *elbow method*).

Each clusters is represented by a centroid chosen either randomly or through some other heuristic and each data point is assigned to the cluster whose centroid is the closest, based on a distance metric. After all points have been assigned to a cluster, the centroid of each cluster is updated and the process is iterated until convergence. These methods assume spherical shape and equal size clusters.

Hierarchical Clustering

Hierarchical Clustering creates a tree-like structure, called **"dendrogram"**, that illustrates how data points are grouped or divided. This method can be implemented in two ways: agglomerative (bottom-up) or divisive (top-down).

In the **Agglomerative Clustering** each data point is initially a distinct cluster. In each iteration, the two closest clusters are merged into a single cluster. This merging continues until all data points form one single cluster or are divided in the desired number of clusters (if a cutoff is applied).

In **Divisive Clustering** the process starts with all data points in a single cluster and at each iteration the cluster with the highest internal dissimilarity is split into two smaller clusters. The splitting continues until each data point is in its own cluster, or the desired number of clusters is reached.

Density-Based Clustering

Density-Based algorithms, such as DBSCAN (Density-Based Spatial Clustering of Applications with Noise), fall into the category of flexible clustering techniques capable of detecting clusters of arbitrary shapes. These methods do not require a predefined number of clusters. Instead they use parameters like the search radius ϵ and the minimum number of points (MinPts) to identify dense regions. Points that do not meet the density criteria are considered outliers.

The process starts with an unvisited point in the dataset for which all its neighbors are retrieved (i.e. points within the radius ϵ). If the number of detected neighbors is greater than or equal to MinPts, the point is classified as a **core point** and is assigned to a cluster. Otherwise, it is classified as **noise**, unless it lies within the search radius of a core point.

Once a core point is identified, all points within its neighbors are recursively added to the cluster. If a neighboring point is also a core point, its neighbors are explored and added to the cluster as well (**density reachability).** Points that are within the search radius of a core point but have fewer neighbors points than MinPts are included in the cluster as **border points**.

The algorithm then moves to the next unvisited, not assigned point and repeats the process. Once all points have been processed, the result is a set of clusters, where each point is either assigned to a cluster or labeled as noise.

Probabilistic Distribution-Based Clustering

Probabilistic algorithms are examples of soft clustering methods that provide a more flexible classification of the data.

In the initial stage, the parameters for the probabilistic distributions of each cluster are assigned. The probability that each data point belongs to each cluster (referred to as "responsibility") is calculated based on the current parameters, which are then updated by maximizing the likelihood of the data, using the calculated responsibilities.

Therefore, the responsibility of each point are recomputed, and the process is repeated until the model parameters converge. Finally, each data point is assigned to the cluster with the highest probability of belonging, creating a soft classification where each point can belong to multiple clusters with different probabilities.

Graph Theory-Based Clustering

Graph Theory-Based Clustering algorithms are flexible methods treat the data as a graph, where nodes represent data points and edges indicate similarities. The weight of an edge typically reflects the strength of the similarity between two points.

Therefore, the Laplacian matrix, that encode how nodes are connected, is computed. Subsequent to this stage, the methods diverge in accordance with the approach utilized.

For instance, in the Spectral Clustering, the eigenvectors and eigenvalues of the matrix are calculated and the eigenvectors with the smallest eigenvalues are used to project the data into a lower-dimensional space, where similar points are closer together. Finally, standard clustering methods, are applied to this transformed space in order to assign each point to a cluster.

2.3.3 Dimensionality Reduction for Clustering

The phenomenon referred to **"curse of dimensionality"** arises when dealing with high-dimensional datasets. This phenomenon complicates the identification of meaningful patterns or neighbourhood structures in the data space and also increases computational complexity.

Dimensionality reduction techniques mitigate this issue by transforming data into a lower-dimensional space, preserving the most significant features and improving both clusters visualization and class separability. This capability is crucial in unsupervised clustering, where accurate representation of the data underlying structure significantly impacts the performance of algorithms.

Among the state-of-the-art methods, **PaCMAP** (Pairwise Controlled Manifold Approximation Projection) excels in reducing dimensionality while preserving both local and global data structures [33]. Compared to other techniques, such as t-SNE (t-Distributed Stochastic Neighbor Embedding) or UMAP (Uniform Manifold Approximation and Projection), PaCMAP is often preferred for balancing local interpretability with global structure preservation (example in Figure 2.5).

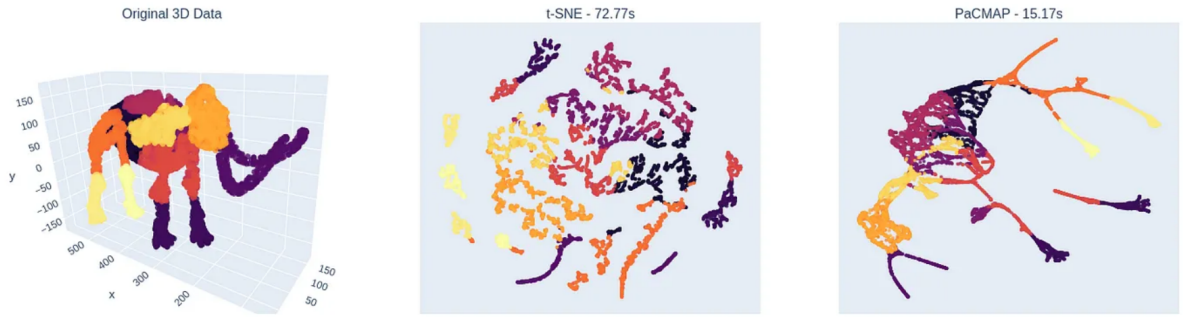

Figure 2.5: Comparison between the application of t-SNE and PaCMAP for dimensionality reduction of the Mammoth dataset [33].

2.4 Artificial Neural Networks - ANN

Artificial Neural Networks (ANN) are supervised learning models inspired by the structure and function of biological neural networks in the human brain. These algorithms aim to simulate the brain's ability to learn patterns and generalize from data.

The fundamental computational units in ANN are called ***"neurons"***, reflecting their functional similarities to biological neurons.

2.4.1 Perceptron

The simplest ANN is the **Perceptron**, a binary linear classifier. It consists of a single layer of neurons that receive inputs and compute a weighted sum, which is then passed as argument to an activation function (scheme in Figure 2.6).

From a mathematical point of view this can be seen as:

$$f\left(\sum_{i=1}^{n} w_i x_i + b\right) = y \tag{2.11}$$

where:

- x_i are the **inputs**;
- w_i are the **weights** associated with the inputs x_i;
- f is the **activation function** which take as argument the outputs of the neurons;
- b is the **bias** of the activation function that improves the model's fit to the data;
- y is the **output**, result of the activation function.

The activation function of the Perceptron is the **Heaviside step function**, which produces a binary output based on whether the input is greater than or equal to a specified threshold:

$$y = f(z) = \begin{cases} 1 & \text{if } z \geq \text{threshold} \\ 0 & \text{if } z < \text{threshold} \end{cases} \tag{2.12}$$

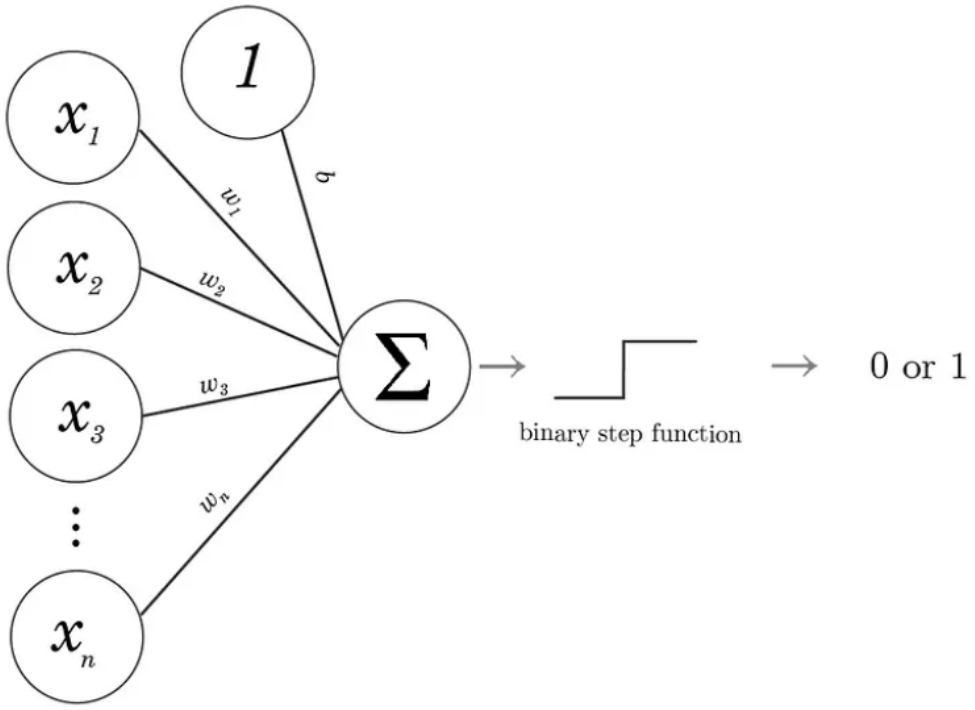

Figure 2.6: Architecture of Perceptron

2.4.2 Feedforward Neural Networks - FNN

More complex neural networks consist of three primary types of layers:

- **Input Layer**: It receives the input data and passes it on to the subsequent layer.
- **Hidden Layers**: They apply non-linear transformations to the input data.
- **Output Layer**: it produces the final prediction.

If the overall architecture is simple, with only a few hidden layers (typically 2 or 3), we refer to them as **Feedforward Neural Networks** (FNN), or **Shallow Networks** (Figure 2.7).

Each neuron in a layer receives inputs from all neurons in the previous layer and uses them to produce an output through the application of Equation 2.11. The outputs are then transmitted as inputs to the neurons in the subsequent layer, forming a dense network of connections.

The fundamental principles of Shallow Network (**activation functions, cost functions, backpropagation**) remain consistent even as the number of hidden layers increase drastically. However, deeper architectures introduce new challenges that necessitate additional solutions. The next section will explore these concepts in detail, keeping in mind that the underlying mechanics of these principles are consistent even in simpler, Shallow Networks.

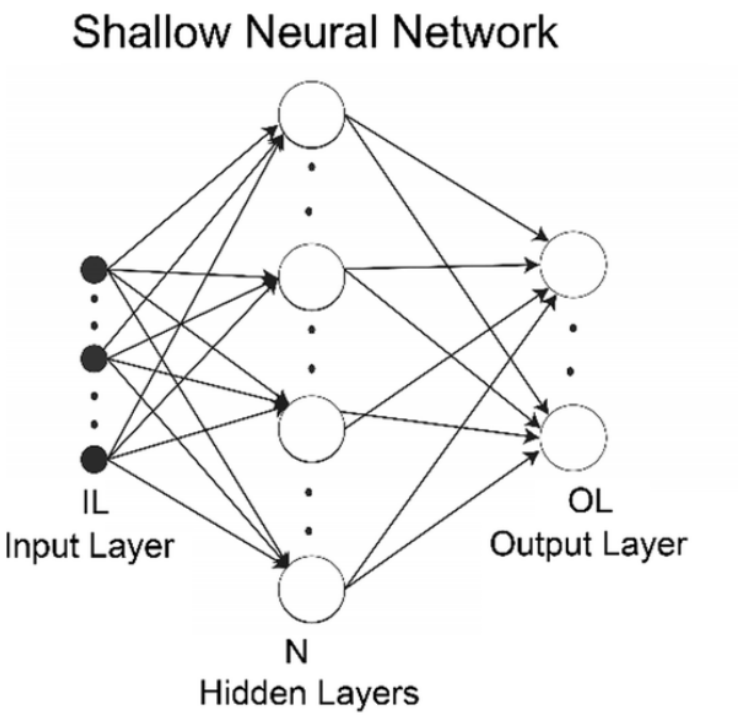

Figure 2.7: Architecture of a Feedforward Neural Network. [34]

2.4.3 Deep Neural Networks - DNN

When the number of hidden layers grows significantly, the network is referred to as a **Deep Neural Network (DNN)** (Figure 2.8).

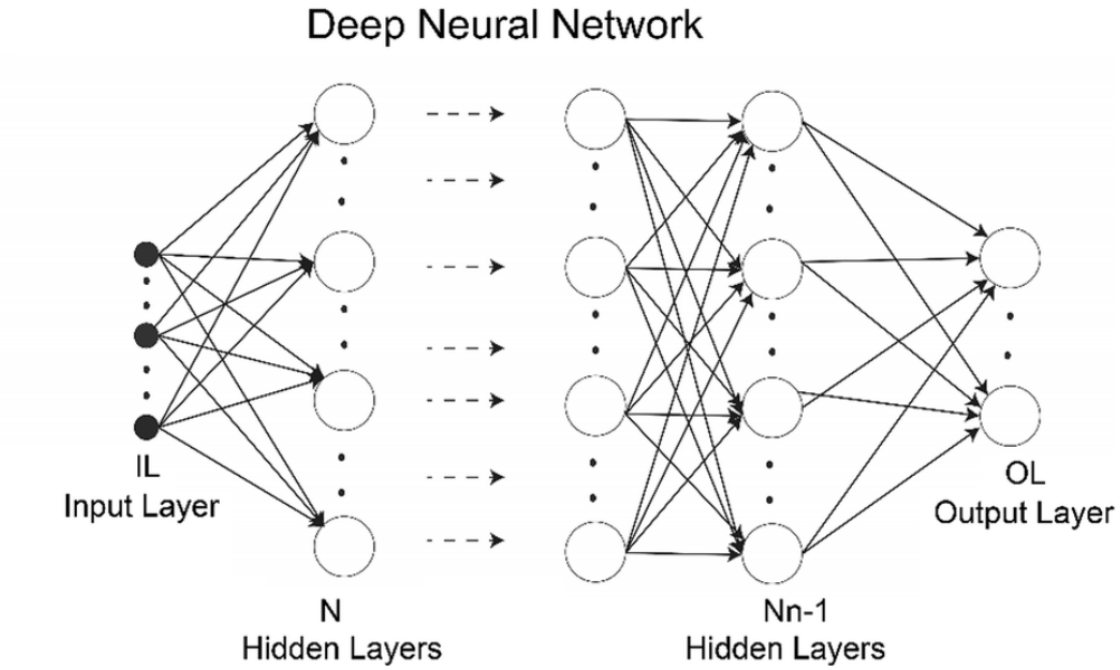

Figure 2.8: Example of architecture of a Deep Neural Network (specifically a MultiLayer Perceptron) [34]

The Deep Neural Networks are ANN that more closely resemble the brain's complex structure and function. The greatest depth of these networks allows them to capture and learn more abstract features from data at each layer, boosting their ability to generalize across complex datasets.

Unlike traditional machine learning models that require manual feature extraction, DNN can automatically learn hierarchical features directly from raw data (Figure 2.9).

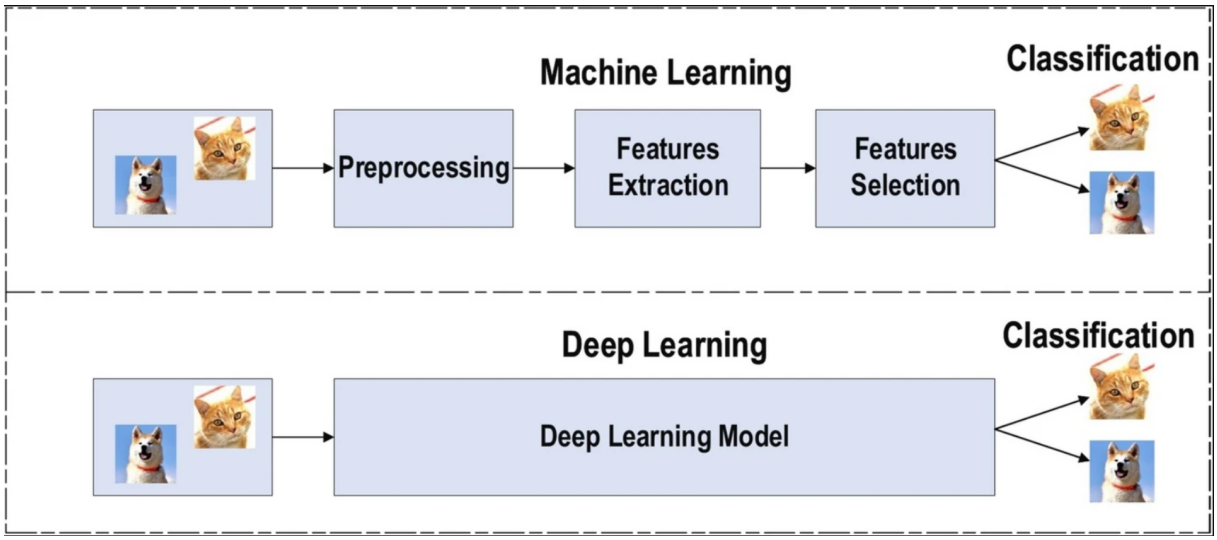

Figure 2.9: Illustration highlighting deep learning ability to automatically learn hierarchical features from raw data, in contrast to traditional machine learning models. Adapted from [35].

Deep Neural Networks encompass a variety of architectures, including:

- **MultiLayer Perceptrons** (MLP): dense fully connected networks, commonly employed for tasks such as classification and regression.
- **Recurrent Neural Networks** (RNN): designed to handle sequential data (e.g. text and time-series signals), incorporate feedback loops that enable the capture of temporal dependencies.
- **Convolutional Neural Networks** (CNN): employ convolution operations to extract spatial features from input data, rendering them particularly well-suited for image analysis.

Activation Functions

The non-linearity in neural networks is introduced through the application of a nonlinear activation function, which enables the network to learn and model complex patterns.

The most commonly used and effective activation functions are:

- **Sigmoid**:

$$\sigma(x) = \frac{1}{1 + e^{-x}},$$

which squashes the input between 0 and 1, often used in binary classification tasks.

- **ReLU (Rectified Linear Unit)**:

$$\text{ReLU}(x) = \max(0, x)$$

outputs the input directly if it is positive, otherwise zero. Typically used in regression problems.

- **Tanh**:

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

is similar to Sigmoid but outputs values between -1 and 1, making it useful in situations where negative values are important (e.g. RNN)

- **Leaky ReLU**:

$$\text{Leaky ReLU}(x) = \begin{cases} x & \text{if } x > 0 \\ \alpha x & \text{if } x \leq 0 \end{cases}$$

a modified ReLU that allows a small, non-zero gradient (α) when the input is negative to avoid *dead neurons*, i.e. neurons that become inactive (only outputs zero).

- **Swish**:

$$\text{Swish}(x) = x \cdot \sigma(x)$$

a smooth function that is defined as the input multiplied by the Sigmoid. It often improves model performance and increase the convergence speed.

- **Softmax**:

$$\text{Softmax}(x_i) = \frac{e^{x_i}}{\sum_j e^{x_j}}$$

a generalization of the sigmoid function for multi-class classification. It takes a vector of raw scores and transforms them into a probability distribution where each value represents the probability that the input belongs to a class.

Cost Functions

The **cost function**, or **loss function**, is a metric that quantifies the discrepancy between the predicted output of the network and the true target values.

In the context of regression tasks, the cost function most commonly employed is the **Mean Squared Error (MSE)**:

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \tag{2.13}$$

where:

- y_i are the true values,
- $\hat{y}_i$ are the predicted values,
- N is the number of samples.

For classification tasks, especially those with multiple classes, the **categorical cross-entropy loss** H is frequently employed:

$$H = -\sum_{i=1}^{N}\sum_{j=1}^{C} y_{ij}\log(\hat{y}_{ij}) \tag{2.14}$$

where:

- N is the total number of samples,
- C is the number of classes,
- y_{ij} is the true label (1 if sample i belongs to class j, 0 otherwise),
- $\hat{y}_{ij}$ represents the predicted probability that sample i belongs to class j.

For classification tasks with imbalanced classes, the **Focal Loss** $\mathcal{L}_{\text{focal}}$ is commonly used to focus on hard-to-classify classes:

$$\mathcal{L}_{focal} = -\alpha(1-\hat{y}_i)^{\gamma}\log(\hat{y}_i) \tag{2.15}$$

where:

- $\hat{y}_i$ is the predicted probability for class i,
- α is a *balancing factor* that adjusts the importance of different classes,
- γ is the *focusing parameter*, which reduces the loss for well-classified examples by down-weighting them by a factor $(1-\hat{y}_i)^{\gamma}$.

Backpropagation Algorithm

The training process of a neural network is fundamentally based on the **backpropagation** algorithm [24], which employs **gradient descent** to iteratively adjust the weights and minimize the cost function.

The algorithm (scheme in Figure 2.10) consists of the following steps:

1. **Forward Propagation**: The input data is processed through the network in order to generate the predicted output.
2. **Error Estimation**: The error is calculated using the cost function.
3. **Backpropagation**: The error is propagated backwards through the network, and the gradients of the weights are calculated.
4. **Weight Update**: The weights are modified according to the formula:

$$w_{ij} = w_{ij} - \alpha\frac{\partial E}{\partial w_{ij}} \tag{2.16}$$

where:

- w_{ij} are the weight between neurons i and j,

- α is the **learning rate**, the parameter that determine the step size along the descent gradient,
- $\frac{\partial E}{\partial w_{ij}}$ is the gradient of the cost function with respect to the weight.

The **learning rate** is a critical factor in the training process: if it is set too high, the model may fail to converge; conversely, if it is set too low, the training process may become unnecessarily slow problem).

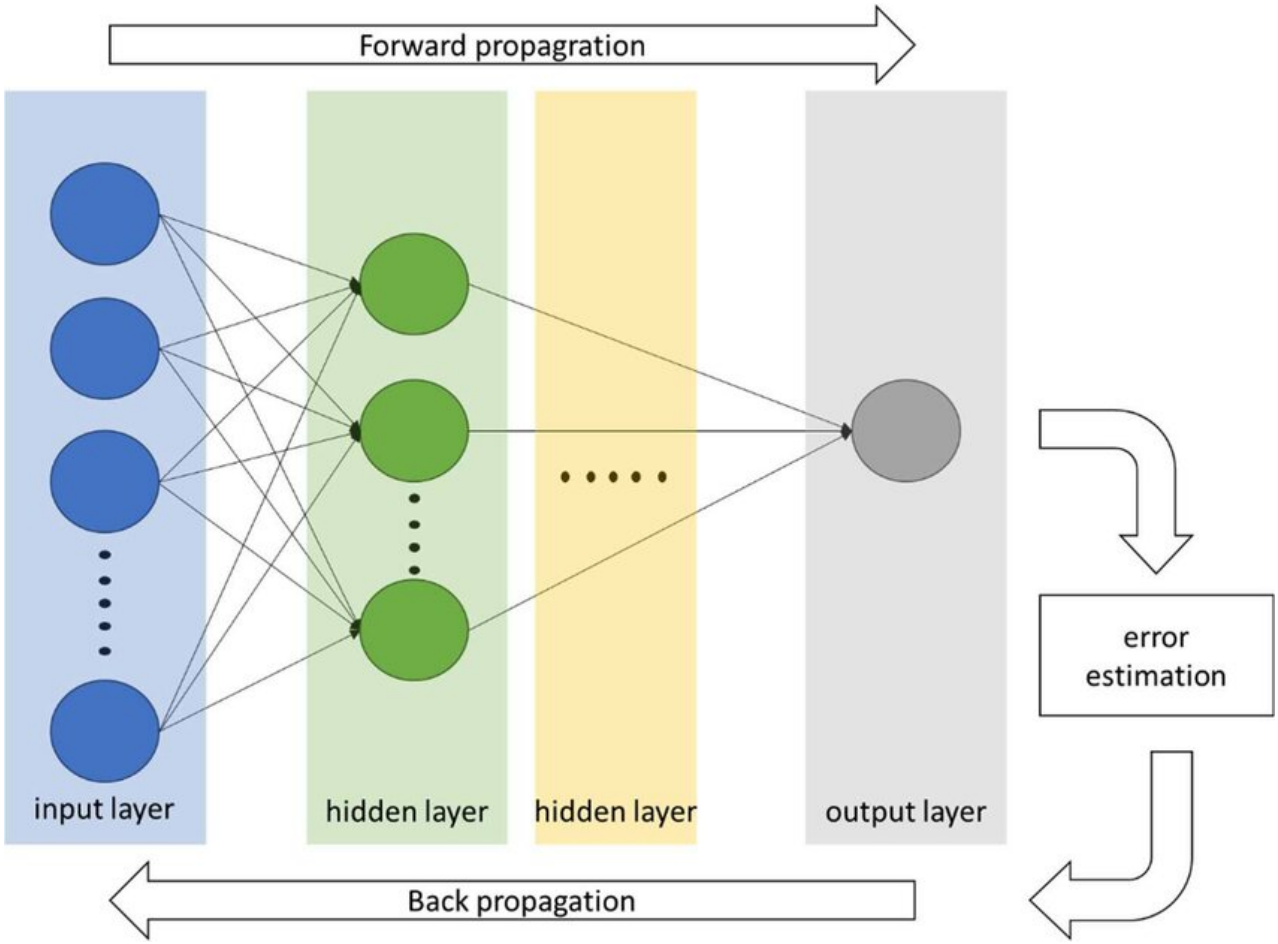

Figure 2.10: Scheme of backpropagation algorithm. [36]

Optimization of Training Process

Deep neural networks present significant challenges in training as their depth increases, mainly due to some critical issues, including **dead neurons** and the **vanishing/exploding** gradient problems [35].

- **Dead neurons** refer to those hat become inactive and no longer contribute to learning, usually because saturated activation functions prevent neurons from significantly updating their weights.
- The **vanishing gradient** issue arises during backpropagation when the gradients become extremely small. This results in negligible weight updates, which severely impairs the network's ability to learn.

- The **exploding gradient** problem occurs when gradients become too large, causing unstable weight updates leading to erratic behaviour and failure to converge.

Advanced weight initialization techniques can mitigate these problems by tailoring the process to the specific activation functions used in the network.
For instance, **Xavier** initialization is particularly effective for activation functions such as *sigmoid* or *tanh*, which have derivatives with a small range (respectively 0.25 and 1), as it ensures that the gradients do not decrease too quickly [37].
Conversely, the **He** method is suitable for the ReLU function because it helps reduce the risk of dead neurons by taking into account the positive bias of ReLU activations. It also helps to avoid the exploding gradient problem that can occur in ReLU-based networks when weights are initialized too large [38].

A technique for improving the flow of gradients through the network is **skip connections**, which allows the output of one layer to be added directly to the output of a deeper layer, bypassing intermediate layers. The skipped layers form a **remaining block** (Figure 2.11). This process is easily expressed by the equation:

$$y = F(x) + x$$

where $F(x)$ and x represent the output and input of the residual block, respectively. This operation ensures a smoother gradient flow, enables the network to learn the identity mapping, and enhances the overall training efficiency [39].

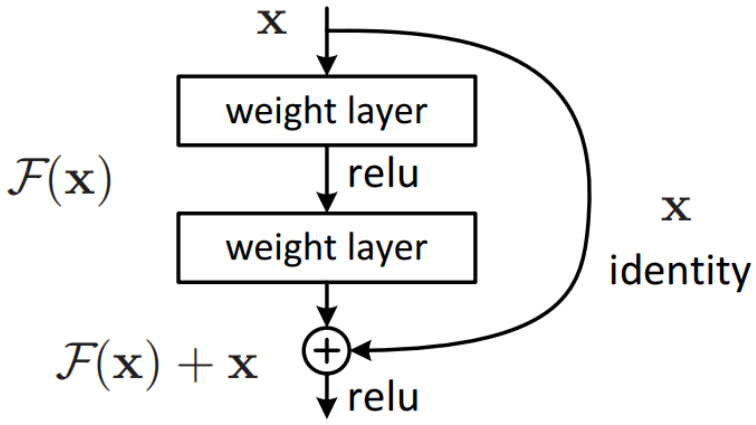

Figure 2.11: Skip connection simple scheme.

One more solution for improving the network's training efficiency is the **Batch Normalization**. This technique normalizes the output of each layer and scaling them to have a mean of zero and a variance of one. As a result, Normalization reduce internal covariate shift, ensuring that the distribution of inputs to each layer remains more stable throughout training [40].

A further technique for preventing overfitting is **Dropout**, which reduces the network's reliance on any single neuron by randomly deactivating a subset of neurons (and their associated connections) at each iteration [41].

Finally, in addition to classic gradient descent, **optimization algorithms** are commonly used to enhance both the convergence speed and stability of the training process. These algorithms adjust the model weights based on the gradient characteristics, computed during backpropagation, by dynamically adapting the learning rate to the gradient behavior. **Adam**, **RMSProp**, and **Adagrad** are popular examples of such optimizers.

2.4.4 CNN for Medical Images Analysis

Convolutional Neural Networks (CNN) are a highly efficient neural network architecture for image analysis, offering a reduction in computational complexity compared to fully-connected neural networks.

In a fully-connected neural network, each neuron is connected to every pixel in the input image. In contrast, CNN utilize neurons that are connected solely to a limited, localized region of the input image, referred to as the **"local receptive field"** (an $n \times n$ matrix of pixels). These neurons perform local feature analysis by applying a convolution operation, shifting the receptive field over different locations in the image (see Figure 2.12) . The map from the input layer to the hidden layer is called feature map.

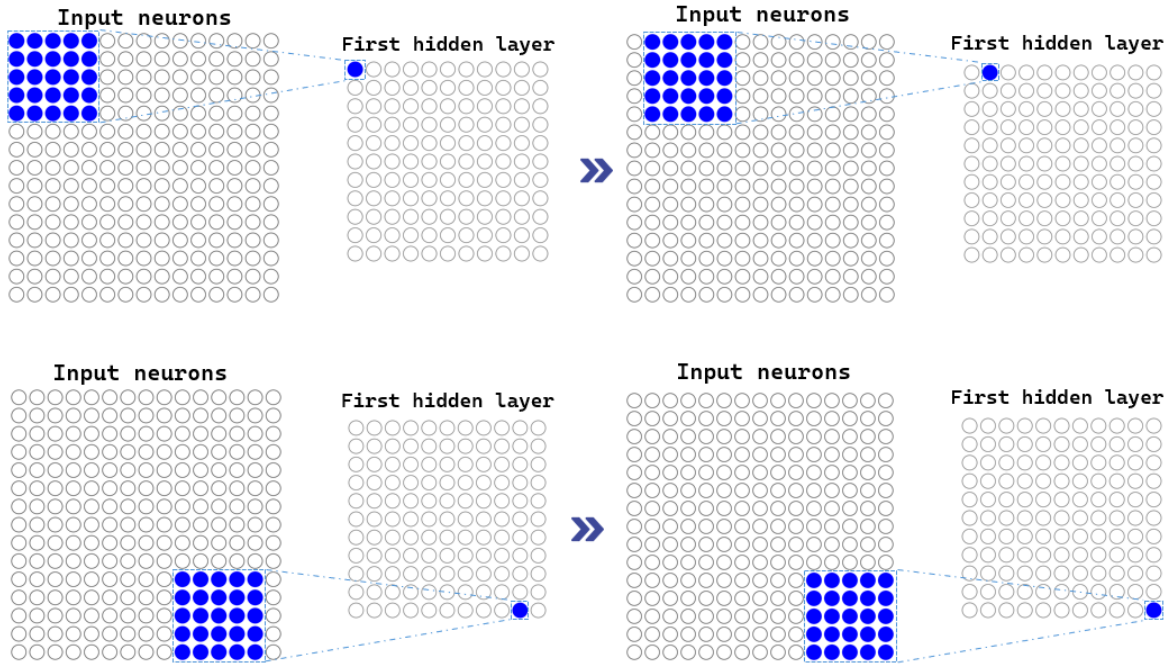

Figure 2.12: For enhanced visual representation, it is common practice to depict convolutional layers arranged in a two-dimensional grid. Each neuron in the hidden layer is only connected to the correspondent local receptive field.

In contrast to fully-connected networks, where each connection is assigned a distinct weight and bias, convolutional neural networks employ **shared weights** and **biases**. These shared parameters constitute a kernel that is applied uniformly across the entire local receptive field within an image. Consequently, each hidden neuron utilizes the identical set of weights as it traverses the image. For a neuron situated at position (i, j) within a given layer, the output is calculated according to the following equation:

$$y = f\left(b + \sum_{l=0}^{n}\sum_{m=0}^{n} w_{l,m}\, a_{i+l,j+m}\right) \tag{2.17}$$

where

- f is the neural activation function,
- b is the shared value for the bias,

- $w_{l,m}$ is a $n \times n$ array of shared weights,
- $a_{x,y}$ is the input activation at position x,y.

The advantage of weight sharing is a notable reduction in the number of parameters, which makes the model more computationally efficient and less susceptible to overfitting. Furthermore, this design choice confers translation invariance upon convolutional networks, enabling them to detect a given pattern in the same manner regardless of its position within the image.

Pooling layers are typically employed in conjunction with convolutional layers. Their function is to condense the feature maps summarizing a specific region of neurons from the previous layer in a single unit. The most common pooling technique is max-pooling, which selects the maximum value from each summarized region of the previous layer.

This operation reduces the size of the feature map while retaining the most important information. Pooling enhances the network's robustness to slight variations or distortions in the input image and further decreases the number of parameters, thereby improving the efficiency of the training process.

Convolutional layers are typically employed in multiple instances, with each layer generating a feature map for the corresponding hidden layer. This allows the network to learn increasingly complex and abstract features as the depth of the network increases.

The final layer of a CNN is a fully-connected layer of neurons, as many as the number of output classes, fully-connected with the penultimate layer.

For an illustrative scheme of a simple CNN see Figure 2.13.

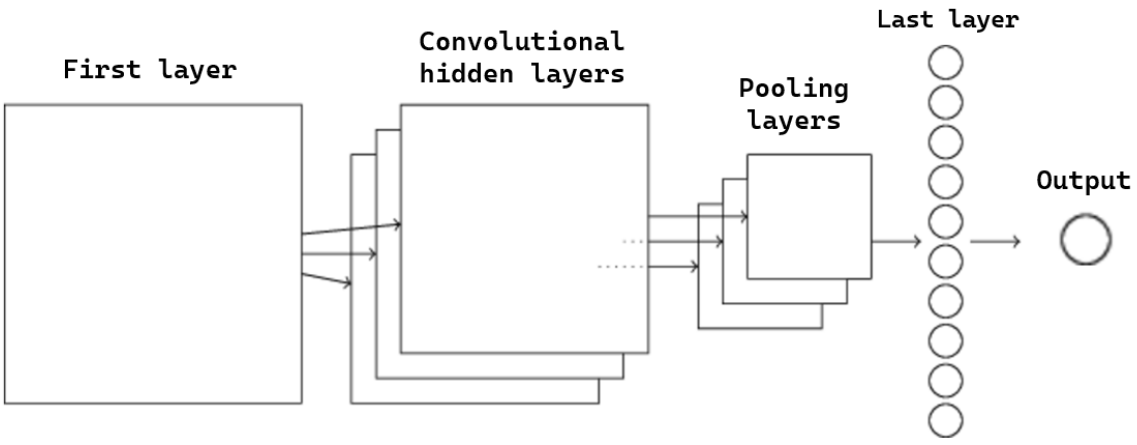

Figure 2.13: Basic architecture of a CNN [42].

U-Net

The U-Net, introduced by Ronneberger et al. in 2015 [43], is a U-shaped fully convolutional network and it is the most employed CNN for image classification or segmentation

in the medical field. Figure 2.14 presents a comprehensive overview of the U-Net architectural scheme and the operations performed by this network.

Nowadays, the backbones of models can vary significantly in their complexity, yet their basic structure always consists of a contracting path (**encoder**) and an expansive path (**decoder**). The encoder compresses the input into an embedding space through convolutions and pooling operations, effectively encoding the primary features of the image. The decoder performs a back projection of the data from the embedding space, reconstructing it to its original dimensions and recovering the fine details of the image that are lost during the encoding process.

The primary distinction from a traditional encoder-decoder architecture is the inclusion of **skip connections**, which serve to retain high-frequency information that could otherwise be lost during the projection phase.

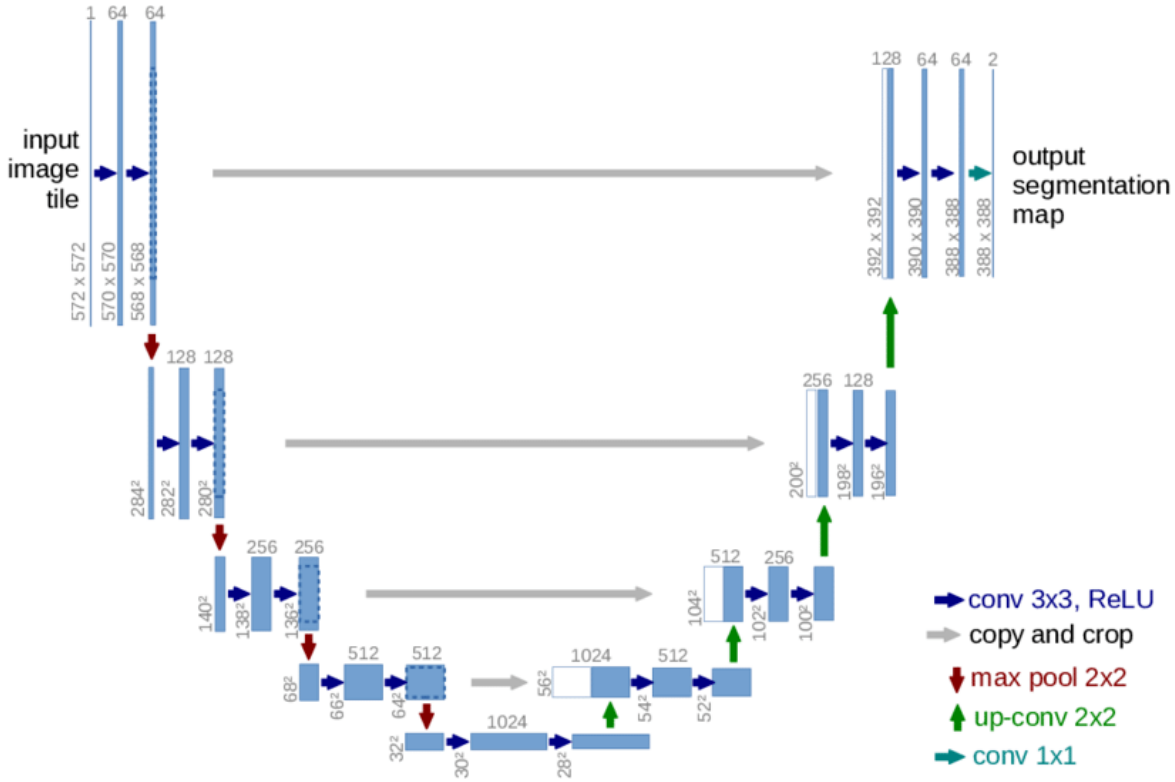

Figure 2.14: Original U-Net architecture for input image 572 × 572

2.5 A Review of Machine Learning for Wound Care

The use of AI in wound assessment and management has seen significant expansion in the last years, improving precision of wound assessment and management, sometimes filling the gaps of traditional treatment. Numerous studies have been, and continue to be, conducted worldwide to implement these methods in clinical practice. I will outline some of the key findings to illustrate the progress made in this field and to emphasize the importance of continuing to advance in this direction.

The greatest advantage of the Machine Learning algorithms is that bases its analysis on digital images, providing a non-invasive approach for patient and bringing more objectivity compared to traditional methods. The latter are, in fact, susceptible to inter-

operator variability, introduced by differences in experience or training, and intra-operator variability, caused by the psycho-physical state of the clinician (stress or fatigue) or by environmental conditions (e.g. illumination and exposure).

Machine Learning algorithms have been able to predict wound healing progress, in some cases surpassing the ability of clinical experts in identifying potential complications or monitoring treatment outcomes [44, 45].

During the COVID-19 pandemic, virtual care received an incredible boost, as it allowed remote patient monitoring, reducing the risk of contagion and minimizing the need for patient travel, and facilitated shared plans among providers. In the same period, the usability and effectiveness of an AI application for objective wound assessment were studied, demonstrating that its implementation in telemedicine can enhance efficiency and result in optimal wound care [46].

Predictive Machine Learning models were already used to analyze various risk factors, including medical history, lifestyle, wound characteristics, and other variables, to identify at-risk patients. These models have shown significant results in wound prevention, particularly for chronic wounds. An example is the model developed by Cross et al. to profile patients and predict the risk of developing chronic wounds, enabling clinicians to focus preventive measures on the most vulnerable patients, optimizing clinical intervention before the wounds become critical [47].

Moreover, the use of AI to predict deep tissue injury events, specifically pressure injuries on the heels, has shown a great potential, with methods achieving an accuracy of about 80%, suggesting the implementation of this technologies in the prevention of hospital-acquired ulcers [48]. Additionally, AI has shown significant potential in predicting deep tissue injuries.

Image segmentation is another area where AI, particularly Convolutional Neural Networks (CNN), has achieved excellent results, allowing precise identification and measurement of lesion areas.

For instance, ASURA (Automatic Skin Ulcer Region Assessment) framework, a deep learning model, reached an accuracy $> 90\%$ in automatic segmentation of skin ulcers and a relative error $\approx 12\%$ in estimation of ulcer area in real units (cm^2) [49].

Furthermore, in 2022 Curti et al. introduced **Deepskin** algorithm, a fully automated model for wound area segmentation that operates using images captured with smartphones [50]. This method, based on a CNN trained with an active semi-supervised learning strategy, is able to segment and extract masks for both wound and patient body from the images. Curti et al. also created a regression model based on textural features extracted from the wound and periwound areas, which predicts the PWAT with remarkable agreement to clinicians' estimations [51].

Machine Learning was also able to predict wound healing time analyzing sets of temporal images of the same wounds. Additionally, networks that incorporate features of tissue types (non-wound, fibrin, granulation, and necrotic) within the wound exhibited superior accuracy in estimating wound healing compared to networks that did not incorporate these features [52].

Chapter 3

Materials and Methods

The chapter will present the wound image dataset used for this thesis, along with the applied processing methods. The Python scripts used for the analysis have been included into the Deepskin GitHub repository (https://github.com/Nico-Curti/Deepskin).

Briefly, I developed an unsupervised learning model that clusters wound images based on features extracted from the periwound area, specifically focusing on the behavior of mean depth profiles from the wound interior to the surrounding skin.

This approach aims to evaluate whether the obtained clustering aligns with scoring methods used in clinical practice, particularly BWAT, and to eventually uncover patterns that could assist clinicians in classifying wound edges.

3.1 Description of Dataset

The dataset of images was acquired during clinical practice at the IRCCS Sant'Orsola-Malpighi University Hospital in Bologna. It consists of 7329 ulcer images, collected from March 2019 to September 2024, but it is daily updated by clinicians with new images.

The dataset includes samples with ulcers at different healing stages and in various anatomical positions. All the photos were simply taken using a smartphone camera and were in a RGB 8-bit format, saved as PNG with a resolution of 1440×1080.

The method proposed in this thesis analyzes data in a novel way; therefore, no standardized protocols for image acquisition were established. The lack of guidelines allowed the identification of the conditions under which the method performs most efficiently.

The number of images available for analysis from the dataset was 489. This reduction is explained by the need for time-consuming manual validation of both wound and body masks, which led to an initial rough filtering of the images. Additionally, images of wounds that were either too small or fully healed, or had inaccurate depth maps were excluded from the dataset.

In this study was used a subset of 159 wound images, which were labeled for the wound edges three times. The size of the dataset was chosen in consultation with the clinician to allow for a quick yet careful labeling process, balancing speed with the need for accurate annotations. Two classifications were performed by a dermatologist with a one-month interval, while the third classification was done by another clinician with more than 3 years of experience in wound management. This allowed to evaluate the model's results through comparison, as well as an assessment of both inter- and intra-operator contingency.

A summary of the edge classifications is provided in Table 3.1. Note that the used edge labels are consistent with BWAT edge scoring (see Table 1.1).

LABELS	**Indistinct**	**Distinct**	**Well-defined**	**Rolled under**	**Fibrotic**
BWAT score	**1**	**2**	**3**	**4**	**5**
DERM	24	41	38	45	11
DERM-1M	21	42	47	32	17
CO	14	47	52	30	16

Table 3.1: Summary of wound edges classification. **DERM** and **DERM-1M** refer to the first and one-month post-classification by the dermatologist. **CO** refers to the classification made by the clinical operator.

3.2 Image Segmentation

Image segmentation is a technique in digital image processing and computer vision that divides an image into distinct regions or segments based on pixel characteristics such as color, shape, or texture. This process helps in object detection and analysis by separating meaningful areas. In the context of digital imaging wound assessment, this implies identifying the wound ROI (Region Of Interest).

3.2.1 Deepskin

Deepskin is an open-source library available on GitHub [53], which contains the Deepskin algorithm, a segmentation model that performs multi-class semantic segmentation. The model segments the image into three regions: the wound ROI, the patient body ROI, and the background ROI.

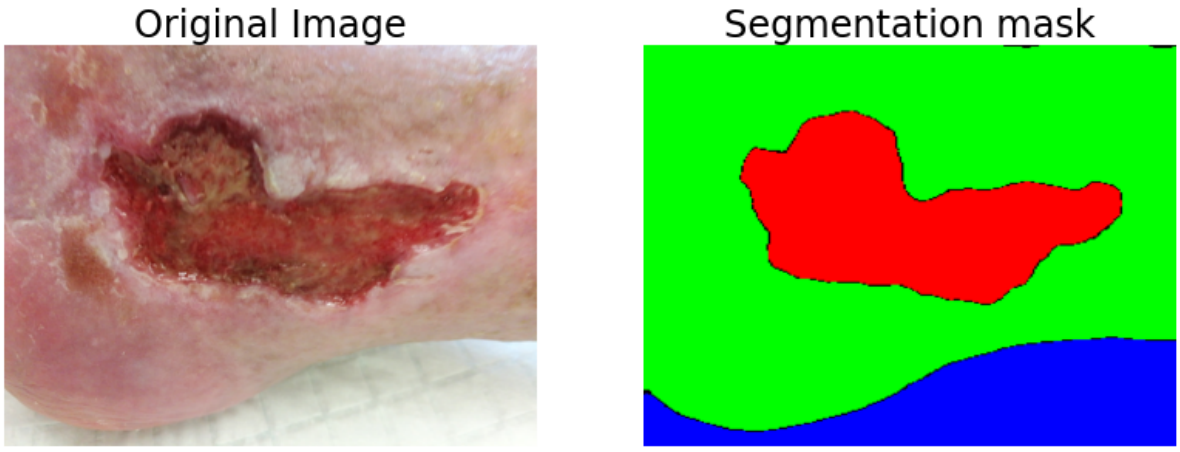

Figure 3.1: Example of Deepskin segmentation. On the left, the original image; on the right, the segmented ROIs: wound (red), patient body (green), and background (blue).

All images in original dataset were processed with Deepskin to extract the wound mask and the body mask. Since the masks were manually validated, an initial rough selection was made, keeping only around 1500 segmented images to facilitate the process. Subsequently, about 500 images were discarded during further validation, leaving only the highest-quality segmentation for subsequent analysis.

3.2.2 Selection of the Largest Mask Component

In cases of multiple wounds or when there are separations between different portions of damaged tissue, the wound mask generated by Deepskin may present several separate components. This can pose a problem in the analysis, as in clinical practice, only the largest wound component is assessed.

In order to address this issue and align with clinical assessment practices, the component with the greatest number of pixels was selected (see Figure 3.2), as it corresponds to the largest wound under the assumption of no substantial image distortion. The selected component was then stored and substituted for the original wound mask extracted by Deepskin for all subsequent analyses.

Only images wherein the largest mask component exceeded 5000 pixels were selected for analysis; hence, about 150 images were excluded.

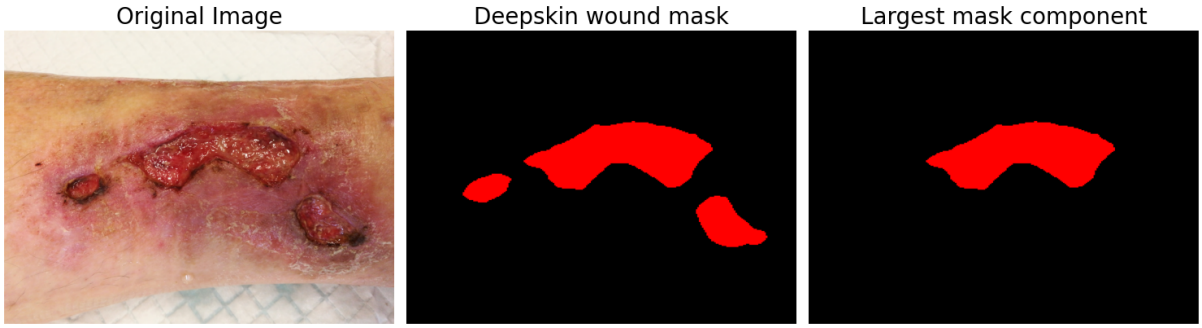

Figure 3.2: Example of multi-components wound. On the left is the original image, in the middle is the wound mask extracted by Deepskin, and on the right is the largest component, which is used as the mask for the analysis.

3.3 Depth Maps Computation

In computer vision, a **depth map** is an image or image channel that contains information about the distance of surfaces of scene objects from a viewpoint. In a depth map, each pixel contains a value indicating how far that point is from the camera. In the context of digital imaging, depth maps are typically stored in uint16 format, meaning each pixel value is represented as a 16-bit unsigned integer. This format allows for a wide range of discrete depth values, from 0 to 65535, providing high precision.

Since the changes in depth along the wound border is needed to evaluate wound edges (see subsection 1.1.1), accurate depth maps of the wound were essential in the analysis.

3.3.1 ZoeDepth

ZoeDepth is an open-source library available on GitHub [54]. It is based on a deep learning model and CNN that estimates the pixel depth in a photo, allowing for the generation of a depth map from a 2D image.

ZoeDepth offers the possibility of using three different pre-trained models, each optimized for a specific type of image:

- **ZoeD_K**: for indoor scenes.

- **ZoeD_N**: for images of people.
- **ZoeD_NK**: for general-purpose applications.

The **ZoeD_NK** model was employed to extract depth maps of the wounds, which were subsequently saved as PNG files. The model was not re-trained and was only used for validation. Despite being trained on natural images and never having been applied to wound images before, the algorithm performed remarkably well on almost all images in the dataset (example in Figure 3.3).

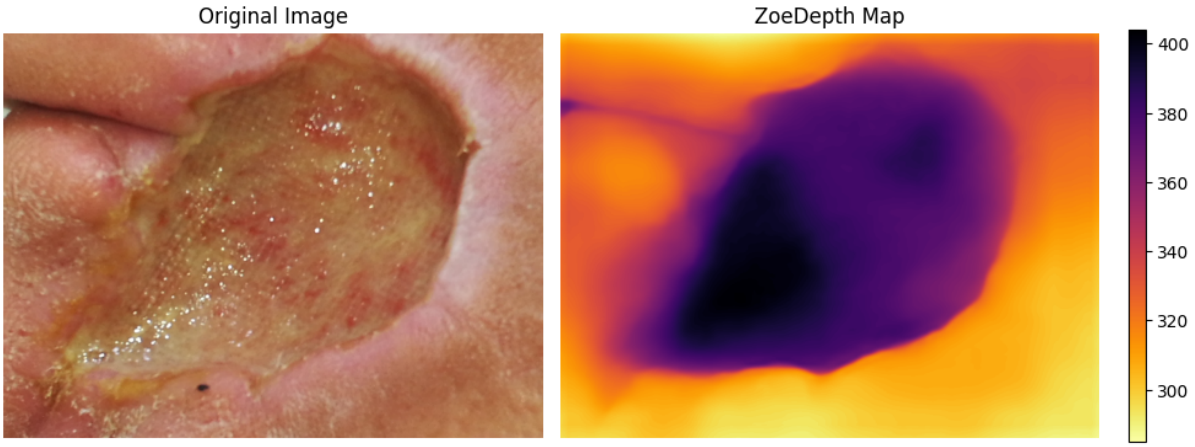

Figure 3.3: On the left is the original image, and on the right is the depth map computed by the ZOED_NK model, presented using the a colormap. It can be observed that the wound depth is accurately represented.

The biggest challenge in applying ZoeDepth to wound images was the lack of a standardized protocol for photo acquisition in the dataset. ZoeDepth performed exceptionally well on photos where the wound bed was positioned perpendicular to the acquisition axis, but it struggled to accurately estimate wound depth when the bed was tilted along that axis. Some issues were also encountered for very blurred images or in the presence of certain artifacts, such as intense light reflections, which result in inaccurate depth estimates.

Not all of the images presenting these issues (about 400) were discarded; some were restored using the correction explained in Section 3.5.

3.4 Wound Border Rectification

The main idea of **border rectification** is to take the periwound region, "cut" and "unroll" it to be represented as a flattened, rectangular strip (as seen in Figure 3.4). This process allows the visualization of wound edges area in a standardized format that does not depend on the wound shape.

The rectification was applied to both the RGB image and the depth map of the wounds, resulting in the extraction of the **"rectified border"** from the RGB image and the **"rectified depth border"** from the depth map.

Although the RGB image (in uint8 format with 3 color channels) and the depth map (in uint16 format with a single channel) differ in format, the processing approach remains fundamentally the same for both and can be broken down into the following steps:

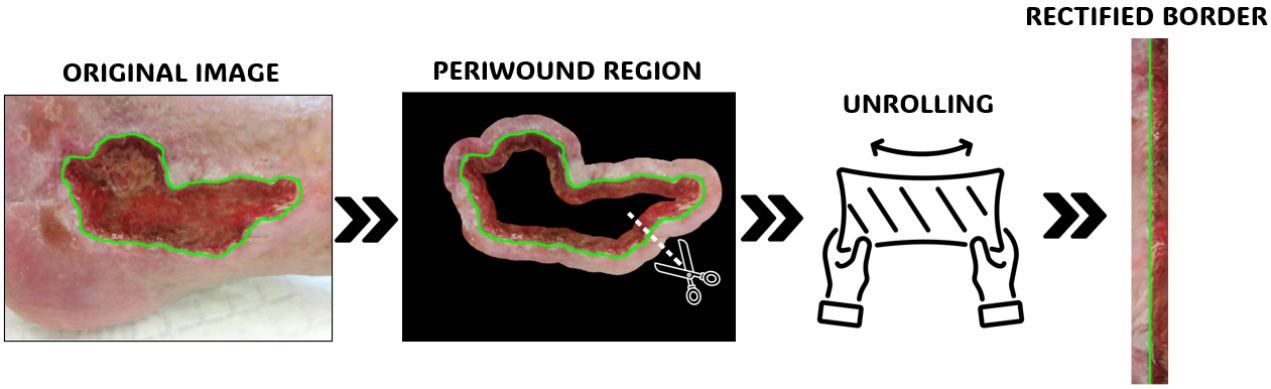

Figure 3.4: Conceptual idea of border rectification: identification of the wound border (green); extraction of the periwound region as a "strip" to be "cut"; unrolling of the strip.

1. **Initial Contour Extraction**: The boundary contour is extracted from the wound mask, and the corresponding pixels are sampled from the original image or depth map; then they are arranged into a column of pixels.

2. **Dilations and Erosions**: An elliptical morphological element was used as kernel for both dilation and erosion of wound mask. At each iteration, the size $(height, width)$ of the element increased linearly as (i, i), where i is the current iteration number. The number of iterations was set to 150.

   The dilations expand the wound boundary outward, capturing the surrounding periwound area.
   The erosions shrink the wound boundary inward, ensuring to cover a region within the wound itself.

   The process operate under two constraints:

   - Dilations must not exceed the body mask, preventing the expansion from going beyond the patient's body region.
   - Erosions stop if the mask is completely eroded.

   If a constraint is violated, the process terminates, considering only the valid iterations while maintaining an equal number of dilations and erosions.

3. **Contour Sampling**: The boundary contour is extracted from each dilated and eroded mask, and the corresponding pixels are sampled from the original image or depth map. These pixels are then arranged into a column and resized to match the size of the initial contour, ensuring a consistent shape across all segments.

4. **Rectified Image Construction**: The sampled strips are finally sorted from innermost (eroded) to outermost (dilated), creating a flattened representation of the wound borders.

3.5 Depth Map Correction

The wound may lie on a curved body surface, and therefore, even though the depth map is correct, it does not necessarily aid in accurately rectifying the depth borders, because in some important regions an offset can be introduced.

For this reason, it was deemed appropriate to correct this offset in all the images by fitting a model to the patient's body, excluding the wound ROI from the body depth map. A simple quadratic fit (Equation 3.1) was used because it effectively models the natural curvature of the healed body (without the wound), accounting for variations in body shape:

$$z_{fit} = ax^2 + by^2 + cxy + dx + ey + g \tag{3.1}$$

where:

- z_{fit} is the depth value calculated by the fit,
- a, b, c, d, e, g are numerical coefficient calculated by the fit,
- x, y are the coordinate of the PNG image.

The fit prediction were then subtracted from the depth map. The minimum value of the difference map was shifted to zero before saving the map in uint16 format, as PNG.

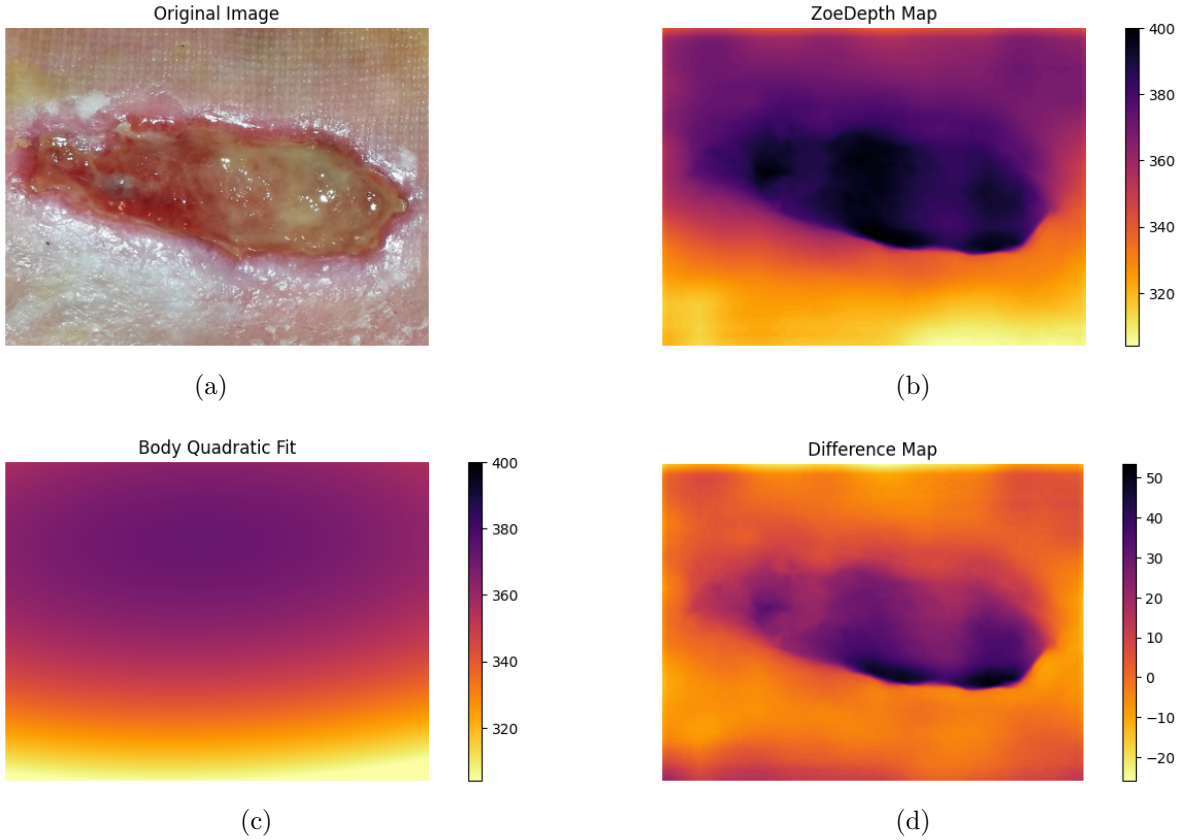

Figure 3.5: (a) Original image; (b) Depth map computed by ZoeDepth; (c) Quadratic fit; (d) Difference between the fit and the ZoeDepth map.

This approach maintained the depth estimation in the already accurate depth maps, improved the reliability of those where the wound lay on a curved surface, and corrected some of the maps from images that had a slight tilt in the photo axis. An example of is shown in Figure 3.5 and Figure 3.6.

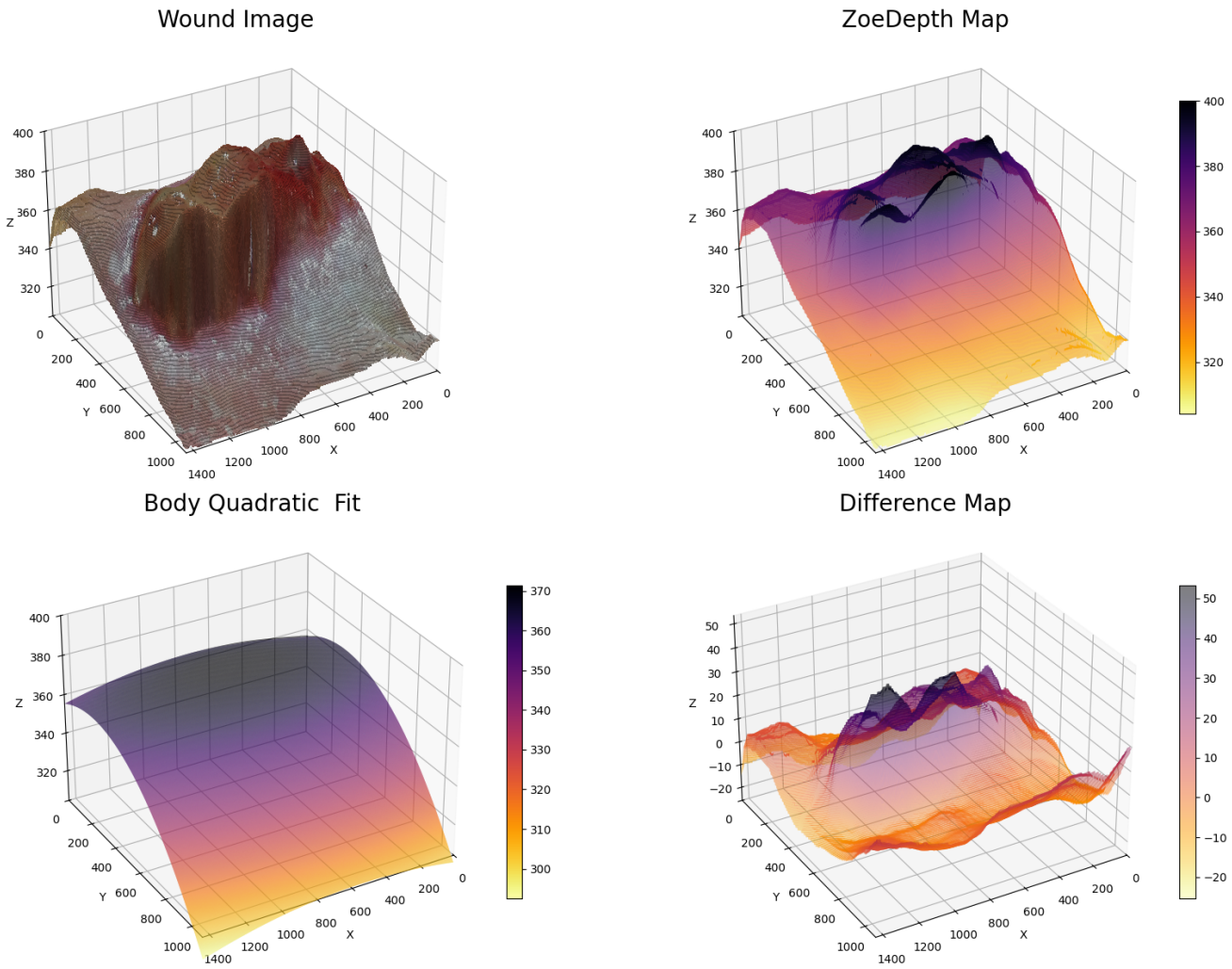

Figure 3.6: 3D representation of the wound data: (top left) original wound image; (top right) ZoeDepth map; (bottom left) quadratic fit; (bottom right) difference between the quadratic fit and the ZoeDepth map.

3.6 Mean Depth Profile

According to the construction of the rectified depth border, the variation in values along its horizontal axis reflects the change in depth across the wound edges.

The mean depth profile along this axis was extracted for each rectified border, providing a simplified representation of the depth transition from the inside to the outside of the wound. Additionally, the standard deviation (STD) profile was used to assess the variability and robustness of the mean depth.

All analyzed rectified depth borders have a minimum width, and so a minimum profile length, of 60 pixels.

It was observed that mean profiles mainly follow three distinct trends, which can be easily identified by the human eye: a horizontal line, a straight line with some inclination, and an inverse sigmoid curve. An example of these behaviors is presented in Figure 3.7.

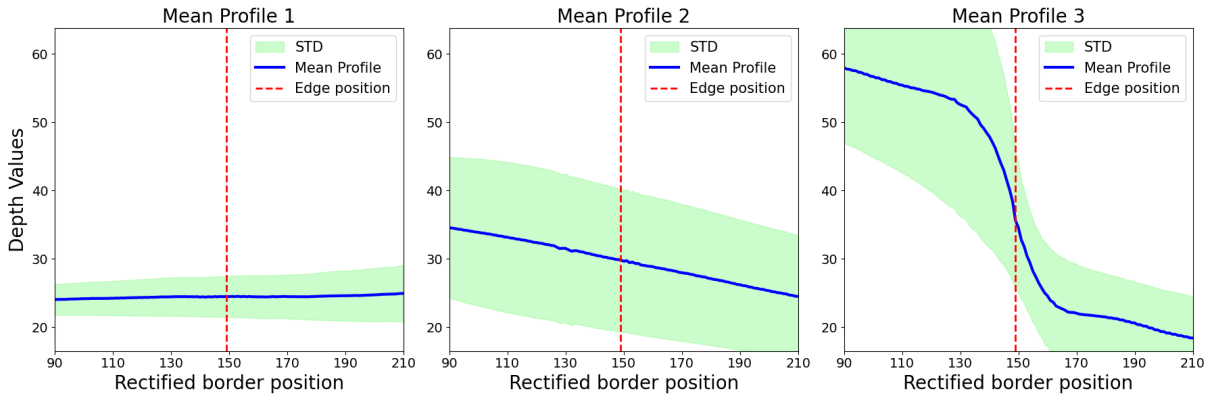

Figure 3.7: Mean depth profiles with standard deviation (STD) for the three different trends.

3.7 Features extraction

The depth profile features were extracted from the mean and STD profiles through a combination of **statistical**, **spectral**, and **correlation** methods (a summary is presented in Table 3.2).

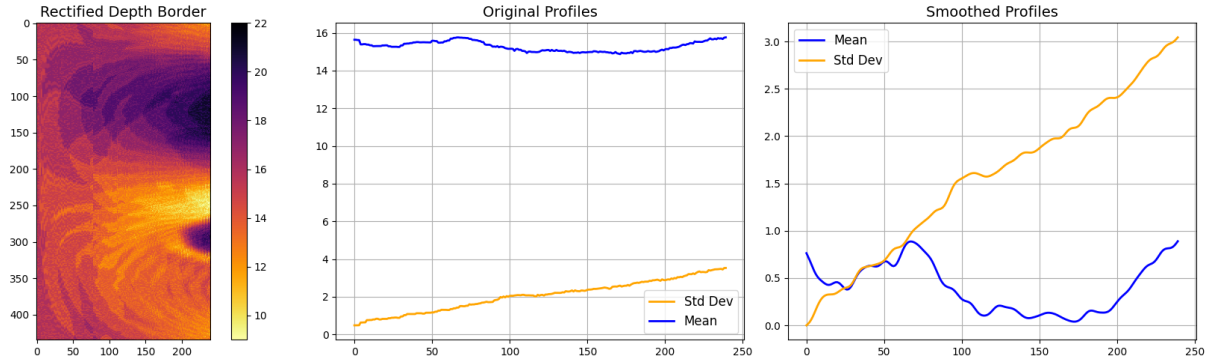

Figure 3.8: On the left, a rectified depth border. In the middle, the original depth profiles. On the right, the profiles the baselining procedure and high-frequency filtering

Prior to feature extraction, the profiles underwent a baselining procedure to establish a consistent reference level. Subsequently, a Butterworth low-pass filter (order 4, cutoff frequency of 0.1) was applied to smooth the profiles, effectively mitigating high-frequency noise (Figure 3.8). This process allowed for more regular spatial profiles, less influenced by rapid fluctuations, thereby improving the quality of spectral analysis and the stability of correlations.

During the feature extraction, the mean profiles were fitted with linear and power law functions (using the Least Squares method), as these models appeared to best capture the observed trends in the data. The fitting process was applied by systematically scanning the central region of the profile with a moving window. Specifically, the 80 central values of the profile (or fewer, if less than 80 were available) were considered. A moving window of various sizes was applied and shifted across this region, and the best fit for each window size was selected based on the highest R^2 value.

Linear fits were performed using window lengths of 10, 20, and 40 pixels, as well as

across the entire central region. Power law fits used intervals of 20 and 40 pixel lengths, and the entire central region.

In the case of power law fitting, x_{min}, the starting point of the fit, was selected to account for potential shifts in the profiles, particularly in cases where the profiles resembled a sigmoid shape.

For both mean and STD profiles, spectral analysis was performed using two main methodologies: **Welch**'s spectral analysis and **Fast Fourier Transform** (FFT). Welch's method was applied to the unsmoothed profiles, enabling the analysis of global features of the spatial signal and highlighting dominant frequencies and their variations. Conversely, FFT was applied to the smoothed profiles to obtain a more stable and less noise-affected spectral representation.

Finally, correlations were performed between the normalized spatial profiles and theoretical functions such as the sigmoid, inverse sigmoid, and horizontal line.

In total, 68 features were extracted; however, parameters whose distributions had excessively low or high standard deviations, as well as those with constant values, were excluded, leaving a total of 55 features for further analysis.

Category	**Method**	**Features**	**Profiles**
Statistical	Linear Fit	Slope Intercept R^2	Smoothed Mean
	Power Law Fit	Exponent α Alpha error σ Constant C R^2 Fit starting point $x_{\min}$	
Spectral	Welch	Dominant Frequency Mean Amplitude Spectral Variance	Mean and STD
	FFT	Centroid Spectral Entropy Maximum Value	Smoothed Mean and STD
Correlation	Sigmoid	Maximum correlation	Normalized Mean and STD
	Inverse Sigmoid	Shift of maximum correlation	
	Horizontal Line	Mean correlation	

Table 3.2: Summary of the extracted features from the depth profiles.

Chapter 4

Results

This chapter presents the results of the analysis of clinician agreement in the classification of wound edges. Furthermore, it explores the application of clustering techniques and supervised machine learning, both based on the extracted features from the depth profiles, with the aim to improve the accuracy and efficiency of wound edge classification.

4.1 Clinician Agreement in Edges Classification

The evaluation of clinician agreement in the edges assessment was conducted by comparing the three different classifications of the analyzed set of images: DERM, DERM-1M, CO. DERM and DERM-1M were performed by a dermatologist, with the second one conducted one-month later. CO was made by an expert clinical operator with more than 3 years of experience.

DERM demonstrated concordance with DERM-1M in approximately 39% of the assessed images. In comparison, CO showed agreement with DERM and DERM-1M in approximately 60% of the cases for both assessments. All three classifications agreed on 53 images ($\approx 33\%$).

The calculation of the Intraclass Correlation Coefficient (ICC) across the classification comparisons revealed reliability levels ranging from poor to good (see Table 4.1).

Figures 4.1 to 4.3 show the normalized co-occurrence matrix for each comparison.

It is evident from these results that it is common for two different clinicians, or the same clinician at different times, to classify wound edges as belonging to distinct yet adjacent categories (e.g., Indistinct - Distinct or Distinct - Well-defined). Furthermore, disagreements may also occur between more distantly related categories, particularly when the Fibrotic class is involved. This indicates variability in the interpretation of wound edges classes, as expected from the calculated ICC.

Comparison	% Concordance	ICC	CI95%
DERM vs DERM-1M	38.9	0.46	[0.33, 0.57]
DERM vs CO	60.38	0.79	[0.73, 0.84]
CO vs DERM-1M	59.12	0.68	[0.58, 0.75]

Table 4.1: Summary of concordance and ICC for each classification comparison.

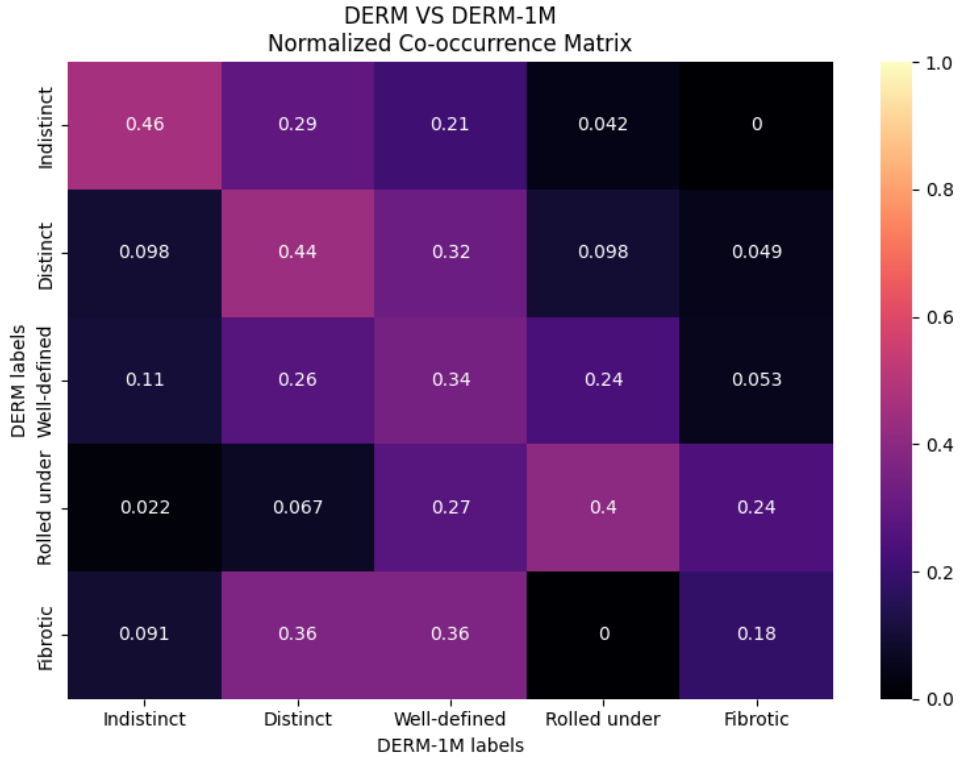

Figure 4.1: Normalized co-occurrence matrix showing the alignment between DERM and DERM-1M classifications.

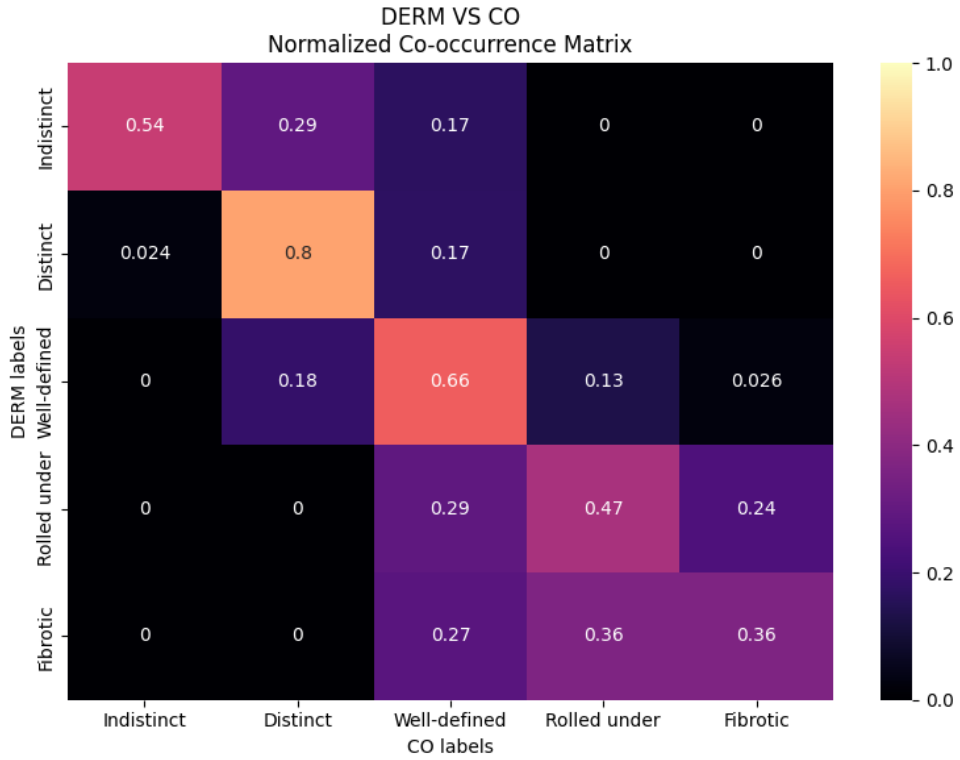

Figure 4.2: Normalized co-occurrence matrix showing the alignment between DERM and CO classifications.

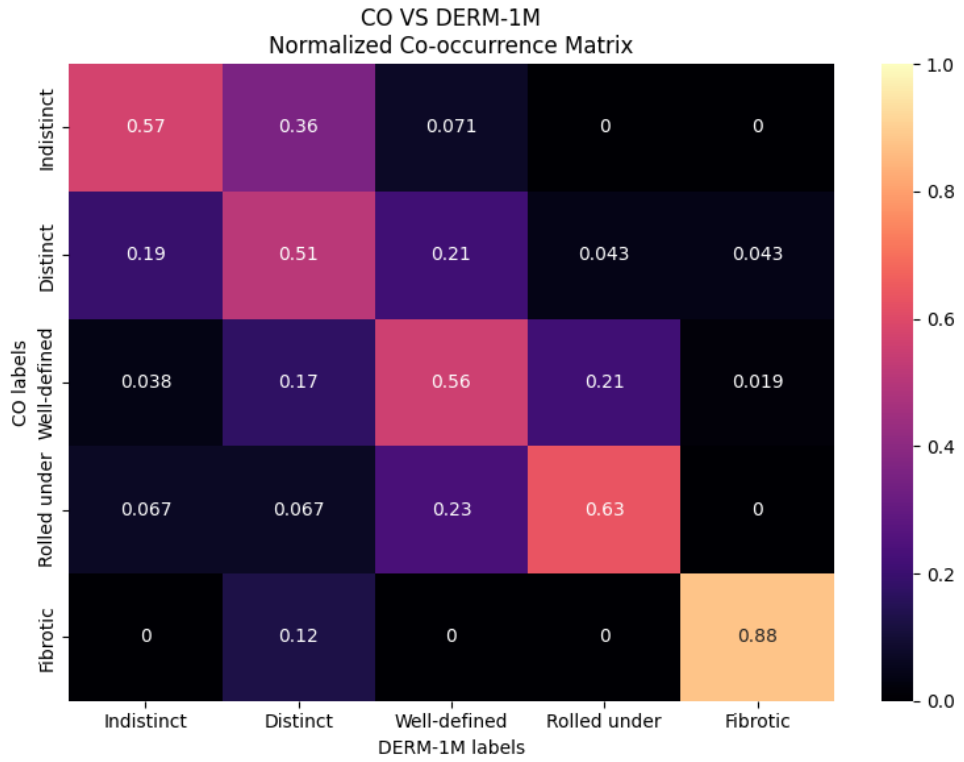

Figure 4.3: Normalized co-occurrence matrix showing the alignment between CO and DERM-1M classifications.

Following consultations with clinicians, certain labels were grouped together because they acknowledged that there can be variability in classifying closely related classes. In particular, the labels Indistinct were merged into Distinct, and Rolled under were grouped with Fibrotic.

These adjustments naturally increase concordance in the comparisons (see Table 4.2) as well as the number of images equally labeled across the three classifications (79, approximately 50%). However, noticeable disagreements still persist when the third group (which includes the Fibrotic) is involved.

Figures 4.4 to 4.6 show the normalized co-occurrence between clinician classification using the grouped labels.

Comparison	% Concordance	ICC	CI95%
DERM vs DERM-1M	52.83	0.18	[0.02, 0.32]
DERM vs CO	74.84	0.56	[0.45, 0.66]
CO vs DERM-1M	67.92	0.47	[0.34, 0.58]

Table 4.2: Summary of concordance and ICC for each classification comparison merging the Distinct to the Indistinct labels and the Rolled under to the Fibrotic labels .

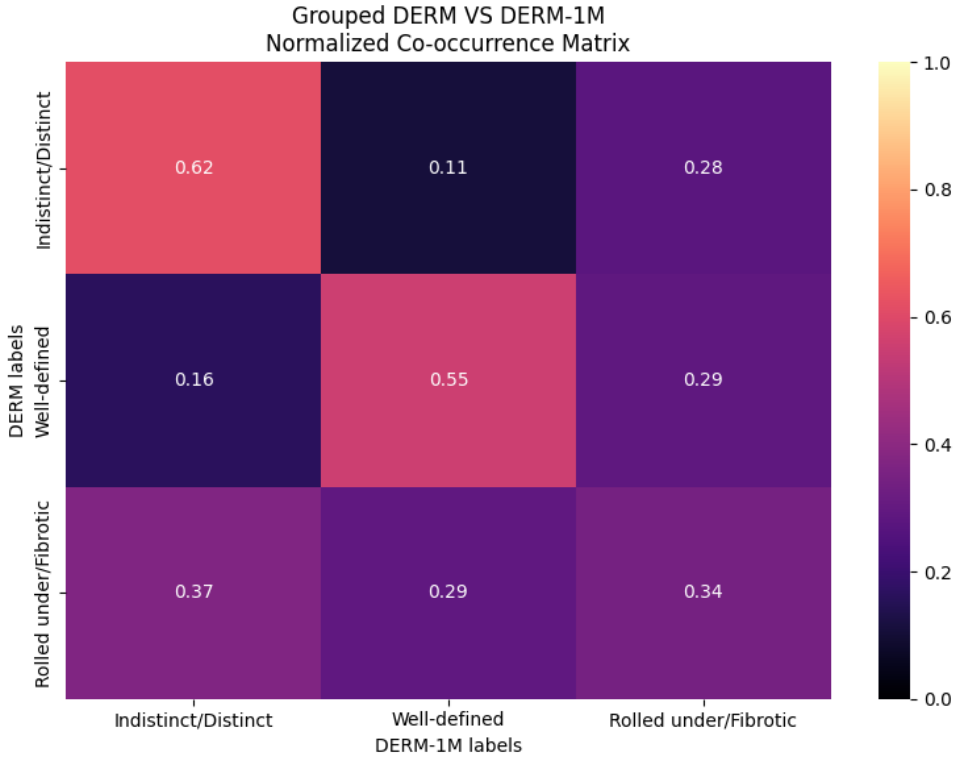

Figure 4.4: Normalized co-occurrence matrix showing the alignment between DERM and DERM-1M classifications merging labels.

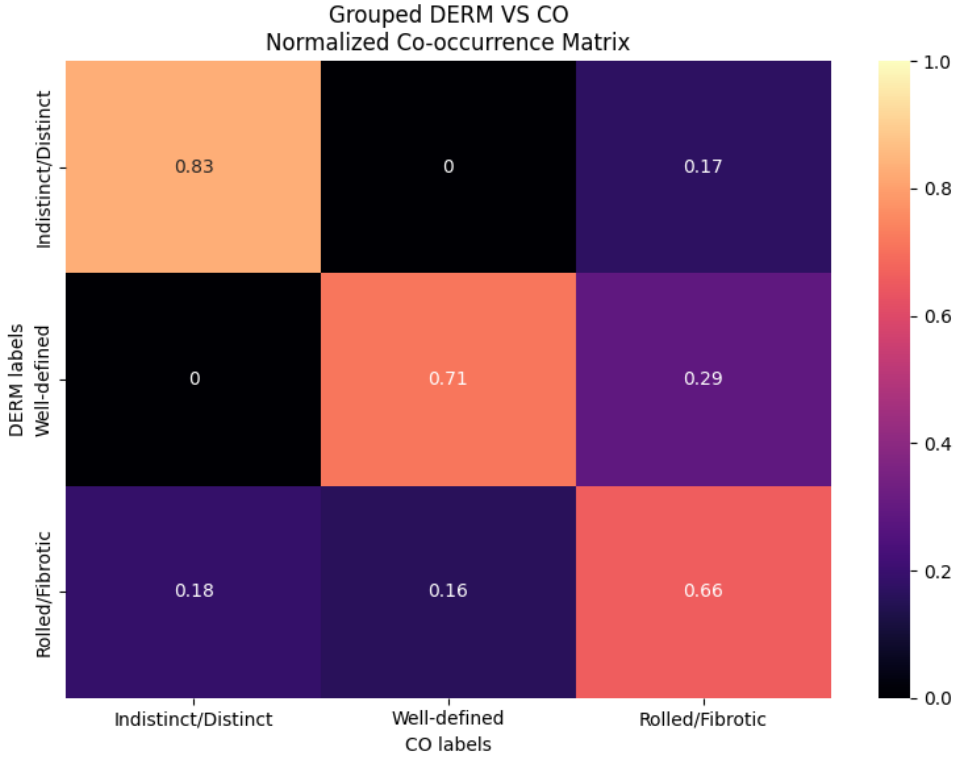

Figure 4.5: Normalized co-occurrence matrix showing the alignment between DERM and CO classifications merging labels.

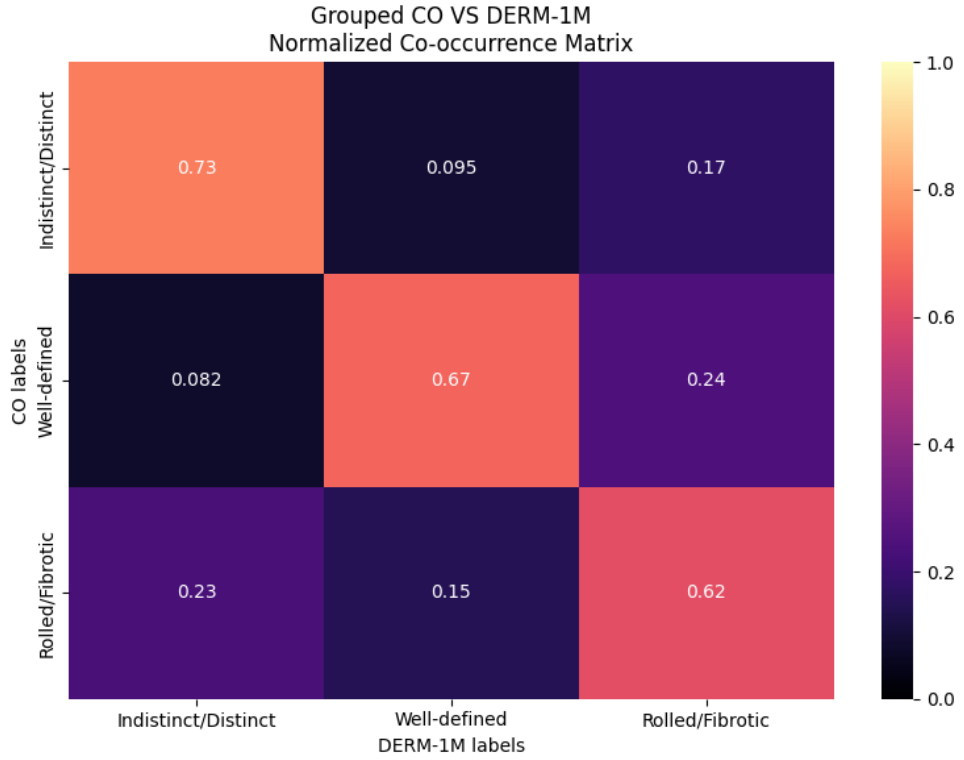

Figure 4.6: Normalized co-occurrence matrix showing the alignment between CO and DERM-1M classifications merging labels.

4.2 Wound Edge Clustering

4.2.1 Features embedding

The features extracted from the depth profiles were used to cluster images based on their edges. Given the large number of selected features (55), the first step involved embedding the data into a two-dimensional space using PaCMAP, with an appropriate fine-tuning of the parameters.

The following PaCMAP configuration was used:

```
pmap = pacmap.PaCMAP(
  n_components=2,
  n_neighbors=None,
  MN_ratio=5,
  FP_ratio=17,
  random_state=42,
  distance= 'euclidean'
  )
```

The resulting distribution in the new feature space is depicted in the Figure 4.7.

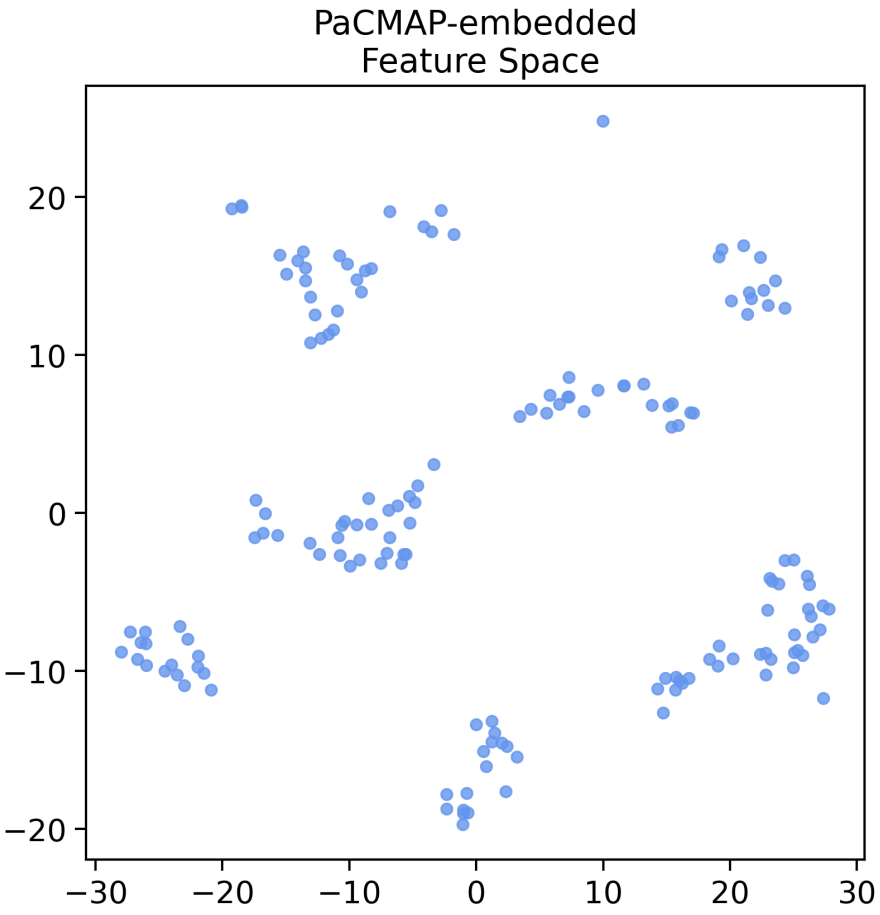

Figure 4.7: PaCMAP embedding of depth profiles features into a two-dimensional space.

4.2.2 Cluster Identification

The two-dimensional representation of the feature space was subsequently analyzed using **HDBSCAN** (Hierarchical Density-Based Spatial Clustering of Applications with Noise), which is particularly suited to real-world datasets, such as those involved in medical image clustering. The adopted HDBSCAN configuration was the following:

```
cl = hdbscan.HDBSCAN(
    min_cluster_size=5,
    min_samples=6,
    cluster_selection_epsilon=.75,
    cluster_selection_method='eom',
    core_dist_n_jobs=-1,
    metric='euclidean',
    )
```

In total, 8 clusters were identified, including noise (see Figure 4.8). The count of points across these clusters is summarized in Table 4.3.

To explore the variability of individual features within the embedded space, each feature was plotted over the 2D space, showing its values within the dataset. The resulting visualizations (Figures 4.9 and 4.10) provide insights into the contribution of each feature to the clustering.

Notably, the slope and R^2 from the linear fit tend to differentiate cluster 6, the group comprising clusters 1, 3, and 4, and the group containing clusters 0, 2, and 5, while

Cluster	Number of Points
-1	1
0	17
1	36
2	27
3	12
4	20
5	29
6	17

Table 4.3: Count of points across the identified clusters. Cluster -1 corresponds to the noise points.

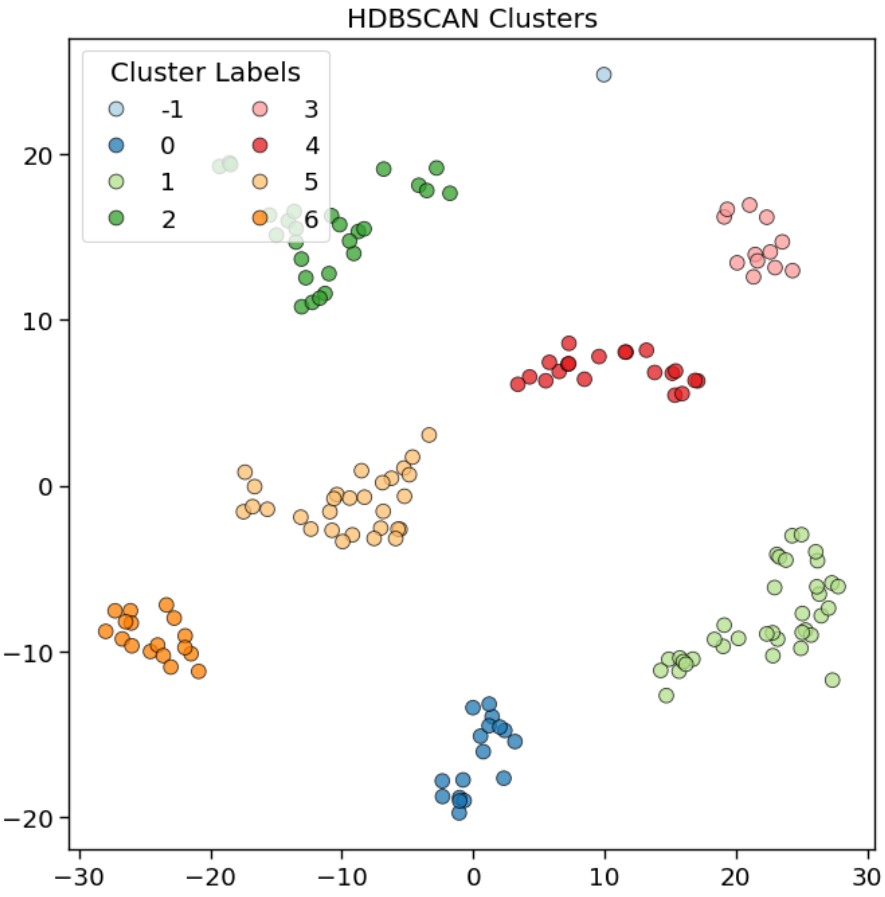

Figure 4.8: Visualization of the clusters identified by HDBSCAN in the embedding space.

showing limited variation within each subgroup.

Similarly, the α, $x_{\min}$, and R^2 of the power-law fit show significant differences between cluster 6, the group formed by clusters 1 and 3, and the remaining clusters.

Certain spectral variables, specifically the Fourier centroid and Fourier entropy of the mean profile, highlight cluster 3 as distinct.

Finally, among the correlation parameters, the shift of the mean profile associated with sigmoid and reverse sigmoid models varies between the groups 1 and 3 and the rest, while the standard deviation shift is primarily distinct for clusters 0, 1, 3, and 4 compared to the others.

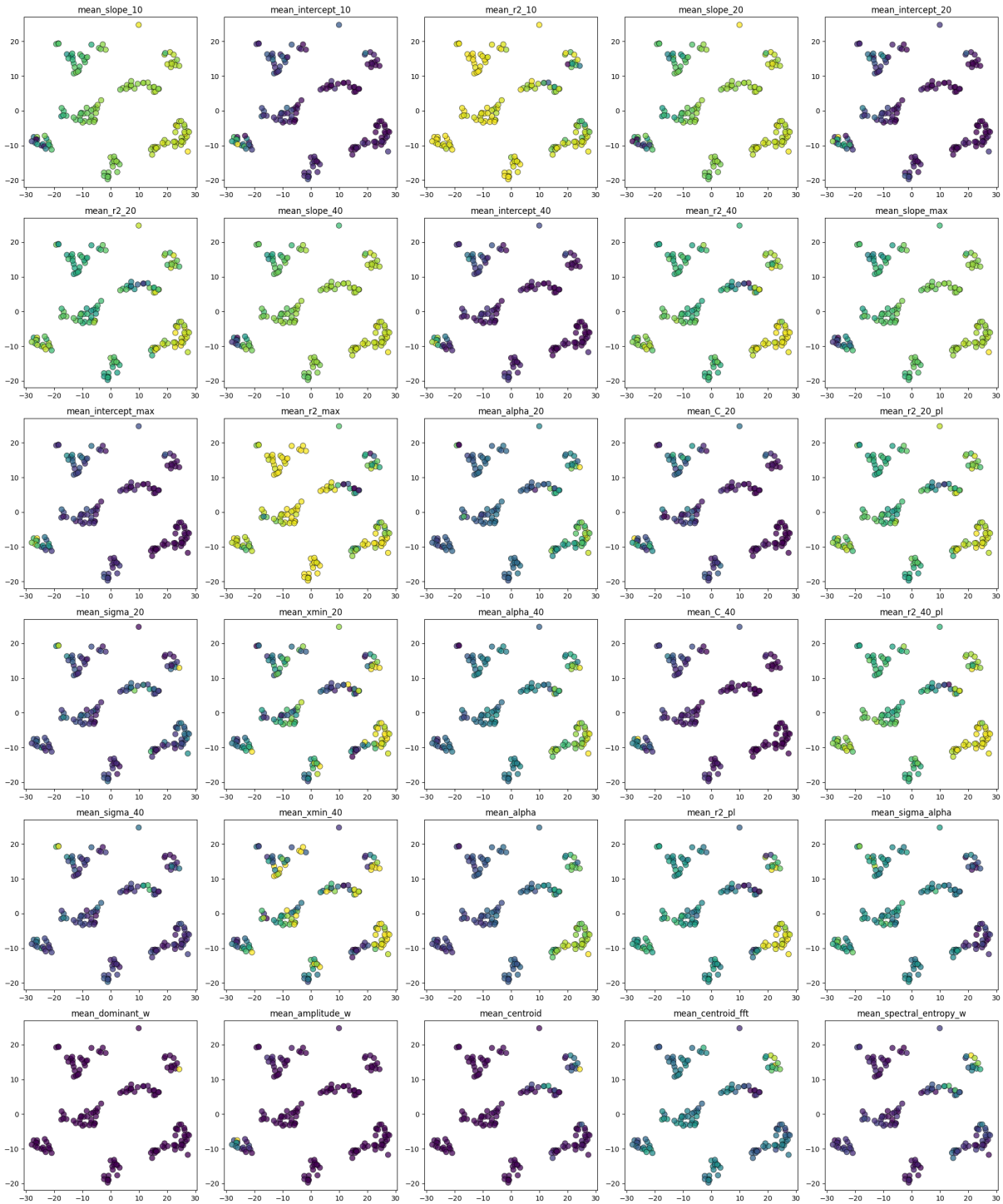

Figure 4.9: Visualizations of 30 individual feature distributions within the 2D embedded space. The color variation represents the change in the feature value across the dataset.

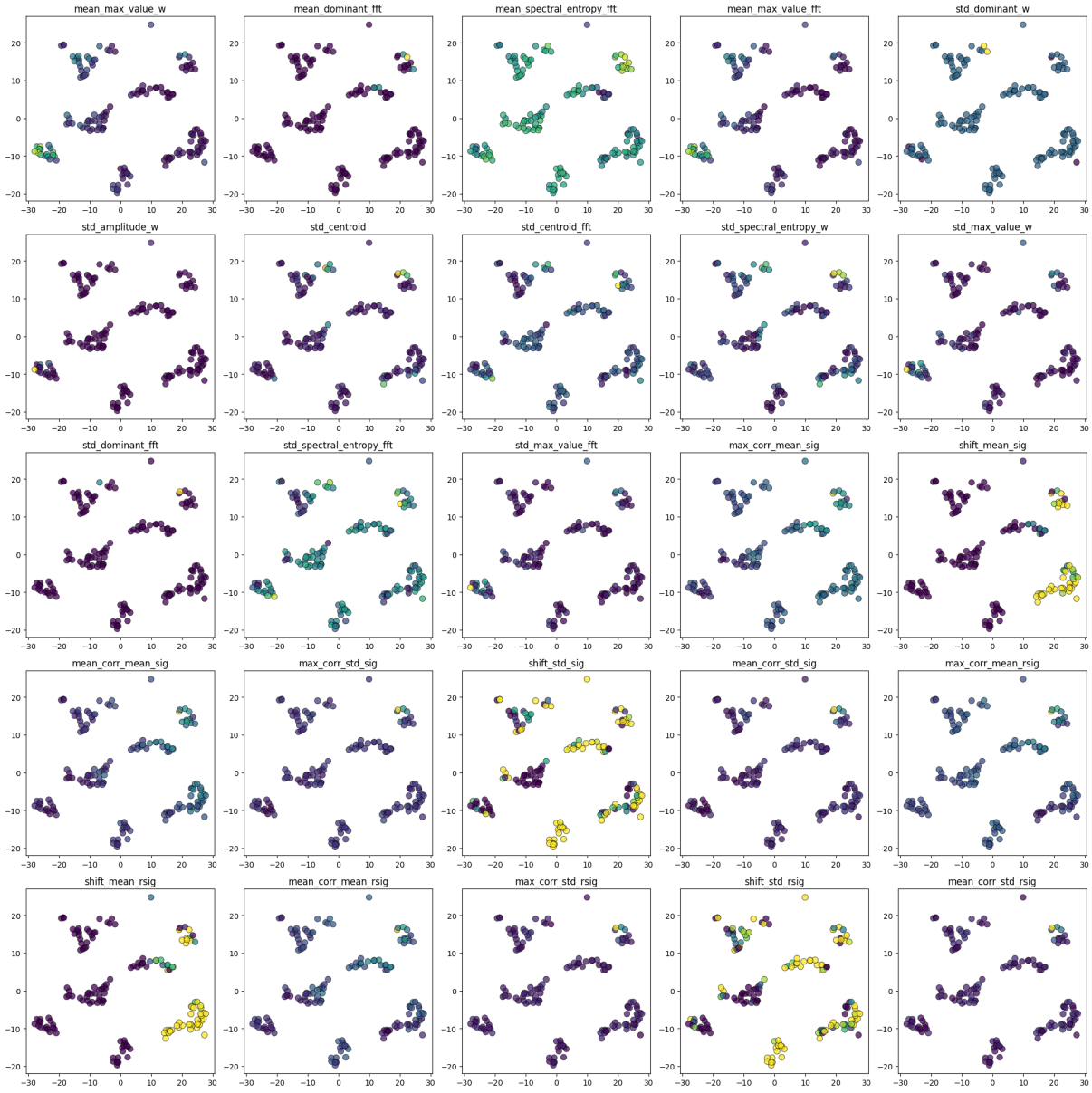

Figure 4.10: Visualizations of 25 individual feature distributions within the 2D embedded space. The color variation represents the change in the feature value across the dataset.

4.2.3 Cluster Composition

Cluster composition was assessed to determine how clinician-labeled data are distributed across clusters, revealing whether specific clusters are populated with certain labels and highlighting the relationships between clustering and clinical classifications.

Figures 4.11, 4.13, 4.15 show the labeled data in the embedding space for each classification (DERM, DERM-1M, CO), while Figures 4.12, 4.14, 4.16 illustrate the correspondent cluster composition.

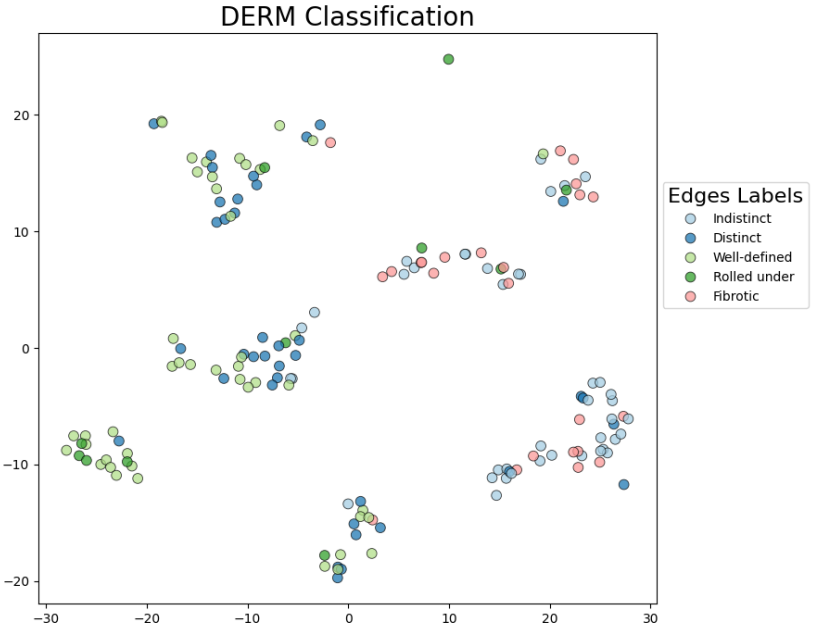

Figure 4.11: DERM labeled data in the embedding space distinguished by color.

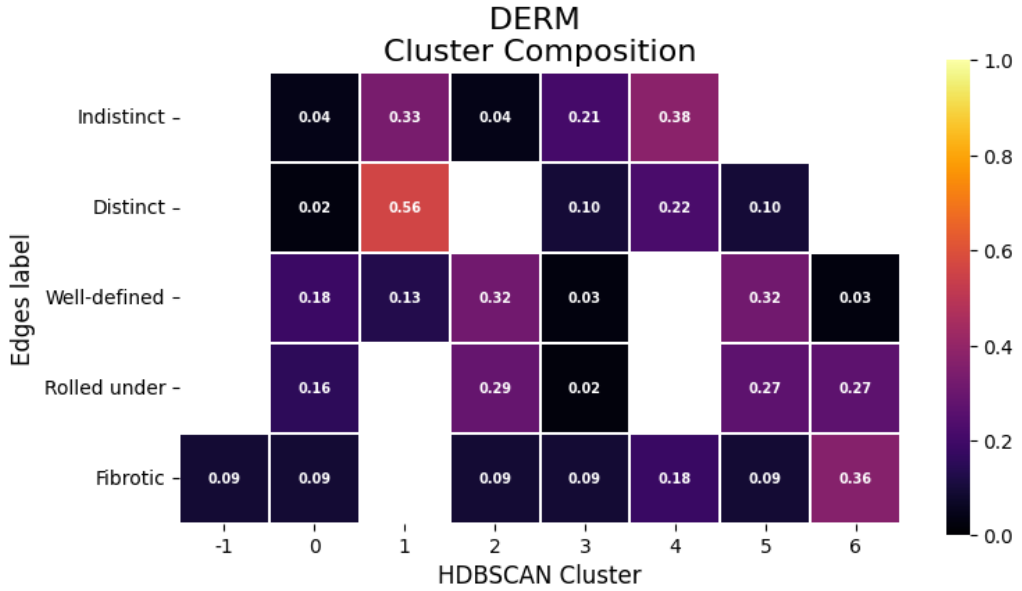

Figure 4.12: Cluster composition matrix that illustrate how DERM classified data are distributed across the clusters. The matrix is normalized across the rows.

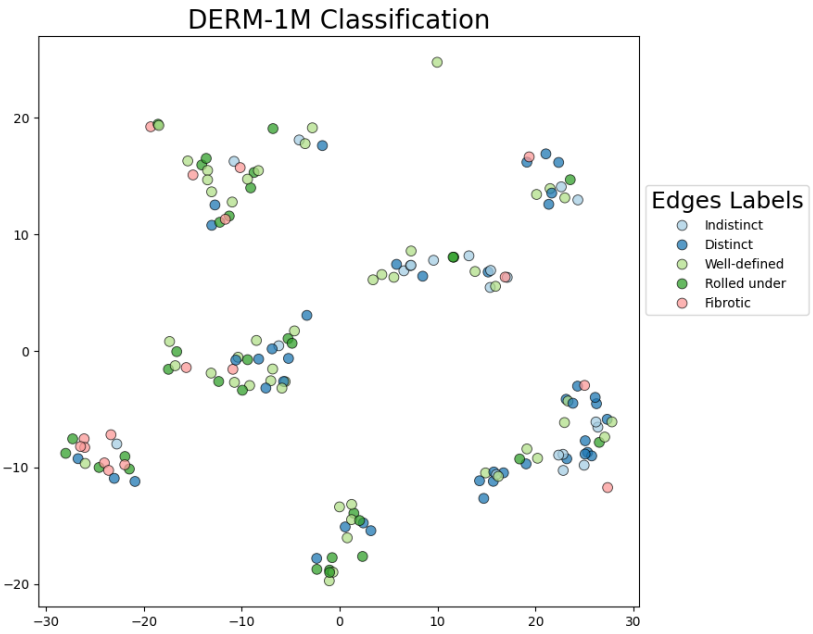

Figure 4.13: DERM-1M labeled data in the embedding space distinguished by color.

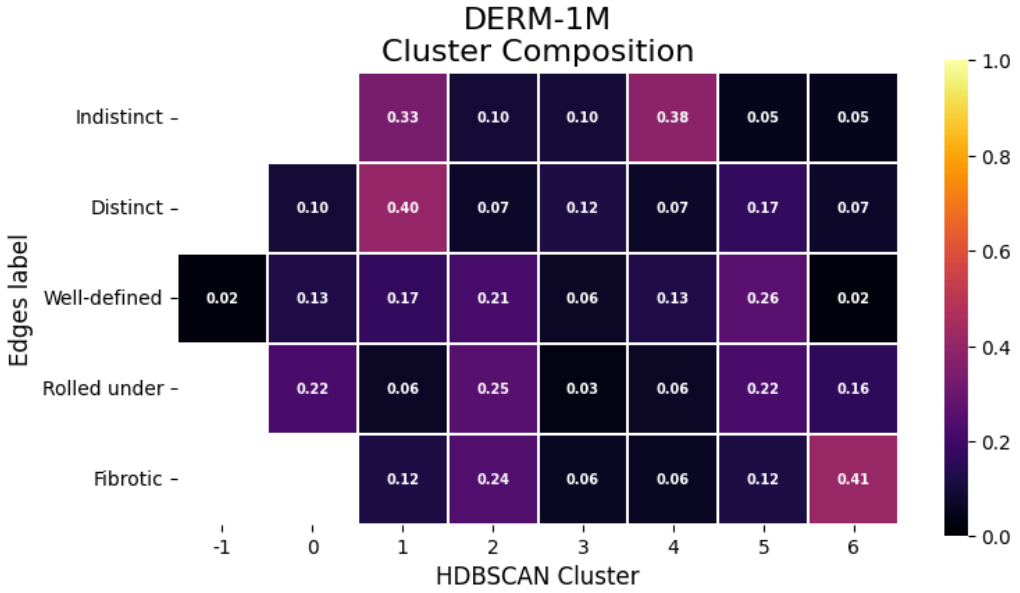

Figure 4.14: Cluster composition matrix that illustrate how DERM-1M classified data are distributed across the clusters. The matrix is normalized across the rows.

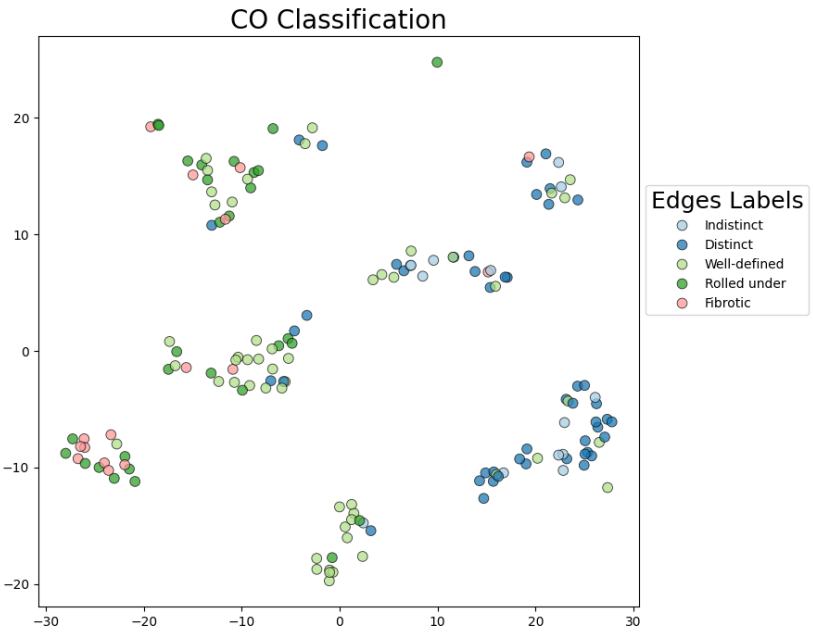

Figure 4.15: CO labeled data in the embedding space distinguished by color.

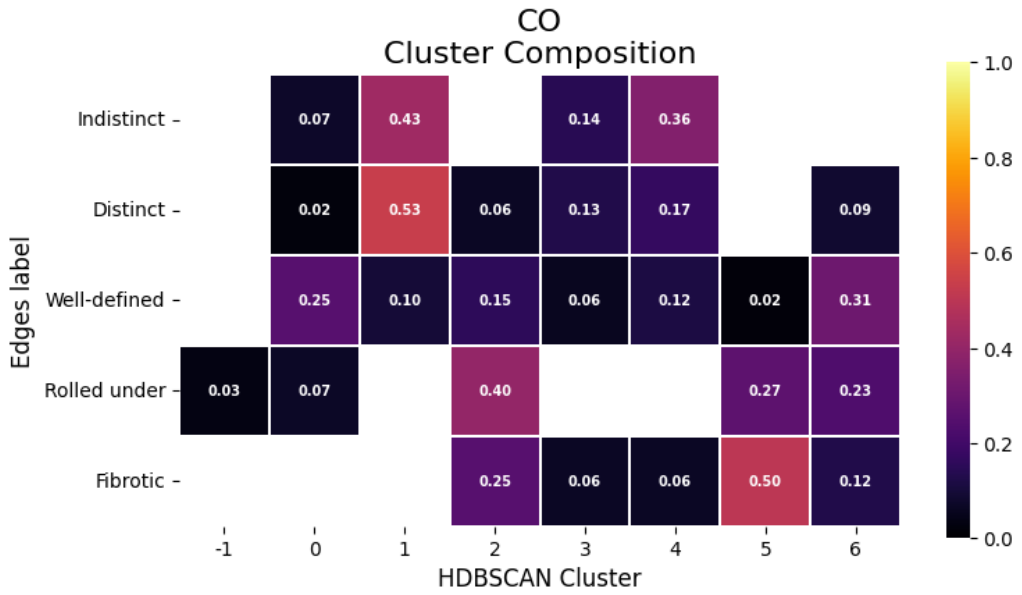

Figure 4.16: Cluster composition matrix that illustrate how CO classified data are distributed across the clusters. The matrix is normalized across the rows.

Analysis of the cluster composition across all three classifications reveals some general patterns. Specifically, Indistinct and Distinct classes are predominantly concentrated in clusters 1, 3, and 4. Meanwhile, the Rolled under and Fibrotic labeled-data predominantly populate clusters 2, 5, and 6. The Well-defined class, however, demonstrates a more scattered distribution across multiple clusters, with its spread varying depending on the specific classification. In general, these cases appear to be distributed throughout all identified clusters.

4.3 Linear Discriminant Analysis for Wound Edge Classification

Linear Discriminant Analysis (LDA) was applied as supervised algorithm based on the clinician scores to visualize the data in a two-dimensional space, focusing on the classifications of wound edges. The method used the extracted features from the depth profiles and was applied to all three available clinical classifications.

LDA scatterplots are presented in Figures 4.17 to 4.19, along with Kernel Density Estimation (KDE) plots to visualize the density of points across the transformed space.

From these plots, it can be observed that in the two-dimensional space, the edge classes exhibit a significant degree of overlap, particularly in DERM-1. In detail, the Indistinct and Distinct labeled-data are almost completely superimposed, while the Well-defined class is distributed across multiple other categories, demonstrating considerable intersection with Rolled under. However, the Fibrotic class appears relatively distinct from the other classifications, with limited overlap.

In order to better visualize the overlap between categories, the Euclidean distance between each point in the LDA space was calculated and organized into a normalized dissimilarity matrix for each case (Figures 4.20 to 4.22).

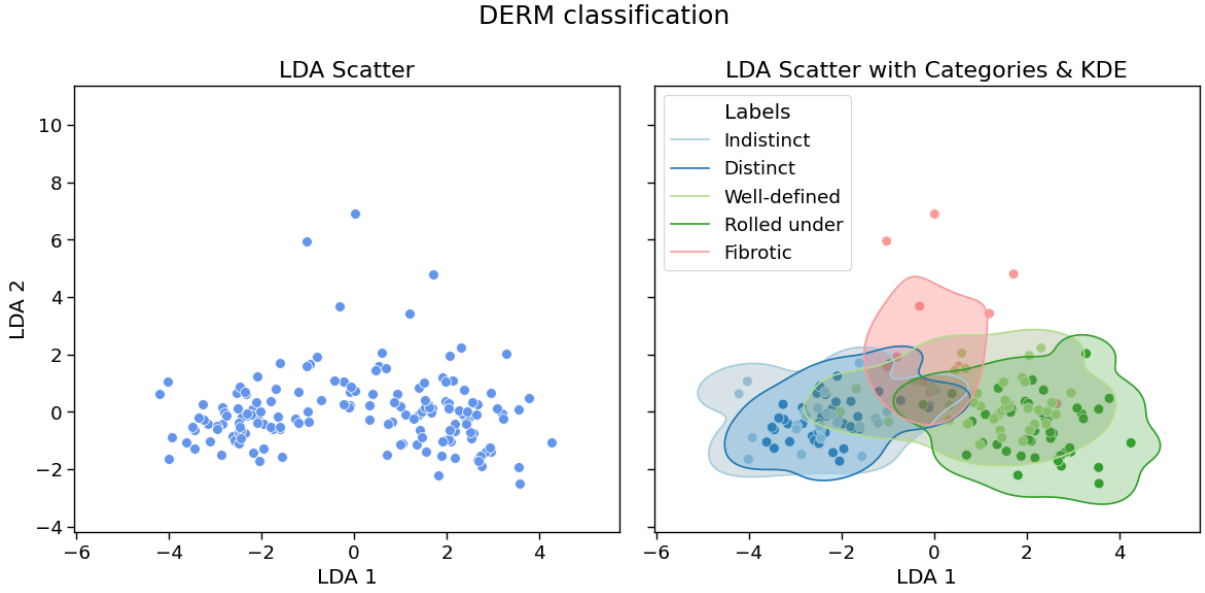

Figure 4.17: On the left, LDA scatterplot in the transformed space for DERM classification. On the right, plot of KDE overlaid to visualize the density of points across the space.

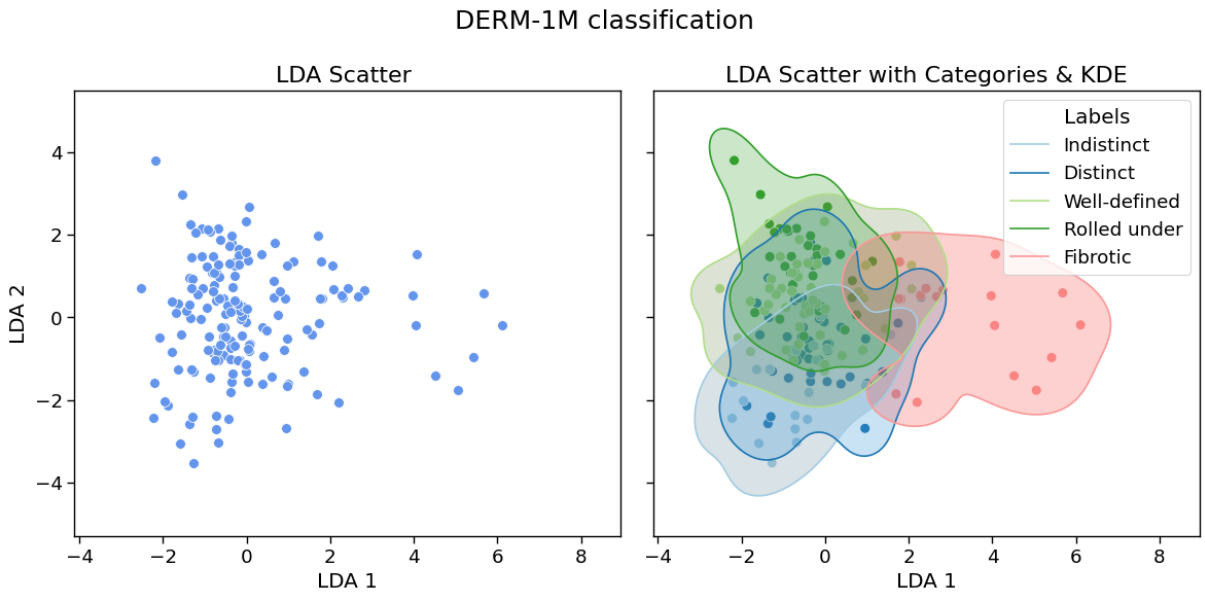

Figure 4.18: On the left, LDA scatterplot in the transformed space for DERM-1M classification. On the right, plot of KDE overlaid to visualize the density of points across the space.

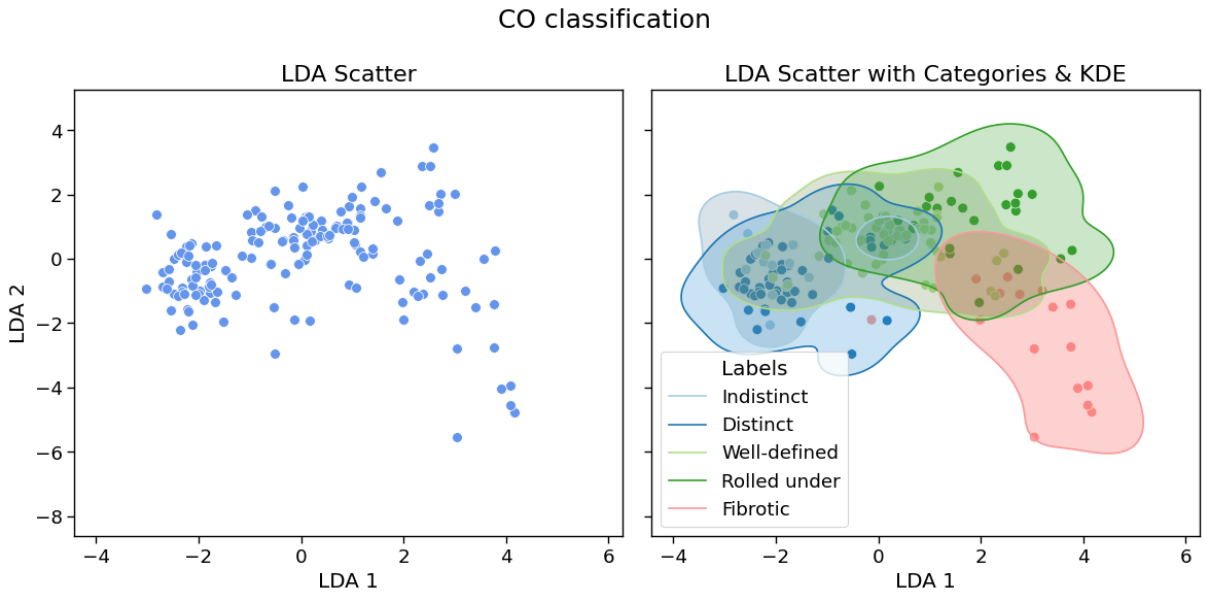

Figure 4.19: On the left, LDA scatterplot in the transformed space for CO classification. On the right, plot of KDE overlaid to visualize the density of points across the space.

The matrices largely confirmed the observations from the scatterplots: in the DERM and CO datasets, the Indistinct and Distinct categories, as well as the Well-defined and Rolled Under categories, are closer together compared to other class pairs. However, in the DERM-1M dataset, these categories appear almost completely overlapped between them. The Fibrotic class remains consistently well-separated from the other categories in all cases.

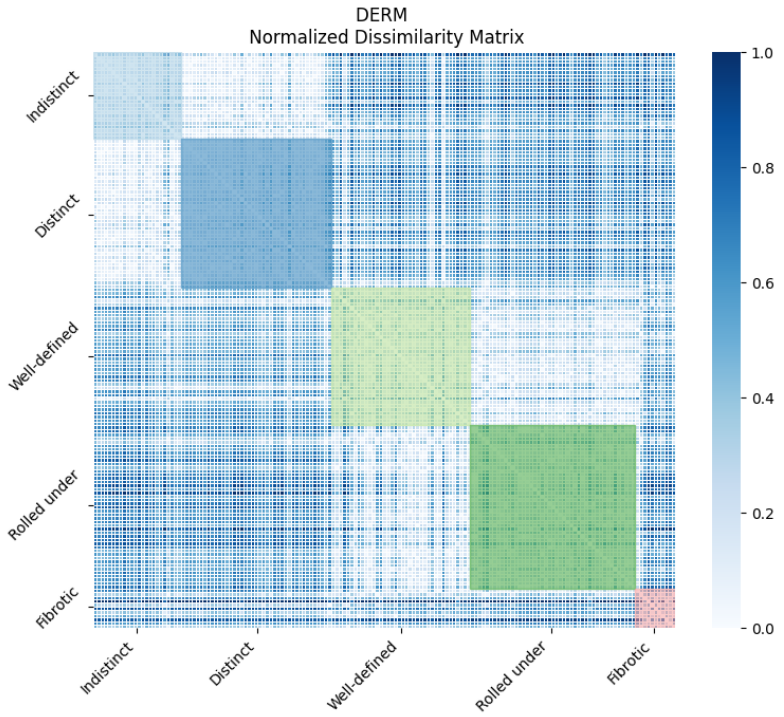

Figure 4.20: Normalized dissimilarity matrix for DERM classification. The colored squares highlight the distances between points with the same label.

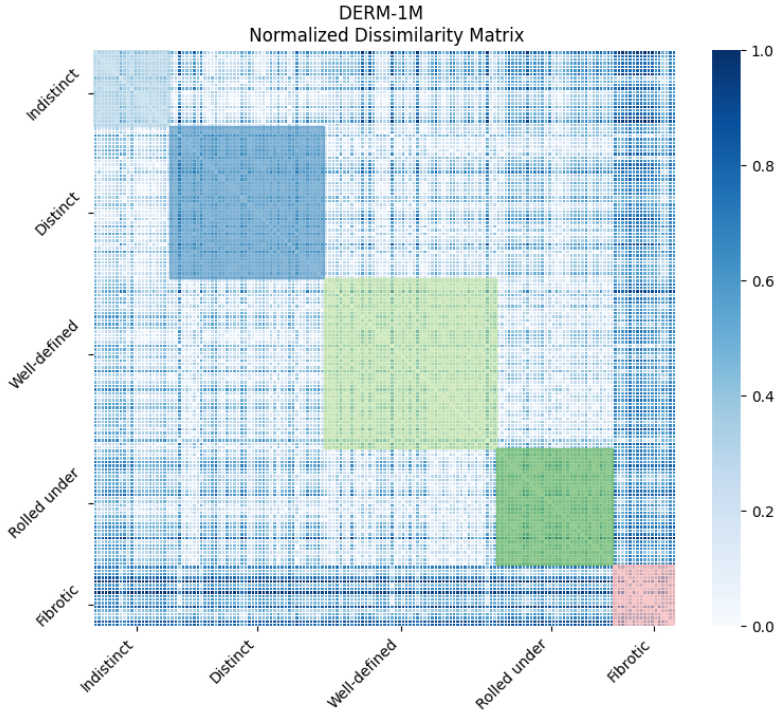

Figure 4.21: Normalized dissimilarity matrix for DERM-1M classification. The colored squares highlight the distances between points with the same label.

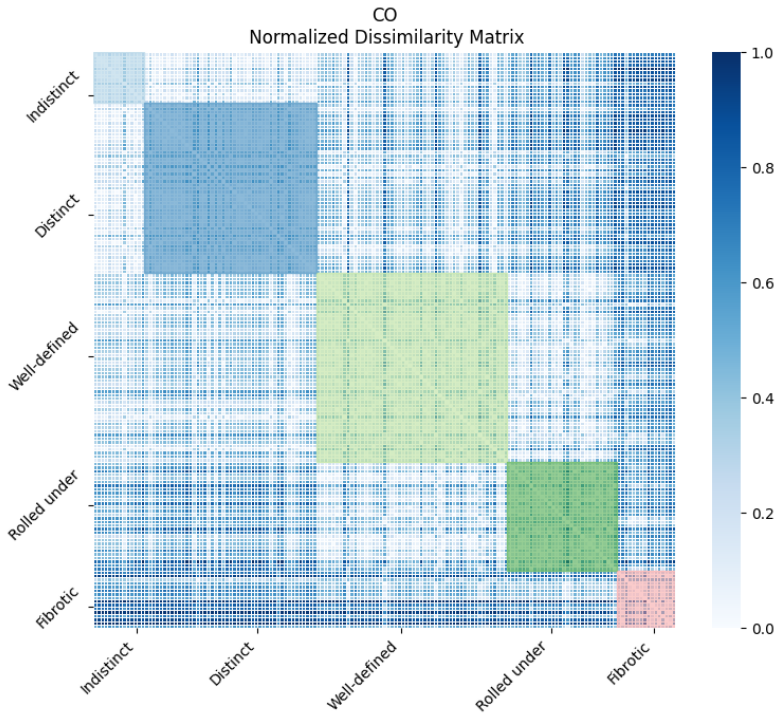

Figure 4.22: Normalized dissimilarity matrix for CO classification. The colored squares highlight the distances between points with the same label.

4.4 Random Forest for Wound Edge Classification

The depth profile features were also used to implement a Random Forest (RF) model for prediction of wound edges, using the three alternative clinical classifications as true values.

Model performance was evaluated through cross-validation:

```
RepeatedStratifiedKFold(
  n_splits=10,
  n_repeats=100,
  random_state=42
  )
```

i.e. splitting the data in 10 subsets for training and testing, and repeating the process 100 times for a more robust evaluation.

Balanced Accuracy was computed to account for class imbalance. The Matthews Correlation Coefficient (MCC) measured the correlation between predicted and true labels.

The model was initially developed to distinguish all five clinical labels. However, following consultations with clinicians, an alternative classification structures was defined merging some labels (Table 4.4), as discussed in Section 4.1.
The model was then retrained to reflect this revised classification.

In Table 4.5 are reported the accuracy (ACC) and the MCC for each trained model. The general pattern indicates that performance improves as the number of classes decreases.

Clinical Labels	Unmodified Labels	Grouped Labels
Indistinct	1	0
Distinct edges	2	0
Well-defined	3	1
Rolled-under	4	2
Fibrotic	5	2

Table 4.4: Label mapping across different class structures analyzed.

	DERM		DERM-1M		CO	
	ACC	MCC	ACC	MCC	ACC	MCC
Unmodified Labels	0.44 ± 0.09	0.43 ± 0.14	0.31 ± 0.12	0.12 ± 0.15	0.41 ± 0.10	0.38 ± 0.15
Grouped Labels	0.71 ± 0.11	0.62 ± 0.15	0.47 ± 0.11	0.24 ± 0.18	0.69 ± 0.11	0.55 ± 0.16

Table 4.5: Accuracy(ACC) and Matthews Correlation Coefficient (MCC) for different classification structures across three datasets DERM, DERM-1M, and CO.

However, merging the Indistinct-Distinct and Well-defined-Rolled under classes, where clinicians might generally disagree, yielded a good concordance with both DERM and CO, with a moderate correlation. This suggests that simplifying the classification scheme not only enhances the model's performance but also aligns better with clinical assessments, where there may be ambiguity or variability in classifying certain features

Furthermore, to evaluate the discriminative power between all possible pairs of classes, an ensemble of RF binary classifiers was implemented. The performance of each classifier is summarized in confusion matrices presenting both Accuracy (ACC) and Matthews Correlation Coefficient (MCC), as shown in Figures 4.23 to 4.25.

The model's performance in distinguishing wound edge revealed both strengths and weaknesses. Differentiating between Indistinct and Distinct edges proved particularly challenging, with accuracy ranging from 50% to 60% and low MCC values (0.01–0.25), indicating limited separability. In contrast, the model performed well when distinguishing Distinct or Indistinct edges from Rolled-under and Fibrotic categories, achieving higher accuracy and MCC scores (using DERM dataset, the distinction Distinct-Rolled under yielded ACC=0.99 ±0.03 and MCC=0.99±0.05). However, the classification of Fibrotic was problematic across all configurations, with poorer results distinguishing with Rolled-under and Well-defined (using CO dataset, the distinction Fibrotic-Rolled under yielded the lowest ACC=0.52 ±0.21) and MCC=0.03±0.44). The Well-defined category also exhibited inconsistent results across datasets, suggesting variability in clinical interpretation.

Performance was also highly dataset-dependent, with DERM-1M yielding the lowest metrics compared to DERM and CO, highlighting the need for improved data quality and refined class definitions.

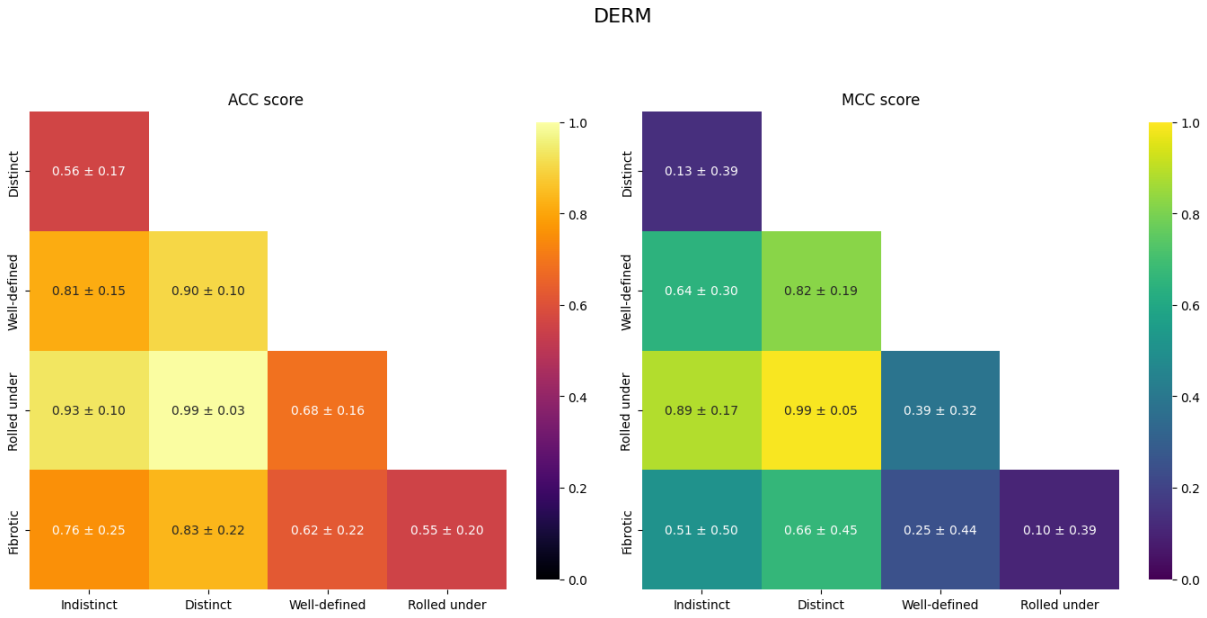

Figure 4.23: Confusion matrices for the ensemble of binary classifiers for DERM dataset, showing both Accuracy (ACC) and Matthews Correlation Coefficient (MCC) for each class pair.

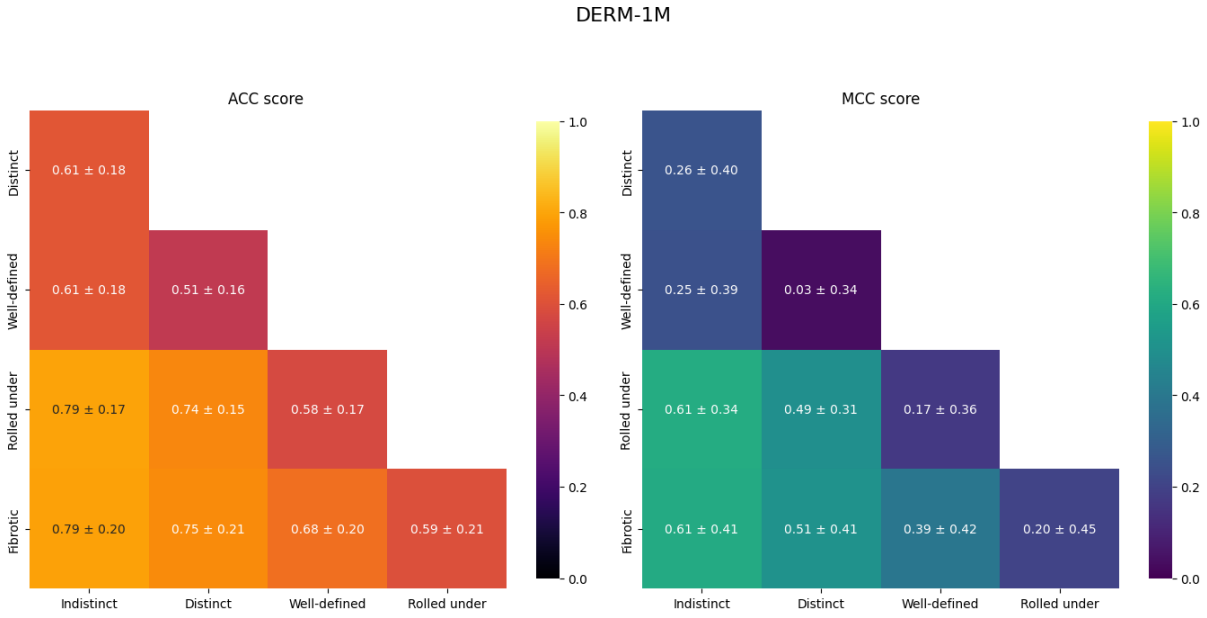

Figure 4.24: Confusion matrices for the ensemble of binary classifiers for DERM-1M dataset, showing both Accuracy (ACC) and Matthews Correlation Coefficient (MCC) for each class pair.

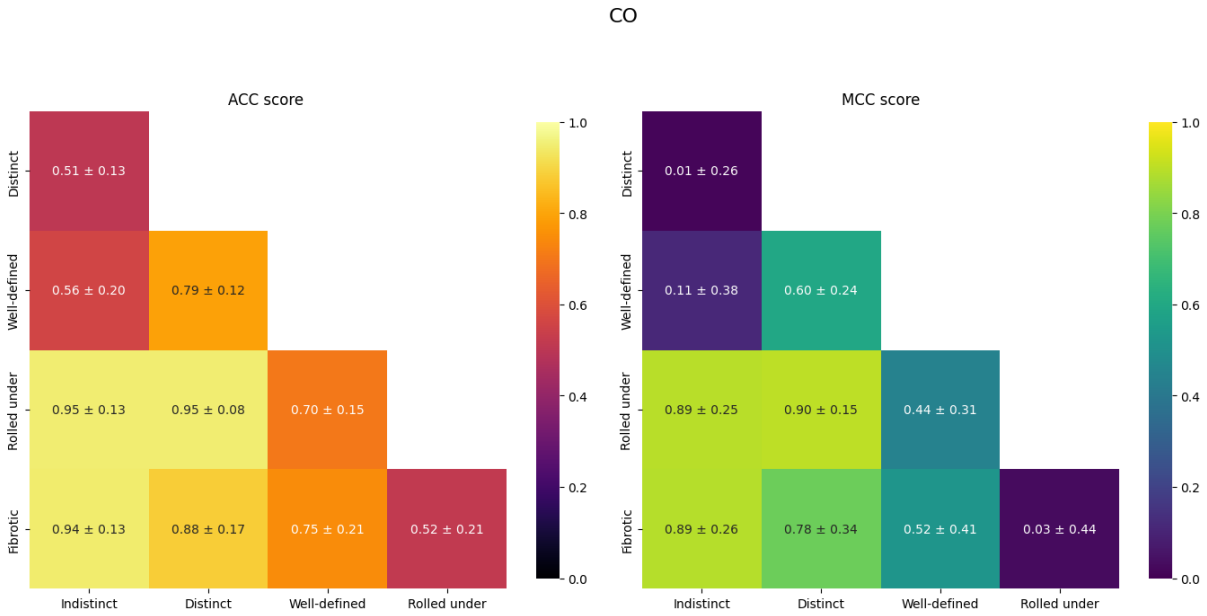

Figure 4.25: Confusion matrices for the ensemble of binary classifiers for CO dataset, showing both Accuracy (ACC) and Matthews Correlation Coefficient (MCC) for each class pair.

Building on the observed dataset dependency, a more robust approach was tested by training the model exclusively on images where all datasets agreed on the classification. Specifically, 53 images with consistent labels across the original labels and 79 images with consistent labels across the grouped labels were used to train the model. The corresponding confusion matrices for these cases are presented in Figures 4.26 and 4.27.

For the original labels, the model struggled to differentiate between Indistinct and Distinct edges, as well as between Rolled-under and Fibrotic. The latter challenge is likely exacerbated by the under-representation of the Fibrotic class in the dataset. The overall performance of the model reflects these limitations (ACC= 0.54±0.20, MCC=0.46±0.27). Some examples of misclassified image are illustrated in Figures to 4.28 to 4.31.

In contrast, using grouped labels improved the model's accuracy significantly (ACC= 0.72 ± 0.17, MCC=0.72 ± 0.18). The Well-defined category was not predicted with high confidence, highlighting its borderline nature and the challenges in its consistent clinical interpretation. However, This approach allowed for great prediction performance without completely sacrificing the finer distinctions of the original labels.

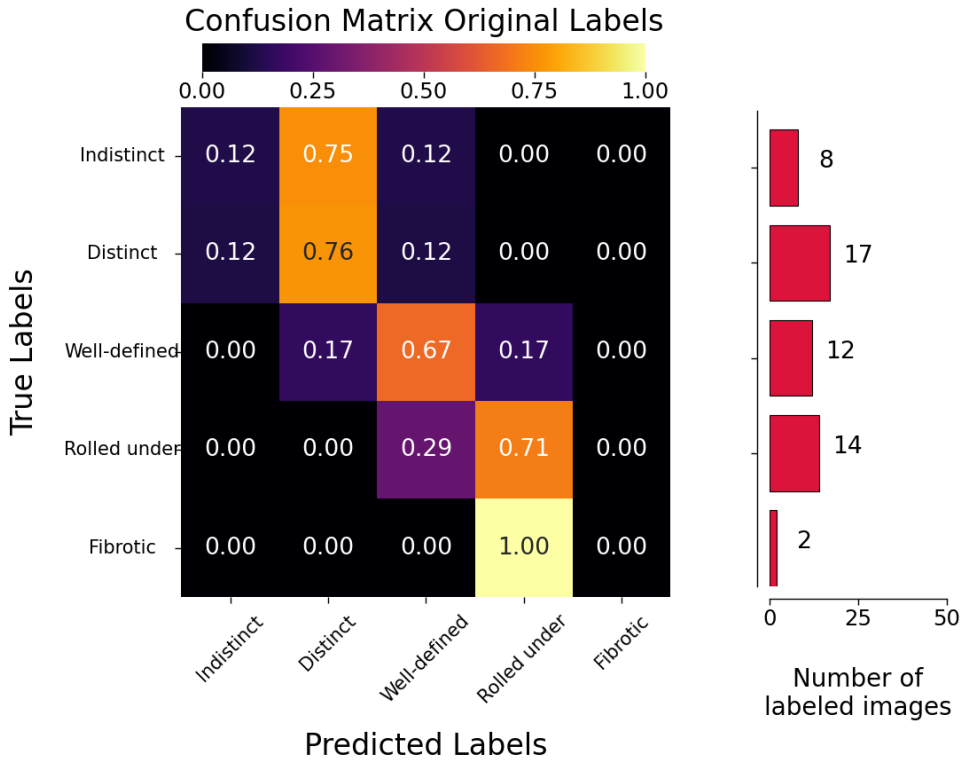

Figure 4.26: Confusion matrix of the model for predicting the original five labels, using only the 53 images where all clinicians agree.

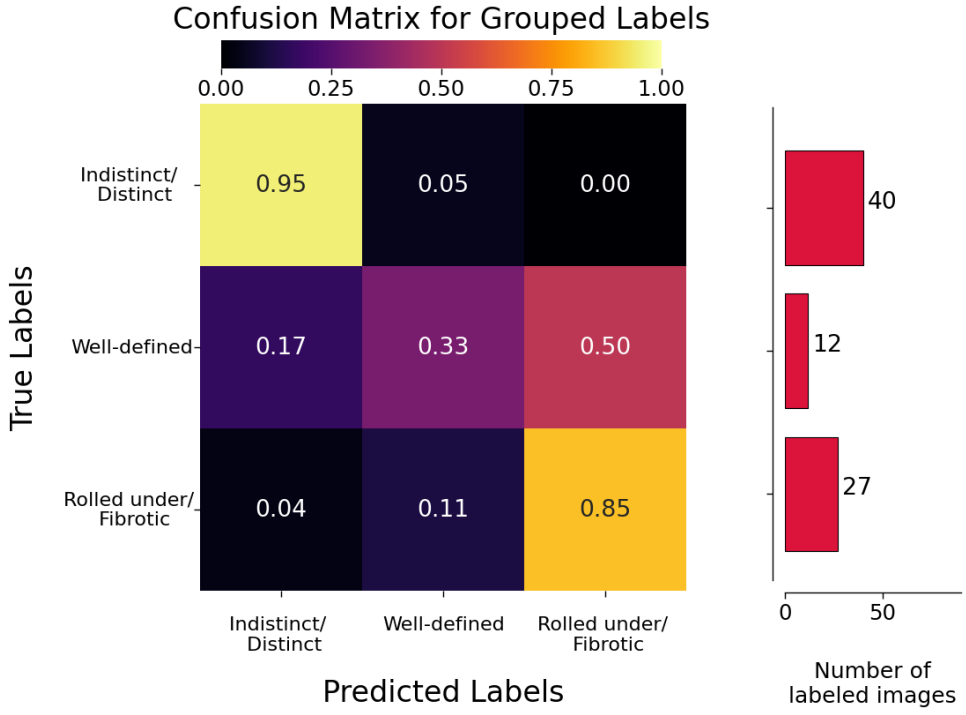

Figure 4.27: Confusion matrix of the model for predicting the grouped labels, using the 79 images where all clinicians agree.

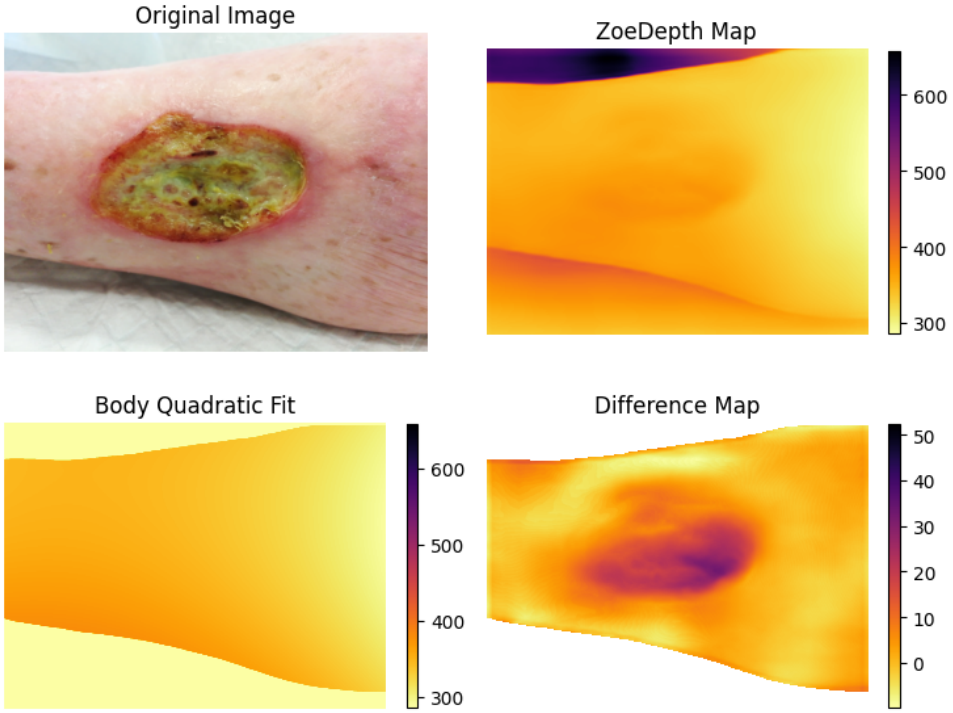

Figure 4.28: Wound with edges classified as Well-defined by clinicians but misclassified as Distinct by the model, likely because the wound bed is filled with slough tissue, which alters the depth change perceived by the model.

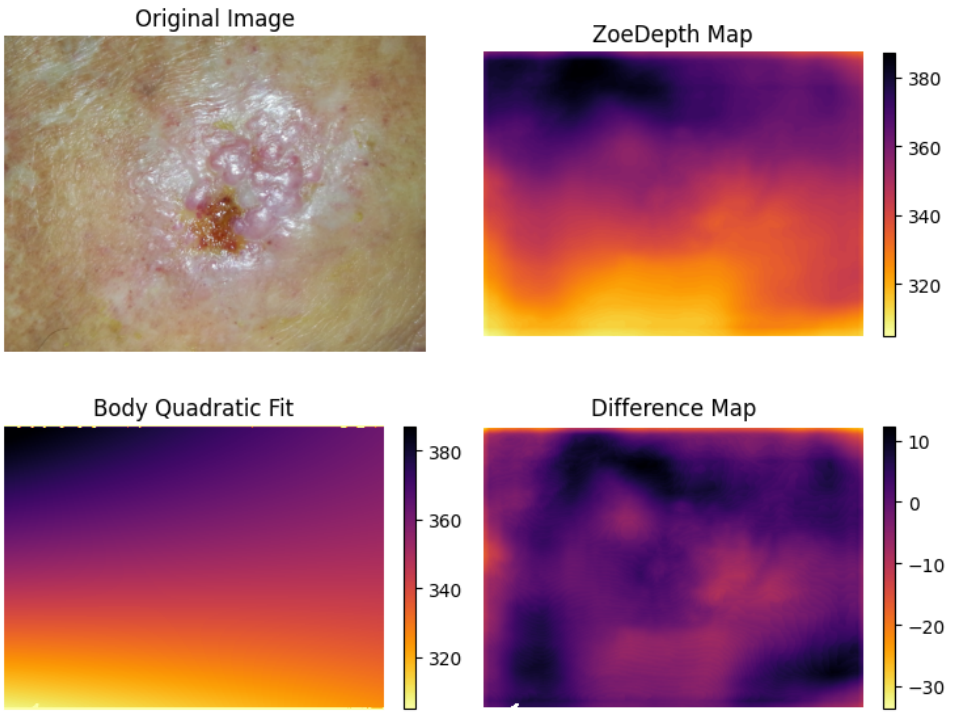

Figure 4.29: Wound image with edges classified as Indistinct by clinicians but misclassified as Distinct by the model, likely because of the scarring around the wound, which alters the depth change measured by the model.

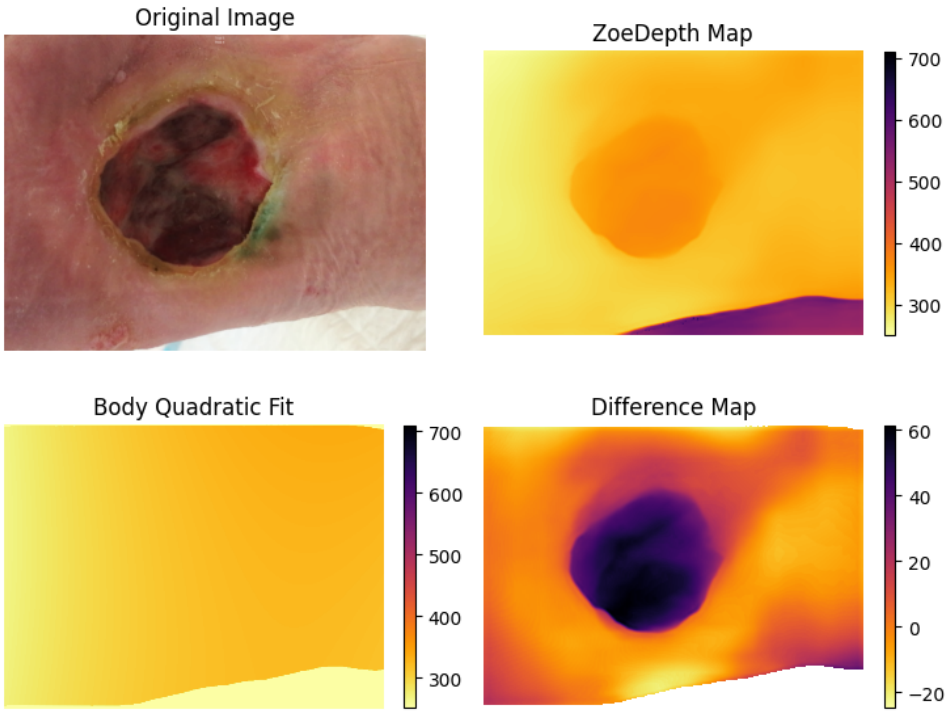

Figure 4.30: Wound image with edges classified as Fibrotic by clinicians but misclassified as Rolled under by the model.

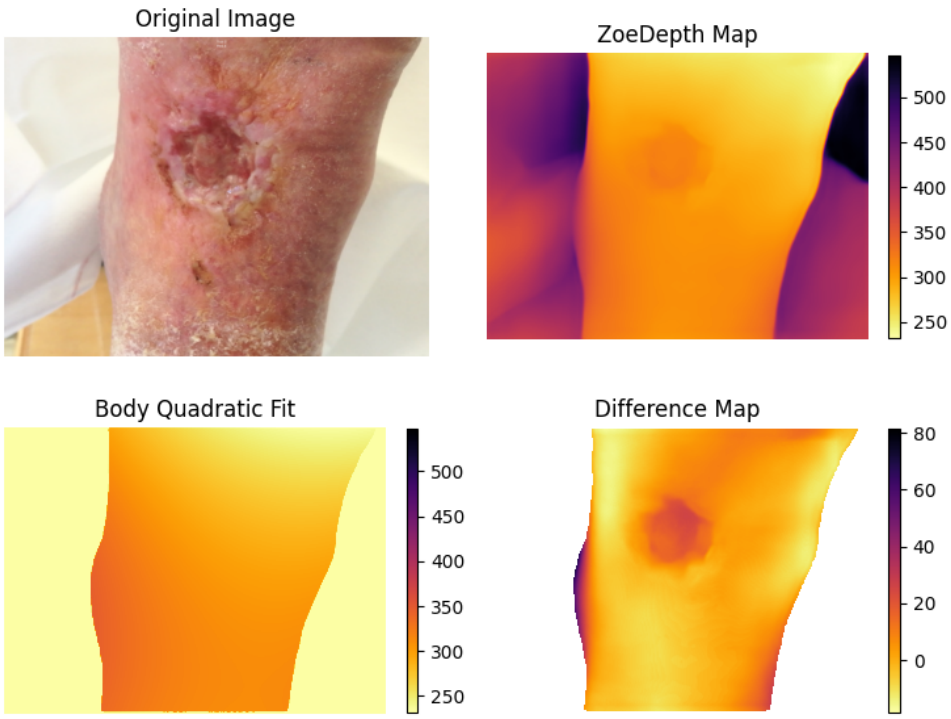

Figure 4.31: Wound image with edges classified as Well-defined by clinicians but misclassified as Rolled under by the model, likely due to jagged edges that hinder an accurate depth evaluation within the wound border.

Chapter 5

Discussions

This chapter will present the results of the study, focusing on the level of agreement among clinicians in wound edge assessment and on the application of machine learning techniques for this task. The findings will highlight the performance of various models, including clustering and classification approaches, and their alignment with clinical evaluations.

5.1 Interpretation of the results

The results of this study underscore the potential of machine learning techniques in advancing the assessment of wound edge for a better management of chronic wounds. The proposed approach aims to cluster and classify wound edges in a manner consistent with clinical evaluations, aligning with established scoring systems.

5.1.1 Clinician Agreement

The analysis of clinician agreement provided valuable insights into the consistency of wound edge classification and allows the validation of the discussed methods. The evaluation of concordance between different clinicians demonstrated a maximum agreement rate of $\approx 60\%$, with an Intraclass Correlation Coefficient (ICC) of 0.79. However, when the same clinician reassessed the same dataset after one month, the agreement rate was $\approx$39%, and the ICC was 0.42. Grouping the labels according to clinician recommendations resulted in an average clinical agreement around 62%. These values highlight both the potential for discrepancies between clinicians and the individual variability in clinical classification, as previously reported in the literature.

5.1.2 Clustering Results

By focusing on features extracted from profile of the depth rectified borders, the clustering process successfully revealed patterns that align with established clinical scoring systems like the Bates-Jensen Wound Assessment Tool (BWAT).

The application of the HDBSCAN clustering algorithm to the PaCMAP-embedded feature space identified eight clusters, including a noise cluster. The evaluation of cluster composition provides insights into the distribution of labeled-data within the identified clusters. Specifically, the categories Indistinct and Distinct were predominantly grouped together in certain clusters (1, 3, and 4), while the categories Fibrotic and Rolled-under exhibited a tendency to concentrate in different clusters (2, 5, and 6). The Well-defined

category was dispersed across almost all clusters (0,1,2,4,5 and 6), indicating variability in its classification.

It is important to note that the clusters identified were clearly distinct and represent more refined classification than the available labels. These clusters could potentially be used not only to provide a finer labeling of the wound edge scale but also to facilitate clearer agreement between clinicians and objective parameters, helping to reduce subjectivity in clinical assessments.

5.1.3 Results of Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) further highlight the degree of overlap between the Indistinct and Distinct labeled-data whitin the LDA embedded space, as well as the tendency of the Well-defined class to disperse. Conversely, the Fibrotic class was observed to exist as a relatively distinct category, exhibiting minimal overlap with other categories.

However, the analysis of normalized dissimilarity matrices revealed that the intra-class distances were not consistently coherent across all datasets.

5.1.4 Classification Results

The Random Forest model demonstrated superior performance by using the grouped edge labels, specifically merging the Indistinct and Distinct classes, as well as the Fibrotic and Rolled classes. This results in good accuracy and Matthews Correlation Coefficient values (average accuracy ACC=0.62 ± 0.11), without entirely losing the finer distinctions of the original clinical classification.

Furthermore, the ensemble of binary classifiers again demonstrated the difficulty of distinguishing between the Indistinct and Distinct classes. The separation of the Well-defined and Fibrotic labeled-data from others exhibited differing performance depending on the data used for the algorithm training. The observed variability indicates that the consistency of model classification may be influenced by subjective clinical assessment.

To address the challenges posed by variability in the datasets, the model was retrained exclusively on images for which all three sources agreed on the labels, thereby creating a more robust dataset. This approach was applied both when using the original five labels and when grouping them into broader categories. In the latter case, the model achieved significantly remarkable performance (ACC=0.72 ± 0.17, MCC = 0.72 ± 0.18) demonstrating a strong alignment with clinical assessments.

5.2 Limitations of the Research

Despite its promising results, this study has several limitations.

The image dataset comes from a single hospital (IRCCS Sant'Orsola-Malpighi Hospital) and may not fully represent patient demographics. Additionally, clinical assessments were conducted by only two operators on a limited number of images, which may introduce bias and reduce robustness of the clinical evaluations.

The method is also sensitive to the image acquisition protocol, as accurate depth map estimation depends on the quality of input images.

Furthermore, the method excludes images where wounds are captured at low pixel resolution, limiting its applicability in cases where the wound occupies a small portion of

the image. Moreover, in cases of multiple wounds, the assessment is performed only on the largest wound to align with clinical protocols.

5.3 Future Developments

In light of the current findings, future research should focus on several key areas to enhance the model's effectiveness.

Exploring additional features from wound edges could enhance classification accuracy of the model, particularly in ambiguous cases where the model may struggle to differentiate wound edges.

Expanding the dataset to include images from multiple centers and regions will improve its robustness and applicability to a broader patient population. Additionally, increasing the number of clinicians involved in the evaluation process will enhance the reliability of clinical assessments.

Establishing standardized acquisition protocols for consistent image capture is crucial for ensuring high-quality input data. Moreover, exploring methods for assessing multiple wounds simultaneously could provide a more comprehensive understanding of patient conditions, ultimately leading to improved treatment outcomes.

Finally, developing user-friendly interfaces for both clinicians and patients will facilitate the model's adoption in routine wound care practices.

Conclusions

This thesis aimed to address the challenge of subjectivity in traditional wound assessment by proposing an innovative machine learning-based approach for wound edge classification. The primary goal was to improve diagnostic accuracy and reproducibility, ultimately assist clinicians to provide more effective care.

To achieve this, I analyzed a dataset of wound images acquired during clinical practice at the IRCCS Sant'Orsola-Malpighi University Hospital in Bologna. This dataset was labeled for wound edges according to the BWAT (Bate-Jensen Wound Assessment Tool) scoring scale by a dermatologist and a clinical operator with more than three years of experience in wound assessment.

The level of agreement among clinicians was estimated, revealing an average concordance of approximately 52% when they use the five labels of the BWAT for wound edges (Indistinct, Distinct, Well-defined, Rolled under, Fibrotic). Notably, when Distinct and Indistinct classes were grouped together, as well as Rolled under and Fibrotic classes, categories in which clinicians acknowledged variability in assessment, the agreement increased to about 62%.

During the analysis process the wound images were segmented using Deepskin algorithm to isolate the wound area and, subsequently, depth maps of images were computed with ZoeDepth model to capture three-dimensional structural details of wound edges.

The unique feature of the proposed methodology was the introduction of the "wound border rectification" technique, which standardizes the periwound area into a rectangular format, regardless of its original shape. This standardization enabled consistent extraction of features from border depth profiles, facilitating more accurate analysis across varying wound geometries.

The extracted features were used for both clustering and classification of wound edges.

The clustering analysis successfully identified distinct patterns in wound edge characteristics, involving dimensionality reduction through PacMAP algorithm to enhance the interpretability of the data. In particular, the analysis revealed that the Distinct and Indistinct classes, as well as the Rolled under and Fibrotic, were grouped within the same clusters, showing a meaningful alignment with clinical assessments. Notably, while the BWAT provides five labels for wound edge classification, the analysis revealed eight clearly distinct clusters. This suggests that the clustering approach could provide a finer labeling of the wound edge classification and also facilitate clearer agreement between clinicians and objective parameters.

The Random Forest model trained on depth features demonstrated limitations in distinguishing between similar edge labels, specifically between Indistinct and Distinct, as well as Rolled under and Fibrotic. The Well-defined category exhibited borderline behavior. The algorithm proved to be highly dependent on the clinician's classification used for training; therefore,the focus was shifted to images where clinicians agreed in classification. By consolidating the ambiguous categories, the model achieved excellent

accuracy (ACC= 0.72 ± 0.17, MCC=0.72 ± 0.18), demonstrating its ability to provide reliable classifications without completely sacrificing the finer distinctions BWAT labels.

Overall, the findings indicate that the integration of the proposed machine learning approach with depth map analysis and border rectification, can significantly enhance the objectivity of wound edges assessment. This approach holds promise for future clinical applications, paving the way for more standardized and efficient wound care practices.

Bibliography

[1] C. Lindholm and R. Searle. "Wound management for the 21st century: combining effectiveness and efficiency". In: *Journal of Wound Care* (2016).

[2] Stefan Stremitzer, Thomas Wild, and Thomas Hoelzenbein. "How precise is the evaluation of chronic wounds by health care professionals?" In: *Journal of Wound Care* (2024).

[3] Ronald Marks. *Skin Breakdown: The Silent Epidemic*. The Smith and Nephew Foundation, 2007.

[4] C. Lindholm and R. Searle. "The Financial and Quality-of-Life Cost to Patients Living with a Chronic Wound in the Community". In: *Journal of Wound Care* 30.5 (2021), pp. 234–240. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7949507/.

[5] S. Kumar and R. Gupta. "Opportunities and Challenges of the Management of Chronic Wounds: A Multidisciplinary Approach". In: *Chronic Wound Care Management and Research* 10 (2023), pp. 1–12. DOI: 10.2147/CWCMR.S123456. URL: https://www.dovepress.com/opportunities-and-challenges-of-the-management-of-chronic-wounds-a-mul-peer-reviewed-fulltext-article-CWCMR.

[6] Robert G. Frykberg. "Challenges in the Treatment of Chronic Wounds". In: *International Journal of Lower Extremity Wounds* 14.3 (2015), pp. 181–192. DOI: 10.1177/1534734615580274. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4528992/.

[7] Laurie McNichol et al. "Assessment of the patient with a wound". In: *Wound, Ostomy, and Continence Nurses Society Core Curriculum: Wound Management*. 2nd. Lippincott Williams and Wilkins, 2021, pp. 53–75. ISBN: 9781975164591.

[8] DC Garbuio et al. "Assessment tools for the healing of wounds: an integrative review". In: *Rev. Eletr. Enf.* 20 (2018). [cited yyyy-mm-dd], v20a40. DOI: 10.5216/ree.v20.49425. URL: https://doi.org/10.5216/ree.v20.49425.

[9] Lisa Gould et al. "Wound healing society 2015 update on guidelines for pressure ulcers". In: *Wound Repair and Regeneration* 24 (2016), pp. 145–162.

[10] D. J. Margolis, J. Kantor, and J. A. Berlin. "Healing of diabetic foot ulcers: Analysis of data from a randomised controlled trial of systemic growth hormone". In: *The Lancet* 355.9201 (2000), pp. 855–856. DOI: 10.1016/S0140-6736(99)08045-0.

[11] J. C. Lantis et al. "Healing rates and time to heal in patients with diabetic foot ulcers". In: *Journal of Wound Care* 22.8 (2013), pp. 407–414. DOI: 10.12968/jowc.2013.22.8.407.

[12] Line Bisgaard Jørgensen et al. “Methods to assess area and volume of wounds – a systematic review”. In: *Journal of Wound Care* 25.10 (2016), pp. 577–582. DOI: 10.12968/jowc.2016.25.10.577.

[13] Jonathan Kantor and David J. Margolis. “Efficacy and Prognostic Value of Simple Wound Measurements”. In: *Archives of Dermatology* 134.12 (1998), pp. 1571–1574. DOI: 10.1001/archderm.134.12.1571.

[14] J. Kantor and D.J. Margolis. “A multicentre study of percentage change in venous leg ulcer area as a prognostic index of healing at 24 weeks”. In: *British Journal of Dermatology* 142.5 (May 2000), pp. 960–964. DOI: 10.1046/j.1365-2133.2000.03478.x.

[15] Laurie Swezey. “How to Assess Wounds for Tunneling and Undermining”. In: *WoundSource* (2014). URL: https://www.woundsource.com/blog/how-assess-wounds-tunneling-and-undermining.

[16] D. Laplaud, J. E. Lechapt, and E. Dap. “Wound debridement: comparative reliability of three methods for measuring fibrin percentage in chronic wounds”. In: *Wound Repair and Regeneration* 18.2 (2010), pp. 201–206. DOI: 10.1111/j.1524-475X.2010.00558.x.

[17] Barbara M. Bates-Jensen et al. “Reliability of the Bates-Jensen Wound Assessment Tool (BWAT) for Pressure Injury Assessment: The Pressure Ulcer Detection Study”. In: *Wound Repair and Regeneration* 27.4 (July 2019). Author manuscript; available in PMC 2020 July 01. Published in final edited form as: Wound Repair Regen. 2019 July; 27(4): 386–395., pp. 386–395. DOI: 10.1111/wrr.12714.

[18] Connie Harris et al. “Wound Care: Bates-Jensen Wound Assessment Tool Pictorial Guide Validation Project”. In: *Journal of Wound, Ostomy and Continence Nursing* 37.3 (May 2010), pp. 253–259. DOI: 10.1097/WON.0b013e3181d73aab.

[19] Karen E Campbell, David H Keast, and M Gail Woodbury. “Photographic assessment of the appearance of chronic pressure and leg ulcers”. In: *Ostomy/Wound Management* 46.5 (2000). Published in May 2000, available on PubMed, pp. 26–34.

[20] Nicole Thompson et al. “Reliability and Validity of the Revised Photographic Wound Assessment Tool on Digital Images Taken of Various Types of Chronic Wounds”. In: *Advances in Skin and Wound Care* 26.8 (2013). Published in Vol. 26, No. 8, pages 360-373, pp. 360–373.

[21] Alan M. Turing. “Computing Machinery and Intelligence”. In: *Mind* 59.236 (1950), pp. 433–460. DOI: 10.1093/mind/LIX.236.433. URL: https://academic.oup.com/mind/article/LIX/236/433/986238.

[22] Arthur L. Samuel. “Some Studies in Machine Learning Using the Game of Checkers”. In: *IBM Journal of Research and Development* 3.3 (1959), pp. 210–229. DOI: 10.1147/rd.33.0210. URL: https://ieeexplore.ieee.org/document/5392560.

[23] Frank Rosenblatt. “The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain”. In: *Psychological Review* 65.6 (1958), pp. 386–408. DOI: 10.1037/h0042519. URL: https://psycnet.apa.org/doi/10.1037/h0042519.

[24] Geoffrey E. Hinton, David E. Rumelhart, and Ronald J. Williams. “Learning representations by back-propagating errors”. In: *Nature* 323 (1986), pp. 533–536. DOI: 10.1038/323533a0. URL: https://www.nature.com/articles/323533a0.

[25] Geoffrey E. Hinton, Simon Osindero, and Yee-Whye Teh. "A Fast Learning Algorithm for Deep Belief Nets". In: *Neural Computation* 18 (2006), pp. 1527–1554. DOI: 10.1162/neco.2006.18.7.1527. URL: https://direct.mit.edu/neco/article/18/7/1527/6982.

[26] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. "Machine Learning Basics". In: *Deep Learning*. Ed. by Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Cambridge, MA: MIT Press, 2016. Chap. 5.

[27] R. A. Fisher. "The Use of Multiple Measurements in Taxonomic Problems". In: *Annals of Eugenics* 7.2 (1936), pp. 179–188. DOI: 10.1111/j.1469-1809.1936.tb02137.x.

[28] Leo Breiman. "Random Forests". In: *Machine Learning* 45 (2001), pp. 5–32. DOI: 10.1023/A:1010933404324. URL: https://link.springer.com/article/10.1023/A:1010933404324.

[29] Leo Breiman. "Bagging Predictors". In: *Machine Learning* 24 (1996), pp. 123–140. DOI: 10.1007/BF00058655.

[30] Bradley Efron. "Bootstrap Methods: Another Look at the Jackknife". In: *The Annals of Statistics* 7.1 (1979), pp. 1–26. URL: http://www.jstor.org/stable/2958830.

[31] Leo Breiman et al. *Classification and Regression Trees*. Belmont, CA: Wadsworth International Group, 1986.

[32] Dongkuan Xu and Yingjie Tian. "A Comprehensive Survey of Clustering Algorithms". In: *Annals of Data Science* 2.2 (2015), pp. 165–193. DOI: 10.1007/s40745-015-0040-1.

[33] Yingfan Wang et al. "Understanding How Dimension Reduction Tools Work: An Empirical Approach to Deciphering t-SNE, UMAP, TriMap, and PaCMAP for Data Visualization". In: *Journal of Machine Learning Research* 22.201 (2021), pp. 1–73. URL: http://jmlr.org/papers/v22/20-1061.html.

[34] A. Bakiya, K. Kamalanand, V. Rajinikanth, et al. "Deep neural network assisted diagnosis of time-frequency transformed electromyograms". In: *Multimedia Tools and Applications* 79.15 (2020). Received 25 May 2018, Revised 08 July 2018, Accepted 15 August 2018, Published 03 September 2018, pp. 11051–11067. DOI: 10.1007/s11042-018-6561-9. URL: https://doi.org/10.1007/s11042-018-6561-9.

[35] L. Alzubaidi, J. Zhang, A. J. Humaidi, et al. "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions". In: *Journal of Big Data* 8.1 (2021), p. 53. DOI: 10.1186/s40537-021-00444-8. URL: https://doi.org/10.1186/s40537-021-00444-8.

[36] Katarzyna Szramowiat-Sala. "Artificial Intelligence in Environmental Monitoring: Application of Artificial Neural Networks and Machine Learning for Pollution Prevention and Toxicity Measurements". In: (July 2023). DOI: 10.20944/preprints202307.1298.v1.

[37] Xavier Glorot and Yoshua Bengio. "Understanding the Difficulty of Training Deep Feedforward Neural Networks". In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. PMLR, 2010, pp. 249–256.

[38] Kaiming He et al. "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2015, pp. 1026–1034.

[39] Guoping Xu et al. *Development of Skip Connection in Deep Neural Networks for Computer Vision and Medical Image Analysis: A Survey*. 2024. arXiv: 2405.01725 [eess.IV]. URL: https://arxiv.org/abs/2405.01725.

[40] Sergey Ioffe and Christian Szegedy. *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift*. 2015. arXiv: 1502.03167 [cs.LG]. URL: https://arxiv.org/abs/1502.03167.

[41] Nitish Srivastava et al. "Dropout: A Simple Way to Prevent Neural Networks from Overfitting". In: *Journal of Machine Learning Research* 15 (2014). Submitted 11/13; Published 6/14, pp. 1929–1958. URL: http://www.jmlr.org/papers/volume15/srivastava14a/srivastava14a.pdf.

[42] Michael A. Nielsen. *Neural Networks and Deep Learning*. 2015. URL: http://neuralnetworksanddeeplearning.com/chap6.html.

[43] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer. 2015, pp. 234–241. DOI: 10.1007/978-3-319-24574-4_28.

[44] T. D. Pham and D. T. P. Le. "Unveiling the role of artificial intelligence for wound assessment and wound healing prediction". In: *Exploratory Medicine* 4 (2023), pp. 589–611. DOI: 10.37349/emed.2023.0016. URL: https://doi.org/10.37349/emed.2023.0016.

[45] R. S. Howell et al. "Development of a method for clinical evaluation of artificial intelligence-based digital wound assessment tools". In: *JAMA Network Open* 4 (2021), e217234.

[46] M. Barakat-Johnson et al. "Reshaping wound care: evaluation of an artificial intelligence app to improve wound assessment and management amid the COVID-19 pandemic". In: *International Wound Journal* 19 (2022), pp. 1561–1577.

[47] K. Cross and K. Harding. "Risk profiling in the prevention and treatment of chronic wounds using artificial intelligence". In: *International Wound Journal* 19 (2022), pp. 1283–1285.

[48] M. Lustig et al. "A Machine Learning Algorithm for Early Detection of Heel Deep Tissue Injuries Based on a Daily History of Sub-Epidermal Moisture Measurements". In: *International Wound Journal* 19.6 (Oct. 2022). Epub 2022 Jan 12, pp. 1339–1348. DOI: 10.1111/iwj.13728.

[49] Daniel Y.T. Chino et al. "Segmenting skin ulcers and measuring the wound area using deep convolutional networks". In: *Computer Methods and Programs in Biomedicine* 191 (2020), p. 105376. ISSN: 0169-2607. DOI: https://doi.org/10.1016/j.cmpb.2020.105376. URL: https://www.sciencedirect.com/science/article/pii/S016926071931404X.

[50] Nico Curti et al. "Effectiveness of Semi-Supervised Active Learning in Automated Wound Image Segmentation". In: *International Journal of Molecular Sciences* 24.1 (2023). ISSN: 1422-0067. URL: https://www.mdpi.com/1422-0067/24/1/706.

[51] Nico Curti et al. "Automated Prediction of Photographic Wound Assessment Tool in Chronic Wound Images". In: *Journal of Medical Systems* 48.1 (Jan. 2024), p. 14. ISSN: 1573-689X. DOI: 10.1007/s10916-023-02029-9. URL: https://doi.org/10.1007/s10916-023-02029-9.

[52] A. Kolli, Q. Wei, and S. A. Ramsey. “Predicting Time-to-Healing from a Digital Wound Image: A Hybrid Neural Network and Decision Tree Approach Improves Performance”. In: *Computation* 12.3 (2024), p. 42. DOI: 10.3390/computation12030042. URL: https://doi.org/10.3390/computation12030042.

[53] Nico Curti. *deepskin pipeline: Wound analysis using smartphone images*. https://github.com/Nico-Curti/Deepskin. 2023. URL: https://github.com/Nico-Curti/Deepskin.

[54] Shariq Farooq Bhat et al. *ZoeDepth: Zero-shot Transfer by Combining Relative and Metric Depth*. 2023. DOI: 10.48550/ARXIV.2302.12288. URL: https://arxiv.org/abs/2302.12288.