

Ca' Foscari
University
of Venice

Master Degree
in Data Analytics for Business and Society

Final Thesis

Classification of Healthy and Diabetic Patients Using Plantar Foot Infrared Images

Supervisor
Prof. Gaia Bertarelli

Assistant supervisor
Dr. Corin Otesteanu

Graduand
Lorenzo Muscillo
Matriculation Number 1000916

Academic Year
2024 / 2025

Abstract

This study investigates the use of multimodal imaging and machine learning for classifying diabetic foot conditions, with a particular focus on the classification of diabetic foot disease (DFDs) and risk assessment. A total of 415 co-registered Red, Green, Blue (RGB) and thermal (IR) image pairs from the STANDUP dataset were analysed, comprising healthy controls and diabetic patients across three clinical stages (R0, R1 and R2). Deep learning models (ResNet and LSNet) were trained using various input strategies (RGB, IR and RGB+IR fusion) and were evaluated under binary (healthy versus diabetic) and multi-class classification settings. Classification was also performed using binary foot shape masks to assess the diagnostic value of morphological patterns. To capture vascular reactivity, images acquired before and after cold stress tests (T0–T10) were used as input for temporal thermal modelling. IR-based models consistently outperformed RGB-only models, achieving balanced accuracy of up to 99.5% in binary classification. Hybrid approaches that integrated statistical features with CNN predictions significantly improved performance, particularly in the multi-class setting. Interestingly, models trained solely on binary foot shape masks achieved comparable accuracy, indicating that morphological features could have diagnostic value. Furthermore, using T0-T10 cold stress image pairs improved classification, emphasising the importance of temporal thermal dynamics. However, due to the absence of longitudinal patient data, the predictive value of thermal asymmetries, particularly the 2.2°C threshold and z-score analysis, could not be conclusively validated. Their clinical significance should be explored in future prospective studies. This work demonstrates the effectiveness of multimodal deep learning and statistical analysis in classifying diabetic feet. It supports the use of IR imaging and hybrid modelling for the early assessment of risk of DFD, while identifying promising areas for further research, including shape-based analysis, temporal modelling and longitudinal validation.

Contents

Contents 3

List of Figures 4

List of Tables 5

1 Introduction . . . 7
1.1 Machine Learning on Diabetic Foot . . . 8
1.2 Convolutional Neural Networks . . . 10
1.3 Aim of the Thesis . . . 12
2 Materials and Methods . . . 13
2.1 Dataset: STANDUP Project . . . 13
2.2 Segmentation . . . 15
2.3 CNN Classifiers . . . 18
2.4 Dataset Splitting and Cross-Validation . . . 24
2.5 Statistical Analysis . . . 26
3 Results . . . 31
3.1 Deep Learning Models Performance . . . 31
3.2 Results of the Statistical Analysis . . . 35
3.3 Hybrid Approach: Deep Learning and Statistical Features . . . 38
4 Discussion and Conclusion . . . 42

Bibliography 47

List of Figures

1 CNN structure . . . 11
2 Image of FLIR ONE Pro. . . . 14
3 An acquisition example: (a) the RGB image and (b) the corresponding IR image . . . 14
4 Example of a plantar IR image acquired with the FLIR ONE Pro thermal imaging camera. . . . 14
5 Directory structure of the STANDUP dataset. . . . 16
6 An acquisition example: (a) the RGB image and (b) the mask computed by human (c) the mask computed automatically by SAM fine tuned on STANDUP dataset . . . 17
7 Final result: (a) the RGB image and (b) IR image . . . 18
8 ResNet18 scheme. . . . 19
9 LSNet(S) scheme. . . . 21
10 Double Encoder ResNet scheme. . . . 22
11 Late Fusion ResNet scheme. . . . 23
12 Stratified k-fold cross-validation diagram . . . 25
13 An acquisition example: (a) the RGB image, (b) the original image resized and (c) the final mask . . . 26
14 Division in angiosomes . . . 28
15 Box plots of absolute temperature differences ($|left - right|$) across anatomical regions (LCA, LPA, MCA, MPA) at baseline (T0). The dashed red line indicates the clinical threshold of 2.2°C. . . . 35
16 Box plots of absolute temperature differences ($|left - right|$) across anatomical regions (LCA, LPA, MCA, MPA) after 10 minutes (T10). The dashed red line indicates the clinical threshold of 2.2°C. . . . 36

List of Tables

1 Breakdown of patients in the dataset by clinical category. 13
2 Comparison of trainable parameters between ResNet-18 and ResNet-50. 19
3 Comparison of LSNet variants evaluated in this study, showing their full names and number of trainable parameters. 20
4 Summary of hyperparameters and training strategies for CNN models. 24
5 Final hyper-parameters selected for each model, after tuning on validation. 24
6 Detail of statistical feature extraction for thermal difference analysis $|left - right|$ across anatomical regions and time points. 30
7 Quantitative summary of the extracted feature dataset. 30
8 Performance comparison of CNN models (LSNet and ResNet variants) on binary classification using RGB and IR images. 31
9 Performance comparison of CNN models (LSNet and ResNet variants) on multi classification using RGB and IR images. 32
10 Balanced Accuracy comparison between RGB and IR modalities for LSNet and ResNet models, including performance deltas. 32
11 Balanced Accuracy results in a more challenging classification scenario, highlighting RGB vs IR performance gaps for LSNet and ResNet models. 32
12 Model Performance Comparison for RGB-IR Image Type (Rounded to Three Decimal Places) in binary classification. 33
13 Model Performance Comparison for RGB-IR Image Type (Rounded to Three Decimal Places) in multi classification. 33
14 Model Performance Comparison for T0 - T10 Image Type (Rounded to Three Decimal Places) in binary classification. 33
15 Model Performance Comparison for T0 - T10 Image Type (Rounded to Three Decimal Places) in multi classification. 34
16 Model Performance Comparison for SHAPE Image Type (Rounded to Three Decimal Places) in binary classification. 34
17 Model Performance Comparison for SHAPE Image Type (Rounded to Three Decimal Places) in multi classification. 34
18 Mean Left-Right Temperature Difference and Z-score at T0 for Segments with Very Significant Effect 36
19 Mean Left-Right Temperature Difference and Z-score at T10 for Segments with Very Significant Effect 37
20 Comparison of classification performance between different machine learning algorithms and data types. 37
21 Performance of models with features in binary classification (values rounded to three decimal places). 38
22 Performance of models with features in multi-class classification (values rounded to three decimal places). 39

23 Comparative analysis of performance variations in the image + features approach for binary classification. 40
24 Comparative analysis of performance variations in the image + features approach for multi-class classification. 40

1 Introduction

Diabetic foot ulcers (DFUs) are among the most prevalent and severe complications associated with diabetes, often resulting in chronic wounds and infections that frequently precede lower limb amputations. It is estimated that between 19% and 34% of patients diagnosed with diabetes will develop a foot ulcer at some point in their lifetime. Once a DFU occurs, healing is challenging – only about 30–40% of ulcers fully heal within 3 months – and recurrence is alarmingly high (roughly 40% of patients suffer ulcer recurrence within one year of healing). DFUs have been shown to significantly increase the risk of infection, amputation, and even mortality. Indeed, ulcers have been observed to precede approximately 80% of diabetes-related amputations, and the mortality risk for a patient with a DFU is more than twice that of a diabetic patient without an ulcer, over a period of five years. Consequently, DFUs constitute a significant medical concern, impacting millions worldwide annually and imposing considerable financial burdens on patients and healthcare systems [1, 2].

The pathophysiology of DFUs is multifactorial, with peripheral neuropathy playing a central role. The primary trigger for plantar ulceration is the loss of protective sensation in the feet, in combination with repetitive trauma or high plantar pressures from walking [3].
Decreased blood flow, resulting from peripheral arterial disease (ischemia), has been demonstrated to impede wound healing and contribute to ulcer development [4].
These neurological, vascular, and biomechanical factors create a dangerous cycle that can ultimately break down the skin [1].
Fortunately, major diabetic foot injuries are largely preventable with early identification and intervention. Comprehensive foot screening and patient education have been demonstrated to be effective measures for the significant reduction of ulcer incidence. The regular performance of medical examinations can facilitate early recognition of pre-ulcerative signs, such as the presence of calluses or redness. Timely management, which may include measures such as off-loading pressure and treatment for minor injuries, can impede the progression of such lesions towards the development of open ulcers [3].
Skin temperature monitoring has emerged as a promising tool for early warning of ulcer risk. Research by Armstrong et al [1]. demonstrated that the development of a foot ulcer is often preceded by local foot hyperthermia, defined as a temperature difference greater than 2.2°C between corresponding sites on the two feet. The presence of such a hotspot region is an indication of the potential for ulceration in the diabetic foot.
Furthermore, the results of a recent clinical trial demonstrated that patients who regularly monitor their foot temperatures and relieve pressure when hotspots are detected experience a roughly 50% reduction in the risk of ulcer formation in comparison with the standard care group (18.7% vs. 30.8% ulcer incidence) [2].
IR thermography thus offers a fast, non-invasive way to visualize subtle perfusion and inflammation changes in the foot, and interest in its use for diabetic foot care is rapidly growing.
In fact, experts suggest that incorporating IR imaging into routine diabetic foot screenings could further reduce ulcer occurrence.
Nevertheless, the difficulty lies in reliably detecting and interpreting these thermal patterns. In this regard, automation and machine learning can play a pivotal role.[5].

1.1 Machine Learning on Diabetic Foot

The diabetic foot is characterised by the development of chronic ulcers. In the absence of timely recognition and treatment, these ulcers can result in amputation in approximately 80% of cases, and occasionally even result in patient death. [6]. For this reason, computer vision systems based on neural network capable of automatically identifying and classifying ulcers from images of the foot have been developed in recent years. Accurate automatic diagnosis and rapid intervention can in fact prevent amputations and complications by early and reliable detection of ulcers through digital image analysis. In this chapter, we examine two fundamental aspects. Firstly, we consider the recognition of pre-existing ulcers with regard to detection, classification and analysis of injured tissue. Secondly, we explore the early recognition of ulcer risk signs before lesions develop, using thermographic and Red, Green, Blue (RGB) images.

Ulcer Recognition

Automated analysis of diabetic ulcers aims to detect the presence of lesions in the foot and assess their characteristics (extent, tissue status, signs of infection or ischaemia). The latest techniques exploit convolutional neural networks (CNNs) trained on databases of images of diabetic feet, achieving high performance in distinguishing healthy feet from feet with ulcers [7].
For example, Thotad et al. employed the EfficientNet network on a dataset of 844 images of feet (with and without ulcers), achieving an accuracy of 98.97% and outperforming classical models such as AlexNet, GoogleNet and VGG [8].
In addition to RGB images alone, it has been shown that integrating texture information can improve classification. Al-Garaawi et al. [9] propose to combine RGB channels with a mapped binary pattern image (a technique derived from Local Binary Patterns) that emphasises the texture of the ulcer. This texture image, combined with the RGB input, is fed to a customised CNN (DFU-RGB-TEX-Net) significantly improving performance in automatically recognising ulcers. In particular, the model obtains AUC of 0.981 (F-measure 0.952) in distinguishing an ulcerated foot from a healthy one, and also performs very well in recognising clinical signs such as ischaemia (AUC 0.995) and infection (AUC 0.820) in ulcer images.
Another crucial strand is the recognition of tissue within the wound, which is useful for monitoring healing. Semantic segmentation techniques with CNN are used to classify the different tissue types in the ulcer (e.g. necrosis, granulation, fibrin/slough). Niri et al. [10] developed an innovative approach based on superpixels combined with a fully-convolutional network, achieving a significant improvement over traditional models. Their system first automatically segments the ulcerated area in the image and subdivides it into superpixels (using the SLIC algorithm); it then classifies each superpixel into three tissue classes: necrotic, granulation, or fibrinous. This two-stage method outperformed classical end-to-end segmentation models (FCN, U-Net, etc.), increasing the accuracy from 84.5% to 92.7% and the DICE index from 54.3% to 75.7% compared to a standard FCN. This demonstrates the validity of incorporating fine-granular segmentation to assess the status of the ulcer floor and support clinical treatment decisions (e.g. targeted debridement of necrotic tissue) .
To automatically localise where an ulcer is located in the foot image (detection), object detection models are used. A recent study by Sarmun et al. [6] combined multiple detection networks by means of a bounding box fusion (WBF technique),

achieving a mAP = 86.4% in identifying ulcers on the DFUC2020 dataset. Notably, the ensemble of YOLOv8 and Faster R-CNN models (ResNet101) outperformed the previous benchmark on the same challenge by 12.4%. This ensemble approach exploits the complementarities of different architectures to improve the generalisation of results. Such systems can highlight the location of the ulcer on the image with a box or mask, acting as a triage tool for the physician. Overall, the use of deep neural networks has made possible a more objective and reproducible analysis of diabetic ulcers, assisting the clinician in the assessment of severity (e.g. by recognising signs of infection) and in the therapeutic indication (e.g. by monitoring the evolution of tissues towards healing).

Early Recognition of Ulcerative Risk

It is imperative to recognise the significance of preventing the formation of new ulcers in the management of the diabetic foot. The primary objective of early recognition is to identify predictive signs of ulceration prior to the manifestation of clinical symptoms, thereby enabling the implementation of preventive measures (e.g., rest, appropriate footwear). Skin temperature is one of the most extensively studied markers. Numerous studies have demonstrated a direct correlation between thermal abnormalities of the foot and the subsequent development of ulcers. In patients suffering from neuropathic pain, an area of significantly higher temperature (hyperthermia) on the sole of the foot typically indicates inflammation or mechanical overload, conditions that may precede ulceration. Conversely, the presence of cooler areas may be indicative of ischaemia. IR thermography has emerged as a promising tool to detect such temperature changes non-invasively and accurately. For instance, Salazar and Zequera [11] have demonstrated that thermography can highlight plantar thermal differences associated with ulceration risk, thereby improving the detection of early-stage ulcers. The utilisation of IR imaging cameras within a laboratory setting has demonstrated a high degree of sensitivity in identifying plantar regions that exhibit increased temperature, often coinciding with areas susceptible to ulceration if not adequately managed.

In parallel, standard RGB images of the foot can also offer early clues, such as changes in skin colour, calluses or minor wounds. However, these visible features are less predictive of ulceration than internal thermal patterns. Therefore, recent approaches combine both and RGB images to maximise accuracy. Reyes-Luévano et al. [12] developed a dual model called DFU-VIRNet that fuses Visible + IR information: this dual-channel network achieved an accuracy of 97.7% in correctly classifying the state of the foot. The integration of thermograms with RGB photos made it possible not only to recognise existing ulcers, but also to improve the detection of areas at risk of ulceration before the injury worsens. In practice, the model indicates abnormal regions (e.g. excessively hot) on the foot where the tissue could rupture, providing an early warning to the physician.

Numerous studies confirm the effectiveness of AI on IR images for early diagnosis. For instance, Munadi et al. [13] used plantar thermograms of diabetic patients by training two lightweight CNNs (ShuffleNet and MobileNetV2) and combining their decisions: this decision fusion led to correctly classifying 91% of foot cases with signs of incipient ulceration vs. healthy foot in their dataset. Customised CNN architectures were also tested: a parallel convolutional network approach on IR images showed superior performance to AlexNet and GoogleNet in distinguishing diabetic from normal feet. Such high results suggest that IR images contain very

distinctive signals of pathology (abnormal thermal patterns) that neural networks can effectively learn [14].
The combined use of thermographic and RGB images supported by deep learning algorithms represents an advanced frontier for the early recognition of diabetic ulcers. By identifying at-risk areas at an early stage - e.g. an area of hyperthermia indicative of underlying inflammation - these systems allow preventive interventions to be implemented before an established ulcer forms [6, 11]
This approach has the potential to significantly reduce the incidence of new ulcers and subsequent amputations by complementing traditional clinical monitoring with highly sensitive automated screening tools. Consequently, contemporary research underscores the potential of artificial intelligence (AI) in the domain of diabetic foot imaging, demonstrating its capacity to enhance patient outcomes by offering valuable decision support. This support encompasses the management of existing ulcers and the proactive prevention of future injuries.

1.2 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) represent a class of artificial neural networks that are particularly well-suited to the processing of data that possesses a grid structure (typically images). In a CNN, in a manner analogous to a traditional neural network, neurons receive input from preceding layers, execute an operation (e.g. a scalar product followed by a non-linear function) and iteratively self-optimize through the process of supervised learning. The fundamental difference between CNNs and classical fully connected networks is that CNNs exploit the local spatial information present in images by means of convolution and pooling operations. In these operations, groups of neurons are connected only to small adjacent regions of the image (local receptive fields) by sharing the same weights (filters). This facilitates the automatic extraction of hierarchical visual features, ranging from simple edges and textures to high-level patterns, thereby significantly reducing the number of parameters that need to be learned. Consequently, this enhances the efficacy of the network for computer vision tasks [15].
The typical structure of a CNN consists of a sequence of blocks organized in four main categories:

- Convolutional layers: The core of a CNN consists of the convolutional layers, which apply filters (or kernels) to small local regions of the input. Each filter runs over the image by calculating scalar products between the weights and the regions covered (receptive fields), thus generating feature maps (activation maps). Each filter learns to detect specific patterns such as edges, corners or textures, and does so over the entire image by weight sharing, i.e., using the same weights at multiple points [16]. This exploits local information: the spatial position of pixels is important, and filters allow similar patterns to be captured even at different positions, thus guaranteeing a certain translational invariance. The resulting feature maps are thus new "images" where each pixel represents the response of a filter applied to the same position in the original image.

- Activation layer (non-linearity): After each convolution, an activation function introduces nonlinearity, typically the ReLU (Rectified Linear Unit: $f(x) = max(0, x)$). This allows the network to learn complex relationships between

features, enabling the representation of much more complex functions than linear operators alone [17].

- Pooling layers (sub-sampling): Pooling layers reduce the spatial dimension of feature maps, keeping only the most relevant information in a local region (e.g. by max pooling, which selects the maximum value in each window). This reduces the number of parameters, introduces a form of robustness to small shifts or distortions and controls overfitting [16].

- • Fully connected layers: After several convolution–activation–pooling blocks, the feature maps are 'flattened' and passed to one or more fully connected layers, analogous to those in classical neural networks. These layers combine all learnt features to produce the final classification (e.g. by assigning probabilities to possible classes via a softmax function).

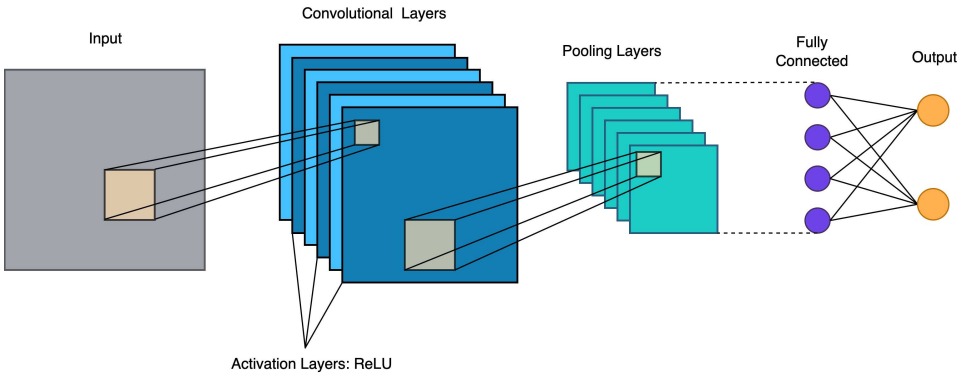


Figure 1: CNN structure

In a CNN, each neuron in the convolutional layers is connected to only a small region of the input (local receptive field), so that local relationships between pixels are captured. The same filter (i.e. the same set of weights) is applied over the entire image (weight sharing), drastically reducing the number of parameters compared to a dense network. The first layers specialise in detecting simple patterns such as edges or corners, while the deeper layers combine this information to extract increasingly abstract and complex features such as shapes, textures or objects. Learning occurs through supervised optimisation of filter weights: the model updates parameters to minimise a cost function (e.g. classification error), using gradient back-propagation to correct the output according to the desired labels [18].
Digital images have strong local correlations: neighbouring pixels tend to have similar values. By exploiting this structure, CNNs are able to capture relevant patterns reducing the risk of overfitting and improving generalisation. Due to the translational invariance introduced by weight sharing, CNNs recognise patterns regardless of their position in the image, and the hierarchical composition of features allows the recognition of complex structures [19].
Furthermore, the presence of pooling layers reduces sensitivity to small translations, rotations or distortions, making the model robust to variations typical of real images. This makes CNNs extremely effective in all major computer vision applications, from classification to segmentation
Although the principle of CNNs was formalised as early as the 1980s-1990s (with the Neocognitron [20]

and then LeNet-5 [21]), the deep CNN revolution started with AlexNet (2012) [22], the first to demonstrate the scalability of convolution on large datasets thanks to GPUs and innovations such as ReLU activation and dropout. VGGNet (2014) [23] introduced systematic depth and the use of filters 3×3; GoogLeNet /Inception (2015) [24] proposed multi-scale modules and parameter reduction via convolutions 1×1; ResNet (2016) [18] solved the problem of gradient degradation in deep networks with skip connections, allowing networks with more than 100 layers to be trained. Today, numerous variants exist (DenseNet [25], MobileNet [26], EfficientNet [27], etc.), but all are based on the structural principles described here.

1.3 Aim of the Thesis

The objective of this thesis is to evaluate the efficacy of deep learning neural networks in predicting patients with diabetic ulcers. The primary objective of this study is to evaluate the efficacy of IR images in enhancing the accuracy of prediction when compared to RGB images, utilising various deep learning models.
A statistical analysis will also be carried out to understand features other than just the images. The thermal mapping for each pixel from the IR images will be extracted, and machine learning methods will be used to attempt classification. The entire project was developed in Python to be easily reproducible and to take advantage of the extensive support available online. The structure of the thesis is outlined as follows. Chapter 1 provides a contextual foundation for the study, encompassing a review of machine learning for diabetic foot assessment and convolutional neural networks. It also delineates the overarching objective of the study. The second chapter of this thesis sets out the materials and methods that were employed in this study. These include the STANDUP dataset, the segmentation procedure, the CNN classifiers, the cross-validation strategy, and the statistical analysis. In Chapter 3, the results obtained are reported. These results include model performance, statistical findings and an examination of the hybrid approach that integrates deep learning with handcrafted features. In conclusion, Chapter 4 provides a detailed discussion of the results and draws definitive conclusions to wrap up the central argument of the thesis.

2 Materials and Methods

2.1 Dataset: STANDUP Project

The STANDUP dataset [28] was developed as part of the European Horizon 2020 STANDUP (H2020-MSCA-RISE Project 777661) project with the objective of supporting the early diagnosis of the diabetic foot using IR and RGB images of the sole of the foot. This database is unique in its scope and approach, as it is the first of its kind to encompass multiple spectral domains. It has been developed with the primary objective of facilitating research in the field of machine learning models. These models, in turn, are designed to identify the early signs of plantar hyperthermia and ulcerative risk.
The dataset under consideration consists of 415 image pairs (IR + RGB) acquired from the soles of the feet. Of these, 125 pairs are from healthy non-diabetic subjects and 290 pairs from type II diabetic subjects. Each pair consists of a IR image and a colour (RGB) image, both of which are captured simultaneously and spatially recorded on the same frame. The healthy subjects are volunteers recruited from two research centres (namely, the PRISME laboratory at the University of Orléans in France and the IRF-SIC laboratory at Ibn Zohr University of Agadir in Morocco). The healthy sample comprised 82 volunteers (35 male, 25 female, average age 30 years) from whom 125 multispectral images of the sole of the foot were obtained. The subjects with diabetes mellitus (DM) were recruited from the Hospital Nacional Dos de Mayo (HNDM) in Lima, Peru, where a dedicated acquisition campaign was conducted as part of a clinical study on the diabetic foot. A total of 145 type II diabetic patients (91 women, 54 men, average age 63 years) participated in the study, each subject undergoing two acquisitions (as described below) for a total of 290 IR images with corresponding RGB. A detailed breakdown of patients by category is provided in Table 1.

Category	**Number of patients**
Healthy	82
Diabetic "normal" (R0)	69
Diabetic neuropathy (R1)	34
Diabetic neuroischemic (R2)	42

Table 1: Breakdown of patients in the dataset by clinical category.

It should be noted that the normal diabetic foot is defined as that of a diabetic patient with no signs of significant peripheral neuropathy or peripheral arteriopathy. The neuropathic foot presents diabetic neuropathy (reduced or absent protective sensitivity), while the neuroischemic foot presents simultaneous peripheral neuropathy and peripheral vascular insufficiency. According to clinical evaluations conducted prior to acquisition, the sample of diabetic patients encompasses subjects in all three of these stages of clinical foot risk.

Instrumentation and acquisition modes

The images were acquired using a FLIR ONE Pro (2) handheld thermal imaging camera connected to a smartphone. This device integrates two sensors: an IR thermal sensor with 160×120 pixel resolution (8-14 μm spectrum) and sensitivity to temperature differences of 0.1°C, and a conventional 1440×1080 pixel RGB camera.

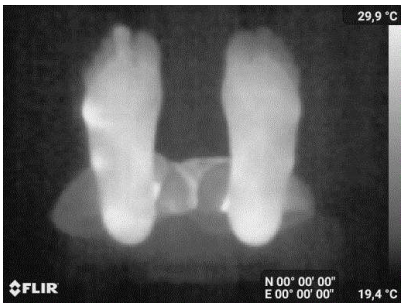


Figure 2: Image of FLIR ONE Pro.

Thanks to the Multi-Spectral Dynamic Imaging (MSX) technology integrated in the FLIR software, the details of the visible images are superimposed on the IR images, producing sharp IR images that are spatially co-registered with the corresponding RGB photos (Figure 3).



(a) RGB



(b) IR

Figure 3: An acquisition example: (a) the RGB image and (b) the corresponding IR image

In this way, a perfectly aligned thermogram and a photograph of the sole of the foot (same framed area) are obtained simultaneously for each subject, an important condition for possible combined analyses. The acquisition distance was chosen so as to frame both feet in full, respecting a margin of approximately 5cm around each foot. This results in an approximate field of view of 40×40 cm^2(Figure 4), which is sufficient to contain the soles of both feet side by side with adequate border.

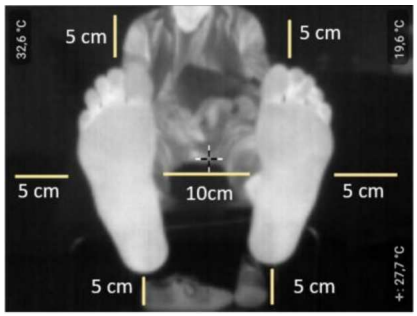


Figure 4: Example of a plantar IR image acquired with the FLIR ONE Pro thermal imaging camera.

Acquisition protocols

For healthy subjects, the acquisition of images was conducted in controlled environments with shielded lighting and a stable temperature (approximately 20 ± 1°C). Following the removal of shoes and socks and a 10-minute acclimatization period, the volunteers were placed in a supine or seated position on a couch, with their feet uncovered and spaced approximately 10cm apart. The IR and RGB images were acquired simultaneously by means of a smartphone connected to the FLIR ONE Pro thermal camera, which successfully framed both soles of the feet according to the expected margins. For a subgroup of 43 subjects, a cold thermal stress test was also performed, with images acquired both before and after the feet were immersed in cold water.
For subjects with diabetes, the protocol involved the use of a dedicated room with controlled light and temperature, following a 10-minute acclimatization period. Patients were positioned in a manner consistent with healthy volunteers and imaged using the same equipment. In addition to the initial acquisition (T0), a cold stress test was administered to all subjects. This entailed the immersion of the feet in water at 15°C for a duration of 60 seconds, followed by drying and a 10-minute rest period prior to the second acquisition (T10). The T0 and T10 images facilitate the evaluation of vascular response and thermal recovery, which is pertinent information for the identification of neuropathic or vascular deficits. The protocol was administered without incident and was conducted under the supervision of medical professionals [28].
The images contained within the STANDUP dataset are organised in a hierarchical structure, with the files divided into two primary folders. The terms "diabetic" and "healthy" are used to denote two discrete states. It is evident that each folder is subdivided into two subfolders. The first of these contains the RGB images, while the second contains the IR images. In the case of subjects with diabetes, an additional subdivision is made according to the clinical stage of the disease:

- R0: Diabetics devoid of neuropathy
- R1: neuropathic diabetics
- R2: Diabetics with a neuroischemic etiology

The nomenclature employed for the images is consistent with the folder to which they belong, with the images being identified by an identification code (e.g. "R0_001" identifies the first image of a normal diabetic subject). This configuration is designed to facilitate navigation and utilisation of the data, thereby enabling straightforward selection for comparative analyses across diverse clinical categories.
The dataset is freely accessible on the official website of the STANDUP project, along with the technical documentation and references to all scientific papers that have utilised it.

2.2 Segmentation

The removal of the background from the foot images was adopted as a pre-processing step to enhance the performance of the classification network. A complex or variable background has the potential to introduce noise and confounding elements, thereby diverting the attention of the CNN from the region of interest (the foot itself). Studies in the literature have confirmed that the isolation of the primary object from

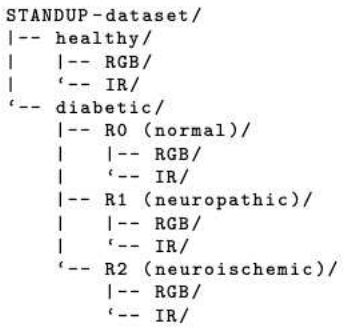


Figure 5: Directory structure of the STANDUP dataset.

its background can significantly increase the accuracy of a classifier. Furthermore, faster convergence during training has also been observed [29]. After those considerations, it was determined that the images would be segmented in order to exclude the background, which was uniformed to black. This approach was adopted to ensure that only relevant information was presented to the classification model.

Initially, manual masks were created for each image in the dataset, delineating the outlines of both feet by hand. This resulted in binary masks (in shades of black and white) in which the foot region is shown in white while the background is black. The manual segmentation thus provided a ground truth for training an automatic segmentation model. However, the exclusive utilization of hand-drawn masks for all images can engender issues of bias and potential data leakage. For instance, the human operator may unconsciously provide insights about the foot's condition during segmentation or segment images from different classes somewhat differently. In order to circumvent such risks and enhance scalability, an automated solution founded upon an advanced segmentation model was implemented.

The model employed for the purpose of automatic segmentation is SAM (Segment Anything Model), a development of Meta AI (by Kirillov et al. [30]). SAM is a universal segmentation model that has been trained on a dataset comprising over 1 billion masks from approximately 11 million images. This dataset is the largest segmentation dataset available to date. The model under discussion is of the foundation-type, and has been designed to be promptable. By this it is meant that it is able to generalise to new objects and new domains without further training, once an appropriate prompt (e.g. one or more reference points on the object to be segmented) has been provided. The architecture of SAM is composed of a powerful image encoder and a light mask decoder conditioned by prompts. This configuration enables the production of segmentation masks in real time (approximately 50 milliseconds per mask). Additionally, the system possesses the capacity to address ambiguous scenarios by predicting multiple masks if necessary.

In the present workflow, manual masks were utilised as ground truth to train SAM on the specific domain of foot images. In practice, the pre-trained SAM model was fine-tuned on the dataset using the initial binary masks, with the objective of specializing it in the separation of the feet from the background. Subsequently, the trained version of SAM was applied to all images (both RGB and IR) in order to automatically generate the final segmentation masks without the need for manual

intervention. The performance of the segmentation model was evaluated using the Intersection over Union (IoU) 1 and DICE score 2 metrics, which measures the overlap between the predicted and ground truth masks

$$IoU = \frac{|A \cap B|}{|A \cup B|} = \frac{TP}{TP + FP + FN} \tag{1}$$

$$DICE = \frac{2 \cdot |A \cap B|}{|A| + |B|} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \tag{2}$$

where:

- A represents the ground truth mask, B the predicted mask;
- TP are true positives (correctly predicted foot pixels);
- FP are false positives (background pixels incorrectly predicted as foot);
- FN are false negatives (foot pixels incorrectly predicted as background).

IoU values closer to 1 indicate better segmentation accuracy, while higher DICE values reflect stronger agreement between predicted and ground truth masks. In this study, both metrics yielded consistently high scores across the dataset, confirming the robustness and reliability of the automated segmentation approach.
The process is illustrated in Figure 6, which shows the progression from the original image to the hand-drawn binary mask (white-black) and finally to the mask automatically predicted by SAM. It is evident that the mask predicted by SAM coincides almost exactly with the actual foot contour.



(a) RGB (b) Manual mask (c) SAM mask

Figure 6: An acquisition example: (a) the RGB image and (b) the mask computed by human (c) the mask computed automatically by SAM fine tuned on STANDUP dataset

This approach has several advantages. Firstly, the employment of SAM to create the masks circumvents the potential for data leakage that may arise from manual segmentation. The validation and test images are fully segmented automatically, thereby ensuring that no unauthorized information or human bias is introduced into the data provided to the classifier. Additionally, the observation that the automatic masks do not perfectly align with hand-drawn masks, introduces a modicum of realism and variability. This can, in turn, enhance the robustness of the final classification model. Therefore, upon completion of the procedure, two segmented versions of each image are obtained (one for the visible RGB channel and one for the IR channel) in which only the feet are shown against a completely black background (Figure 7). These segmented images constitute the final input for the training and evaluation phase of the classification CNN (see next section), ensuring that the model focuses exclusively on foot characteristics.



(a) RGB Segmented (b) IR Segmented

Figure 7: Final result: (a) the RGB image and (b) IR image

2.3 CNN Classifiers

In this study, two primary CNN architectures, ResNet and LSNet, were examined through the evaluation of several variants:

- ResNet-18
- ResNet-50
- LSNet-T/S/B

Furthermore, the exploration of models incorporating dual inputs was conducted to combine information from both RGB and IR images. For each architecture under consideration, experiments were conducted on both RGB and IR thermographic images independently. Additionally, the efficacy of "multi-temporal" configurations was evaluated by employing two successive acquisitions (T0 and T10) as inputs to a single model. In practice, the two images were stacked on channels (obtaining a 6-channel image for RGB and a 2-channel image for IR) so that the network could learn from temporal differences between consecutive frames. Shape classification based on B/W masks was also performed to understand whether the shape of the foot can be a discriminant. Finally, for the fusion of both modalities (RGB+IR), two distinct approaches were implemented: early fusion by means of a dual encoder model (simultaneous input of both modalities in one architecture) and late fusion by means of a decision-level combination of two separate networks.

ResNet

The Residual Networks (ResNet) proposed by He et al. [18] enabled the training of significantly deeper networks than traditional CNNs through the incorporation of residual blocks with skip connections (identity connections). In a ResNet block, the output of convolutional layers is added directly to the input of the block (shortcut operation), allowing the network to learn a residual function rather than a full mapping. This approach alleviates the problem of performance degradation as network depth increases. The authors demonstrated that a 34-layer ResNet achieves lower error rates than an 18-layer version, reversing the trend previously observed in "plain" networks without skip connections. Skip connections facilitate gradient flow throughout the network, mitigating the vanishing gradient problem and enabling successful training of very deep architectures.
In the present study, two standard ResNet variants were utilized:

- ResNet-18:

  - Architecture: 17 convolutional layers and 1 fully connected layer
  - Structure: 8 "basic" residual blocks
  - Block composition: Each block contains two 3×3 convolutions

- ResNet-50:
  - Architecture: 50 layers total
  - Structure: Incorporates bottleneck-type residual blocks
  - Block composition: Each block contains three layers, 1×1 convolution (dimension reduction), 3×3 convolution (feature extraction) and 1×1 convolution (dimension restoration)

A comparison of the number of trainable parameters between the two variants is reported in Table 2.

Variant	N **parameters**
ResNet-18	11.2×10^5
ResNet-50	23.9×10^5

Table 2: Comparison of trainable parameters between ResNet-18 and ResNet-50.

Both ResNet variants share the following characteristics:

- Normalization: Batch Normalization (BN) layers after each convolution
- Activation: ReLU activation functions following each convolution
- Final layers: Global average pooling layer and Fully connected classifier
- Training approach: Models were initialized with ImageNet pre-trained weights and subsequently fine-tuned on the study dataset

The bottleneck design in ResNet-50 represents a more sophisticated approach compared to ResNet-18's basic blocks, allowing for increased network depth while maintaining computational efficiency through the strategic use of 1×1 convolutions for dimensionality management.

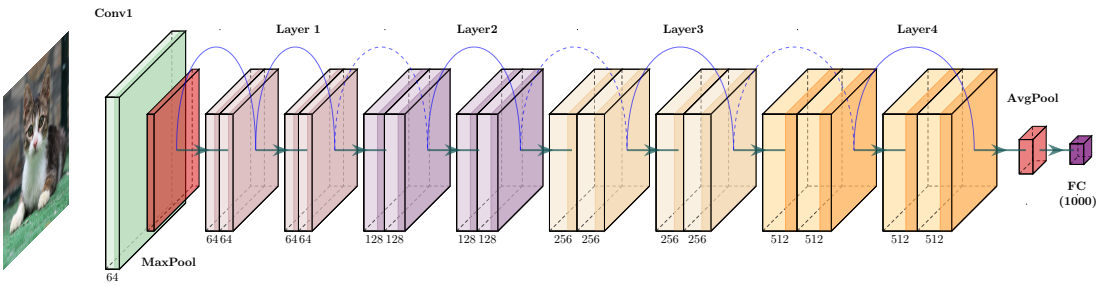


Figure 8: ResNet18 scheme.

LSNet

LSNet (Large-Small Network) is a family of lightweight CNN (Convolutional Neural Network) models proposed by Wang et al. [31]. These models are specifically designed to maximize computational efficiency while maintaining high performance through innovative architectural design. The fundamental concept underlying LSNet is the novel LS convolution, which implements the "See Large, Focus Small" strategy:

- Large-kernel perception: each block executes a separable large-windowed convolution to capture extensive contextual information;
- Small-kernel aggregation: subsequently focuses on relevant local details through small-kernel operations.

The LS convolution combines:

- Large-kernel depth-wise convolution (e.g., 7×7) to perceive broad spatial regions;
- Dynamic weights applied to smaller regions (e.g., 3×3) by clustering channels into sub-groups;
- Efficient integration of both global and local information.

The LSNet architecture follows a hierarchical design that begins with an initial stem layer performing patch embedding through stacked convolutions. The network then progresses through four distinct stages with decreasing spatial resolution. In the first three stages, multiple LS blocks are stacked to process features at progressively reduced resolutions, leveraging the "See Large, Focus Small" strategy for efficient feature extraction. The fourth and final stage represents a key architectural innovation: when spatial resolution becomes very low, LSNet transitions from LS convolutions to Multi-Head Self-Attention (MHSA) blocks. This strategic replacement allows the network to capture global dependencies at the final representation level, where the reduced spatial dimensions make self-attention computationally feasible and effective. This hybrid structure, combining LS convolutions for early-stage processing with MHSA for final-stage global modelling, enables LSNet to achieve an excellent trade-off between accuracy and inference speed, consistently outperforming other efficient models with comparable computational complexity.
In this study, three LSNet variants of different computational capacity were evaluated, as summarized in Table 3.

Variant	**Full Name**	N **parameters**
LSNet-T	LSNet-Tiny	12.15×10^5
LSNet-S	LSNet-Small	16.88×10^5
LSNet-B	LSNet-Base	24.65×10^5

Table 3: Comparison of LSNet variants evaluated in this study, showing their full names and number of trainable parameters.

This strategy involves the execution of a separable large-windowed convolution by each block to capture a large context (large-kernel perception), followed by a small-kernel aggregation that focuses on relevant local details (small-kernel aggregation).

In practice, the LS convolution combines a large-kernel depth-wise convolution (e.g. 7×7) to perceive a large region, with dynamic weights that are then applied to small regions (e.g. 3×3) by clustering the channels into sub-groups, thus efficiently integrating global and local information. Each LS block incorporates a residual skip connection (as in ResNet) to facilitate optimisation, an additional depth-wise convolution and SE layer (Squeeze-and-Excitation) to introduce local bias, and a small feed-forward network (MLP) for channel fusion.

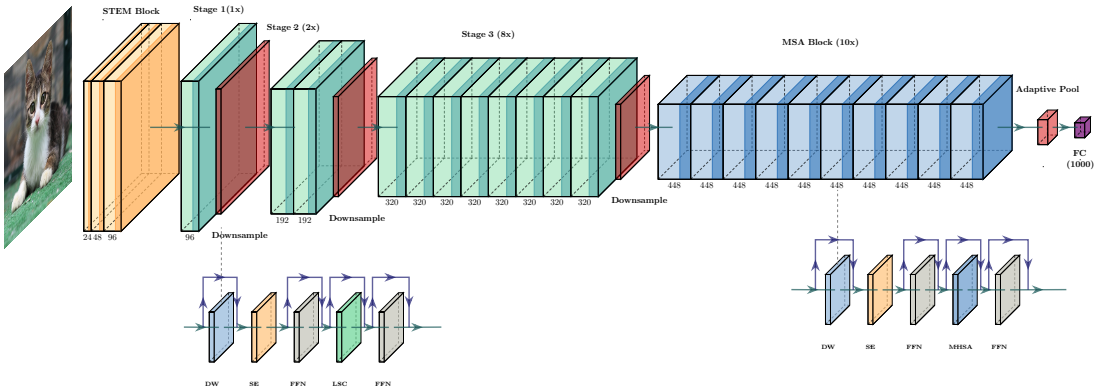


Figure 9: LSNet(S) scheme.

Each LSNet variant was also evaluated using two distinct adaptive convolution functions to assess their performance characteristics under different computational paradigms:

- The **Spatial Kernel Attention (SKA)** function implements a GPU-accelerated mechanism that applies spatially-adaptive kernel weights to input feature maps. This approach enhances the "Focus Small" component of LSNet by providing fine-grained spatial attention at the kernel level, where weights are dynamically adjusted based on local spatial context. The implementation leverages Triton kernels for optimal GPU utilization and supports mixed precision training.
- The **Channel-Wise Dynamic Convolution (CWDC)** function offers an alternative approach through channel-specific kernel adaptation. Unlike SKA's spatial focus, CWDC generates dynamic convolution kernels independently for each channel using a compact predictor network. This mechanism enables flexible feature extraction that adapts to varying channel-specific patterns while maintaining computational efficiency through grouped convolutions.

Both evaluation functions provide complementary perspectives on adaptive convolution mechanisms: SKA emphasizes spatial attention and position-specific weight adaptation, while CWDC focuses on channel-wise dynamic behavior and input-responsive kernel generation. This dual evaluation approach allows for comprehensive assessment of how different adaptive strategies affect the performance of each LSNet variant within the "See Large, Focus Small" architectural framework.
Similar to the ResNet models, LSNet variants were initialized with ImageNet pretrained weights provided by the original authors and subsequently fine-tuned on the study dataset. This transfer learning approach leverages the robust feature representations learned on large-scale data while adapting to the specific characteristics of the target dataset.

Dual input models (RGB + IR)

When the image RGB and the corresponding IR image are available simultaneously, both modes can be exploited to try to improve classification. Two fusion strategies were implemented and compared: (a) the Double-Encoder ResNet model, in which a single network integrates the two inputs via dual coding paths, and (b) the Late Fusion ResNet model, in which the two modes are processed separately and combined only at the decision level.

The **Double-Encoder ResNet** (DE-ResNet) model draws inspiration from the Double Encoder concept, as pioneered by Bouallal et al. [32] for the purpose of segmenting RGB+IR plantar images. In practice, DE-ResNet employs two parallel convolutional feature extractor branches (both based on shared ResNet architecture): one receives as input the RGB image, the other the IR image. Each branch extracts latent features from its mode through a series of residual blocks. The high-level representations of the two branches are then fused, for example by concatenation or summation, and passed to common fully connected layers (the fusion layer and the final output), which produce the combined class prediction. The training process is conducted in a manner that facilitates end-to-end learning, thereby enabling the network to collectively acquire the complementary characteristics of both sources (i.e., colour and temperature). This collaborative learning approach is undertaken with the objective of enhancing the classification performance. This approach, termed "early fusion", leverages the integration of features already present at the intermediate representation level, with the potential to capture correlations between RGB and IR. In this study, the dual encoder structure was implemented through the duplication of the ResNet backbone (starting with ImageNet pretrain weights for the RGB branch and random or pretrain weights on IR datasets for the IR branch, if available) and the merging of the two feature vectors prior to the final layer.

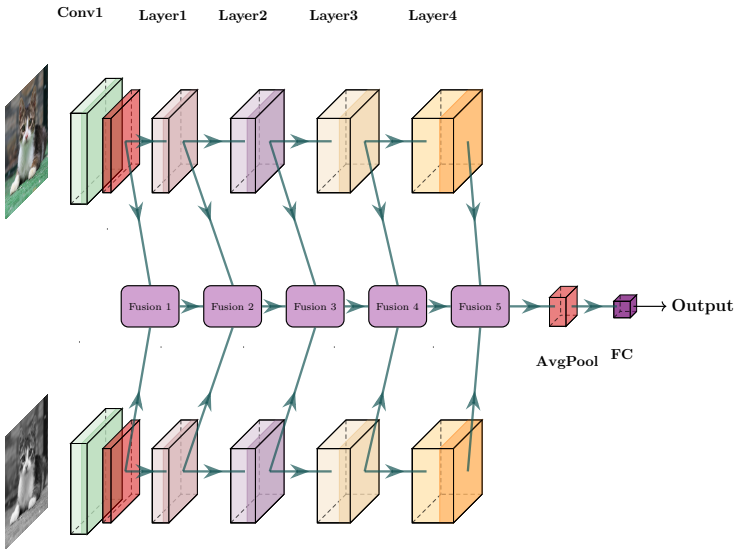


Figure 10: Double Encoder ResNet scheme.

The **Late Fusion ResNet** refers two distinct models that are employed, with one

trained on RGB images and the other on IR. The outputs of these models are combined after individual inference. In practice, late fusion involves the aggregation of the predictions of the two unimodal networks, for example by averaging the probabilities or using a simple classifier on the output concatenates. In this study, we investigated the efficacy of two distinct methodologies for the aggregation of output probabilities. The first approach involved the arithmetic averaging of the output probabilities, while the second entailed the concatenation of the penultimate feature vectors, followed by the addition of a small fully connected layer to produce the final decision. In contrast to the previous model, in this model the networks for each mode do not share weights or layers during training; they are trained separately on their respective domain, and the combination occurs post-hoc. This approach is characterised by its simplicity and modularity, with each network specialising in a distinct modality. However, it is important to note that this method may not adequately capture subtle inter-modal relationships if these relationships are not clearly manifested in the individual predictions. The extant literature suggests that late fusion can yield favourable results when the modalities are heterogeneous, with the fusion phase mediating between models that have already been optimised for each domain. In the context of multimodal medical imaging, this approach also ensures greater robustness against possible inconsistencies between channels (as each model can be independently validated on its own modality) [33].

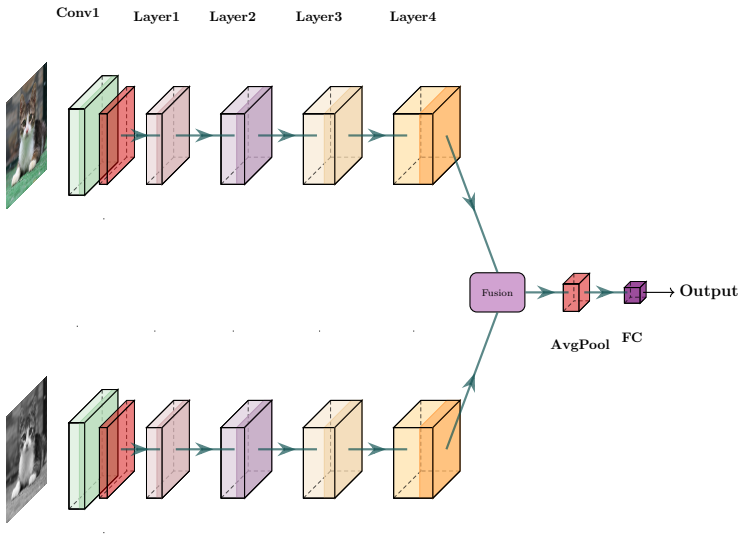


Figure 11: Late Fusion ResNet scheme.

Hyperparameters and Training Setups

We conducted extensive research on hyperparameters to optimise the performance of the models described above. Table 4 summarises the main hyperparameters explored and the values considered. The selection process was carried out through a grid search over the listed ranges. The initial choice of candidate values was primarily guided by prior personal experience and by insights obtained from the original papers describing the architectures tested. Below we list the categories of hyperparameters and the range of values tested:
All experiments were conducted using the same loss function (categorical cross-entropy), class balancing strategy (inverse-frequency class weights computed on the

Parameter	Values Tested
Optimizer	Adam, AdamW, SGD
Learning Rate (LR)	1×10^{-3}, 5×10^{-4}, 1×10^{-4}
Batch Size	8, 12, 16, 24, 32, 64, 128
Epochs	20+30 (two-phase), 30–50
Scheduler	CosineAnnealingLR, ReduceLROnPlateau
Dropout (FC layers)	None, 0.3, 0.4, 0.5
Weight Decay (L2)	0.001, 0.005, 0.01, 0.05

Table 4: Summary of hyperparameters and training strategies for CNN models.

training set and applied in the loss function), and hardware (NVIDIA A100 GPU with 80 GB VRAM).
After a hyperparameter tuning phase, different sets of optimal hyperparameters were selected for each model, taking into account architectural specificities and types of input data. These settings reflect the specific optimisation of each model and input type, as summarised in the table below.

Model	Batch	Epochs	Opt.	LR	W.Decay	Scheduler	Dropout
DE-ResNet	16	50	AdamW	1e-4	0.05	Red.LROnPlat.	0.3
LF-ResNet	16	50	AdamW	1e-4	0.05	Red.LROnPlat.	0.3
ResNet50	16	20	SGD	1e-4	0.01	Cos.Ann.LR	0.5
ResNet18	16	20	AdamW	1e-4	0.01	Cos.Ann.LR	–
LSNet (t/s/b)	64	20+30	AdamW	1e-3/1e-4	0.01	Cos.Ann.LR	–
ResNet18 (T0-T10)	32	30	AdamW	1e-3	0.05	Cos.Ann.LR	–
ResNet50 (T0-T10)	32	30	AdamW	1e-3	0.05	Cos.Ann.LR	–

Table 5: Final hyper-parameters selected for each model, after tuning on validation.

2.4 Dataset Splitting and Cross-Validation

In order to robustly evaluate the performance of the classifiers and avoid overfitting, a cross-validation strategy with splitting by patient was adopted. Indeed, it is imperative to segment the data at the patient level (rather than at the individual image level) to prevent data leakage: different images of the same subject must never fall into the training and testing sets simultaneously, otherwise the model will learn patient-specific features instead of generalising. For the purpose of this study, all images were grouped by patient and assigned to the various splits.
In detail, the dataset was divided as follows:

- External test set (hold-out): We initially reserved 15% of the total patients as the final test set. All images (RGB and IR, all timepoints) of these patients were set aside and not used in any training or validation phase, ensuring a truly independent final evaluation.

- Training/Validation cross-validation (85% patients): The remaining patients (i.e. 85% of the total) were used for an in-house 5-fold cross-validation. In each fold, approximately 70% of the total is used for training and the remaining 15% for validation. We created 5 different subdivisions by moving the block of patients in validation each time, so that overall validation covered 75% of the patients in the dataset (every patient in the 85% training appeared in validation in one of the folds). Important: the validation folds were defined at

the patient level, so a patient is only present in one validation fold and in no other, avoiding any overlapping.

- Each model was trained on the corresponding training data and evaluated on the external split test associated with that fold. After the five cross-validation iterations had been completed, the test metrics were aggregated by calculating the mean and standard deviation across the five tests. This resulted in a robust estimate of test performance. This ensures a fair comparison between models, as each was tested on the same external splits with no overlap with training or internal validation data.

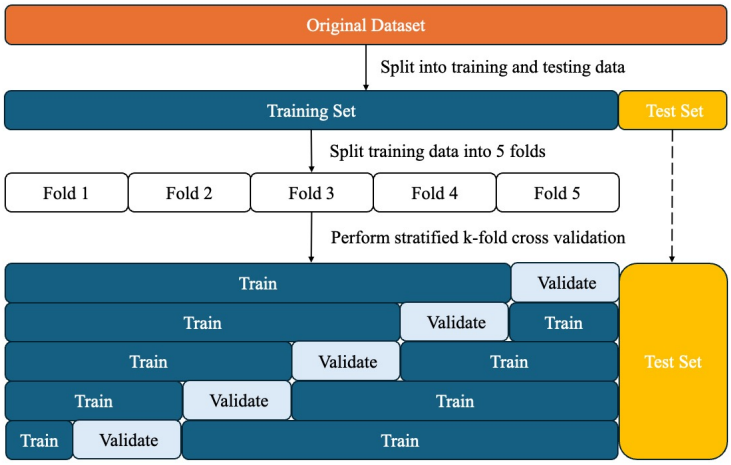


Figure 12: Stratified k-fold cross-validation diagram

During the preparation of the data, the visual information was preserved without the introduction of distortions. The original images exhibited heterogeneity in both dimensions and orientation, with some images captured vertically and others horizontally. In order to equalise the input to the network, we adopted an aspect-ratio preserving resizing approach. This entailed resizing each image to a square of approximately 640x640 pixels, with the shorter side being compensated for by adding black padding bands until the desired size was reached (the process is explained in Figure 13). This aspect-ratio preserving resize approach (analogous to letterboxing) avoids "stretching" or deforming the image while maintaining the original proportions. Subsequent to this operation, all images (RGB and IR) are centred on a black background in a 224×224 square. The pixels were normalised by scaling them to a value of 0.1 and then standardising them according to the mean and standard deviation of the RGB channels as determined by ImageNet. A similar process was repeated for the IR channel.

In order to enhance the variety of training data and improve generalisation, a number of data augmentation techniques were applied during the training phase. In particular, photometric transformations such as colour jitter (i.e. random variation of brightness, contrast and saturation) were used on RGB images. Geometric transformations such as random horizontal flip (i.e. RandomHorizontalFlip) were also employed on both modes. The flip probability was set to 50 per cent, while the colour jitter was calibrated with moderate changes so as not to excessively alter clinically relevant features. Augmentation was applied on-the-fly to each epoch



(a) RGB (b) Black bands (c) SAM mask

Figure 13: An acquisition example: (a) the RGB image, (b) the original image resized and (c) the final mask

on each batch, so that slightly different views of each image were presented to the network at each time step. This approach led to a substantial expansion of the training space, thereby reducing the occurrence of overfitting, a phenomenon that is particularly salient in cases where the available dataset is limited in scope.
The performance of the classifiers was evaluated using the following metrics: balanced accuracy (3), recall (4), precision (5) and F1-score (6) on each fold and on the final test [34]. In particular, balanced accuracy was chosen as the main metric because it considers accuracy on positive and negative classes equally, being appropriate under conditions of class imbalance (it is defined as the average between sensitivity and specificity). In addition, the precision and recall for the positive class of interest, as well as the F1-score representing its harmonic mean, are reported. This approach provides a comprehensive assessment of the models' efficacy in preventing false negatives and containing false positives. The metrics were calculated both at the cross-validation level (average over the validation folds) and on the external test set

$$BalancedAccuracy_k = \frac{1}{K}\sum_{k=1}^{K}\frac{\mathrm{TP}_k}{\mathrm{TP}_k + \mathrm{FN}_k} \tag{3}$$

$$Recall_k = \frac{\mathrm{TP}_k}{\mathrm{TP}_k + \mathrm{FN}_k} \tag{4}$$

$$Precision_k = \frac{\mathrm{TP}_k}{\mathrm{TP}_k + \mathrm{FP}_k} \tag{5}$$

$$F1Score_k = \frac{\mathrm{Precision}_k \cdot Recall_k}{\mathrm{Precision}_k + \mathrm{Recall}_k}. \tag{6}$$

2.5 Statistical Analysis

Extraction of thermographic map from IR images

for each thermographic image acquired, an absolute temperature map was extracted using the thermal scale provided in the image margin. Specifically, the minimum and maximum temperatures indicated in the image were utilised to perform a linear calibration of the pixel intensity values (in greyscale) to corresponding temperature values. The linear transformation function is defined as:

$$T(x,y) = T_{\min} + \frac{I(x,y) - I_{\min}}{I_{\max} - I_{\min}} \cdot (T_{\max} - T_{\min}) \tag{7}$$

where:

- $T(x, y)$ is the calibrated temperature value at pixel coordinates (x, y) in °C
- $I(x, y)$ is the grayscale intensity value at pixel coordinates (x, y)
- $I_{\min}$ and $I_{\max}$ are the minimum and maximum intensity values in the image
- $T_{\min}$ and $T_{\max}$ are the minimum and maximum temperature values indicated in the thermal scale

Consequently, it was feasible to correlate each pixel in the image with its temperature value (in °C). Segmentation was then employed to isolate the foot region by eliminating the background, resulting in a heat map that is a matrix of the same size as the original image. This heat map contains numerical temperature values for pixels belonging to the foot and NaN values for background pixels (i.e. pixels that do not belong to the foot). This procedure was applied to every available IR image. Moreover, given the presence of two feet in each thermogram, each image was separated into the two components corresponding to the right and left foot. The images of the individual feet were then realigned and oriented so that the feet were positioned in parallel with consistent proportions. This pre-processing step was necessary in order to subsequently perform a consistent division into vascular regions (angiosomes) on all subjects.

Regional analysis by angiosomes

The analysis focused on the regional thermal distribution within the foot, subdividing each foot into anatomical areas roughly corresponding to plantar angiosomes. The angiosomes of the foot are defined as the vascular regions irrigated by specific arteries (e.g. medial and lateral plantar angiosome from the posterior tibial artery, calcaneal angiosomes from the fibular and posterior tibial artery, etc.). It has been proposed by several authors that these anatomical divisions can be used as a reference for the thermographic analysis of the diabetic foot. This is with a view to detecting localised changes in perfusion or inflammation that may be masked in the analysis of the whole foot. In this study, an algorithm was implemented for the automatic subdivision of the foot mask into four main regions: medial forefoot, lateral forefoot, medial hindfoot and lateral hindfoot [35] referred to as:

- MPA: Medial Plantar Angiosome;
- LPA: Lateral Plantar Angiosome;
- MCA: Medial Calcaneal Angiosome;
- LCA: Lateral Calcaneal Angiosome.

In summary, given the configuration of the foot (binary mask), the algorithm delineates a horizontal line transverse to a specified proportion of the foot's length (the cut of the heel, demarcating the separation between the forefoot and the calcaneal region) and a vertical median line along the longitudinal axis of the foot (distinguishing the medial from the lateral side, inverted for the right and left foot). This results in a subdivision of the body into four anatomical quadrants. The regionalisation procedure by angiosomes draws inspiration from analogous methodologies outlined in the extant literature, yet it is adapted to the images in question and visually validated on each patient to ensure correct separation of the regions (the

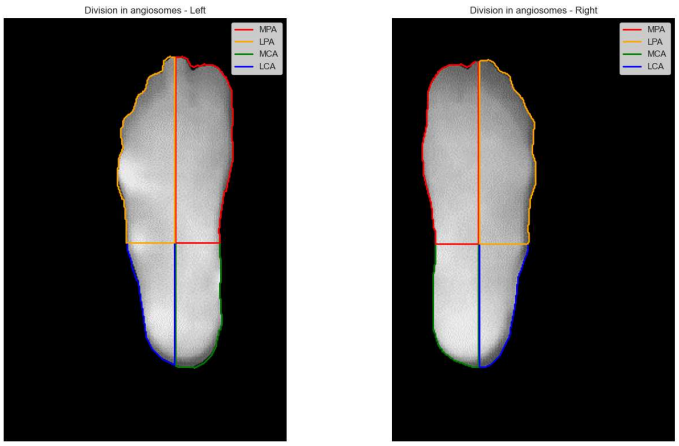


Figure 14: Division in angiosomes

feet have been previously realigned with precision to ensure a consistent division in terms of percentage area).

For each of the four angiosome regions of each foot, mean temperatures and other local statistical indicators (e.g. regional maxima, minima) were calculated. We assessed, for each patient, whether regional differences above the threshold of 2.2°C emerged between corresponding angiosomes. This threshold of 2.2°C is extensively documented in the literature as an indicator of potential ulcerative risk, with numerous studies on neuropathic diabetic patients adopting a difference >2.2°C as an alarm signal for an incipient ulcer. This criterion is derived from evidence that areas of the foot that will develop ulcers present, approximately one week earlier, a significantly higher local thermal elevation (up to 5 times higher) than corresponding areas [36]. The use of the threshold at the regional level makes it possible to identify potentially significant localised asymmetries that might be attenuated or not evident when considering the foot as a whole. It is important to note that recent studies have cautioned against the use of a single measurement greater than 2.2°C as an isolated predictor. For instance, Featherston et al. [37] concluded that a single finding of a difference greater than 2.2°C, taken in isolation, is not sufficiently valid to predict an ulcer (only approximately 0.4% of measurements were confirmed on the following day), although validity improves by requesting confirmation on subsequent days. Nevertheless, the criterion remains a useful indicator of potential thermal abnormality for investigation, and in the present context it served to identify regions with strong thermal asymmetries.
A series of descriptive statistical features were extracted from the thermal maps to quantitatively characterise the thermal profile of each region. The mean temperature of the plantar area was calculated for each angiosome region and for each time (T0, T10). In addition, the standard deviation (std) and the combined standard deviation (pooled std) were calculated, as were the 5th and 95th percentiles (p05, p95) of the thermal distribution. Furthermore, the skewness and kurtosis index of the distribution were calculated, as were the inter-decile range (IDR, difference between the 95th and 5th percentiles) and a z-score (8) of thermal anomaly. The z-score was calculated as an additional confirmation metric to quantify the degree of temperature asymmetry between corresponding angiosome regions, with higher values indicating deviations that exceed normal variability. Unlike a fixed temperature threshold, the z-score is a dimensionless measure that ranks asymmetry severity;

values of 2 or 3 correspond to differences approximately 2–3 standard deviations from the mean, reflecting relatively uncommon occurrences. In our context, we used it qualitatively: larger z-scores flagged more pronounced anomalies.

$$ZScore = \frac{|p95_1 - p95_2|}{\sqrt{(std_1)^2 + (std_2)^2}} \tag{8}$$

Where $p95_1$ and $p95_2$ represent the 95th percentiles of corresponding angiosome regions, and std_1 and std_2 the respective standard deviations.
In particular, indicators such as contrast, homogeneity, energy and thermal entropy, among others, were calculated based on the Gray-Level Co-Occurrence Matrix (GLCM) of the foot map. The integration of elementary statistical characteristics and texture GLCM features aligns with established methodologies documented in the extant literature for the differentiation of diabetic feet from healthy feet through the analysis of thermograms. These regional features were then combined (e.g. by considering absolute differences between homologous regions) and provided as input to the supervised classification algorithms (machine learning models) to ascertain the feasibility of discriminating between distinct patient groups based on their thermal profile. In particular, the study trained and compared a Random Forest classifier, Support Vector Machine (SVM) and a model based on TabPFN (Tabular Prior-Fitted Network) [38], a recent Transformer-type deep learning model pre-trained specifically for tabular data. This model allows accurate classifications on small datasets without requiring hyperparameter optimisation. The expectation was that subdivision into angiosomes, by isolating potential localised thermal "hot spots", could provide more sensitive features to pathological changes, helping to increase the accuracy of classifiers based on both image and thermometric tabular data.

No artificial class balancing was applied in the training of the models on the extracted features; however, given the possible imbalance in the number of subjects per category, the performance of the classifiers was evaluated using balanced metrics (i.e. balanced accuracy) to provide a more unbiased estimate of accuracy.
It should also be noted that some patients were excluded from the analysis at this stage as they had thermographic maps with invalid data due to camera acquisition errors; however, the remaining data included representatives of all patient categories in the study (no clinical categories were omitted from the statistical analysis).

Features Extraction

The feature extraction process resulted in the generation of a comprehensive dataset comprising 120 statistical features, derived from the analysis of absolute thermal differences between left and right feet. As demonstrated in Table 6, features were extracted from four anatomical regions (LCA, LPA, MCA, MPA) at two time points (T0, T10), resulting in 60 features per time point.

Category	Metrics	Regions	Time
Basic Statistics	Mean	LCA, LPA, MCA, MPA	T0,T10
	Standard Deviation		
	Pooled Std		
	5th Percentile		
	95th Percentile		
Distribution Shape	Skewness		
	Kurtosis		
Advanced Metrics	Inter-Decile Range		
	Z-Score		
Texture GLCM	ASM		
	Contrast		
	Correlation		
	Dissimilarity		
	Energy		
	Homogeneity		

Table 6: Detail of statistical feature extraction for thermal difference analysis $|left - right|$ across anatomical regions and time points.

The extracted features are divided into four primary categories: basic statistics (mean, standard deviation, percentiles); distribution shape features (skewness, kurtosis); advanced metrics (inter-decile range, z-score); and texture features based on the Gray Level Co-occurrence Matrix (GLCM). This multidimensional approach facilitates the capture of both global and local characteristics of thermal distributions. A quantitative breakdown of the extracted dataset is provided in Table 7

Item	Count
Total Features	120
Categories	4
Metrics	15
Time Points	2 (T0, T10)
Anatomical Regions	4 (LCA, LPA, MCA, MPA)
Features by Time Point	60
Features by Regions	30

Table 7: Quantitative summary of the extracted feature dataset.

3 Results

3.1 Deep Learning Models Performance

The performance of various models was evaluated on two distinct classification tasks: binary classification and multi-class classification (4 classes). The following analysis summarizes the results based on metrics including Balanced Accuracy, Precision, Recall, and F1-Score. In the tables, results highlighted in bold indicate the best-performing models, with Balanced Accuracy used as the primary decision criterion..

Performance Comparison between RGB and IR Models

A detailed comparison of the performance of different models tested for the binary classification task is provided in Table 8.

Model	Image	Bal. Accuracy	Precision	Recall	F1-Score
LSNet-B	RGB	0.767 ± 0.034	0.794 ± 0.029	0.781 ± 0.033	0.785 ± 0.031
LSNet-B CWDC		0.849 ± 0.020	0.789 ± 0.033	0.849 ± 0.020	0.805 ± 0.036
LSNet-B SKA		0.820 ± 0.037	0.839 ± 0.028	0.831 ± 0.025	0.833 ± 0.026
LSNet-S		0.801 ± 0.039	0.823 ± 0.033	0.809 ± 0.033	0.813 ± 0.032
LSNet-S CWDC		0.772 ± 0.025	0.798 ± 0.022	0.784 ± 0.025	0.788 ± 0.024
LSNet-S SKA		0.774 ± 0.023	0.798 ± 0.020	0.775 ± 0.021	0.781 ± 0.020
LSNet-T		0.781 ± 0.024	0.804 ± 0.020	0.784 ± 0.018	0.789 ± 0.018
LSNet-T CWDC		0.773 ± 0.042	0.798 ± 0.033	0.781 ± 0.022	0.785 ± 0.024
LSNet-T SKA		0.794 ± 0.027	0.817 ± 0.021	0.803 ± 0.021	0.807 ± 0.021
ResNet18		**0.970 ± 0.025**	**0.982 ± 0.014**	**0.981 ± 0.015**	**0.981 ± 0.016**
ResNet50		0.955 ± 0.010	0.973 ± 0.006	0.972 ± 0.006	0.972 ± 0.006
LSNet-B	IR	0.892 ± 0.023	0.896 ± 0.020	0.881 ± 0.023	0.884 ± 0.023
LSNet-B CWDC		0.877 ± 0.029	0.884 ± 0.025	0.872 ± 0.029	0.875 ± 0.028
LSNet-B SKA		0.886 ± 0.022	0.893 ± 0.021	0.884 ± 0.025	0.886 ± 0.024
LSNet-S		0.890 ± 0.022	0.894 ± 0.019	0.875 ± 0.022	0.878 ± 0.021
LSNet-S CWDC		0.891 ± 0.032	0.896 ± 0.027	0.884 ± 0.031	0.887 ± 0.030
LSNet-S SKA		0.874 ± 0.037	0.881 ± 0.031	0.853 ± 0.035	0.858 ± 0.034
LSNet-T		0.883 ± 0.017	0.888 ± 0.014	0.869 ± 0.016	0.872 ± 0.015
LSNet-T CWDC		0.891 ± 0.036	0.895 ± 0.031	0.872 ± 0.039	0.876 ± 0.037
LSNet-T SKA		0.870 ± 0.015	0.878 ± 0.013	0.863 ± 0.018	0.866 ± 0.017
ResNet18		0.990 ± 0.012	0.994 ± 0.008	0.994 ± 0.008	0.994 ± 0.008
ResNet50		**0.995 ± 0.010**	**0.997 ± 0.006**	**0.997 ± 0.006**	**0.997 ± 0.006**

Table 8: Performance comparison of CNN models (LSNet and ResNet variants) on binary classification using RGB and IR images.

The detailed results of the multi-class classification process are presented in Table 9.

Model	Image	Bal. Accuracy	Precision	Recall	F1-Score
LSNet-B	RGB	0.364 ± 0.038	0.383 ± 0.037	0.378 ± 0.030	0.376 ± 0.030
LSNet-B CWDC		0.312 ± 0.033	0.320 ± 0.034	0.337 ± 0.032	0.323 ± 0.033
LSNet-B SKA		0.369 ± 0.046	0.390 ± 0.042	0.388 ± 0.047	0.385 ± 0.044
LSNet-S		0.357 ± 0.060	0.369 ± 0.034	0.369 ± 0.060	0.365 ± 0.047
LSNet-S CWDC		0.324 ± 0.030	0.334 ± 0.036	0.341 ± 0.029	0.332 ± 0.031
LSNet-S SKA		0.396 ± 0.073	0.404 ± 0.071	0.403 ± 0.069	0.398 ± 0.066
LSNet-T		0.356 ± 0.028	0.345 ± 0.030	0.363 ± 0.027	0.351 ± 0.027
LSNet-T CWDC		0.395 ± 0.039	0.401 ± 0.030	0.397 ± 0.036	0.392 ± 0.032
LSNet-T SKA		0.381 ± 0.034	0.402 ± 0.035	0.388 ± 0.033	0.386 ± 0.030
ResNet18		**0.442 ± 0.024**	**0.499 ± 0.018**	**0.478 ± 0.022**	**0.478 ± 0.024**
ResNet50		0.431 ± 0.045	0.487 ± 0.047	0.459 ± 0.044	0.453 ± 0.045
LSNet-B	IR	0.428 ± 0.052	0.453 ± 0.065	0.453 ± 0.044	0.434 ± 0.049
LSNet-B CWDC		0.452 ± 0.009	0.459 ± 0.031	0.484 ± 0.014	0.464 ± 0.028
LSNet-B SKA		0.482 ± 0.037	0.492 ± 0.033	0.513 ± 0.029	0.492 ± 0.034
LSNet-S		0.457 ± 0.030	0.478 ± 0.033	0.484 ± 0.026	0.474 ± 0.027
LSNet-S CWDC		0.434 ± 0.051	0.454 ± 0.039	0.463 ± 0.047	0.446 ± 0.041
LSNet-S SKA		0.396 ± 0.073	0.404 ± 0.071	0.403 ± 0.069	0.398 ± 0.066
LSNet-T		0.475 ± 0.035	0.500 ± 0.031	0.494 ± 0.035	0.486 ± 0.031
LSNet-T CWDC		0.463 ± 0.044	0.471 ± 0.044	0.484 ± 0.040	0.468 ± 0.039
LSNet-T SKA		0.450 ± 0.038	0.485 ± 0.032	0.475 ± 0.036	0.473 ± 0.032
ResNet18		0.476 ± 0.043	0.527 ± 0.044	0.522 ± 0.038	0.509 ± 0.039
ResNet50		**0.498 ± 0.023**	**0.546 ± 0.028**	**0.538 ± 0.023**	**0.513 ± 0.026**

Table 9: Performance comparison of CNN models (LSNet and ResNet variants) on multi classification using RGB and IR images.

Model	RGB Bal. Accuracy	IR Bal. Accuracy	Δ Bal. Accuracy	Δ Std
LSNet-B	0.767 ± 0.034	0.892 ± 0.023	+12.5%	-1.1%
LSNet-B CWDC	0.849 ± 0.020	0.877 ± 0.029	+2.8%	+0.9%
LSNet-B SKA	0.820 ± 0.037	0.886 ± 0.022	+6.6%	-1.5%
LSNet-S	0.801 ± 0.039	0.890 ± 0.022	+8.9%	-1.7%
LSNet-S CWDC	0.772 ± 0.025	0.891 ± 0.032	+11.9%	+0.7%
LSNet-S SKA	0.774 ± 0.023	0.874 ± 0.037	+10.0%	+1.4%
LSNet-T	0.781 ± 0.024	0.883 ± 0.017	+10.2%	-0.7%
LSNet-T CWDC	0.773 ± 0.042	0.891 ± 0.036	+11.8%	-0.6%
LSNet-T SKA	0.794 ± 0.027	0.870 ± 0.015	+7.6%	-1.2%
ResNet18	0.970 ± 0.025	0.990 ± 0.012	+2.0%	-1.3%
ResNet50	0.955 ± 0.010	0.995 ± 0.010	+4.0%	+0.0%

Table 10: Balanced Accuracy comparison between RGB and IR modalities for LSNet and ResNet models, including performance deltas.

Model	RGB Bal. Accuracy	IR Bal. Accuracy	Δ Bal. Accuracy	Δ Std
LSNet-B	0.364 ± 0.038	0.428 ± 0.052	+6.4%	+1.4%
LSNet-B CWDC	0.312 ± 0.033	0.452 ± 0.009	+14%	-2.4%
LSNet-B SKA	0.369 ± 0.046	0.482 ± 0.037	+11.3%	-0.9%
LSNet-S	0.357 ± 0.060	0.457 ± 0.030	+10.0%	-3%
LSNet-S CWDC	0.324 ± 0.030	0.434 ± 0.051	+11%	+2.1%
LSNet-S SKA	0.396 ± 0.073	0.444 ± 0.049	+4.8%	-2.4%
LSNet-T	0.356 ± 0.028	0.475 ± 0.035	+11.9%	+0.7%
LSNet-T CWDC	0.381 ± 0.034	0.463 ± 0.044	+8.2%	+1%
LSNet-T SKA	0.395 ± 0.039	0.450 ± 0.038	+5.5%	-0.1%
ResNet18	0.442 ± 0.024	0.476 ± 0.043	+3.4%	+1.9%
ResNet50	0.431 ± 0.045	0.498 ± 0.023	+6.7%	-2.2%

Table 11: Balanced Accuracy results in a more challenging classification scenario, highlighting RGB vs IR performance gaps for LSNet and ResNet models.

Tables 8 and 9 report the detailed results of binary and multi-class classification.

Balanced accuracy values ranged from 0.767 to 0.995 in binary tasks and from 0.312 to 0.498 in multi-class tasks, depending on model and modality.
Tables 10 and 11 provide a direct comparison between RGB and IR images. For binary classification (Table 10), performance deltas between IR and RGB ranged from +2.0% to +12.5%, with standard deviation variations between –1.7% and +1.4%. For multi-class classification (Table 11), deltas ranged from +3.4% to +14.0%, with standard deviation changes between –3.0% and +2.1%.

Image Fusion

The image fusion approach was evaluated using two different fusion architectures: Late Fusion ResNet (LF-ResNet) and Dual Encoder ResNet (DE-ResNet). These models were designed to leverage complementary information from both RGB and IR imaging modalities simultaneously. The binary and multi-class fusion results are presented in Tables 12 and 13

Model	Image	Bal. Accuracy	Precision	Recall	F1-Score
LF-ResNet	RGB-IR	0.934 ± 0.015	0.941 ± 0.008	0.935 ± 0.013	0.935 ± 0.012
DE-ResNet		**0.954 ± 0.028**	**0.955 ± 0.025**	**0.952 ± 0.027**	**0.953 ± 0.027**

Table 12: Model Performance Comparison for RGB-IR Image Type (Rounded to Three Decimal Places) in binary classification.

Model	Image	Bal. Accuracy	Precision	Recall	F1-Score
LF-ResNet	RGB-IR	0.342 ± 0.025	0.146 ± 0.111	0.310 ± 0.092	0.174 ± 0.093
DE-ResNet		**0.378 ± 0.053**	**0.352 ± 0.059**	**0.465 ± 0.072**	**0.376 ± 0.079**

Table 13: Model Performance Comparison for RGB-IR Image Type (Rounded to Three Decimal Places) in multi classification.

In binary classification (see Table 12), LF-ResNet achieved a balanced accuracy of 0.934 ± 0.015, whereas DE-ResNet achieved 0.954 ± 0.028. Precision ranged from 0.941 to 0.955, recall from 0.935 to 0.952 and F1 scores from 0.935 to 0.953. In multi-class classification (Table 13), LF-ResNet obtained a balanced accuracy of 0.342 ± 0.025, while DE-ResNet achieved 0.378 ± 0.053. The precision values were 0.146 and 0.352, the recall ranged from 0.310 to 0.465, and the F1-scores ranged from 0.174 to 0.376.

Pre-Post Cold Test

The results of the Pre-Post Cold Test strategy are presented in Tables 14 and 15, showing the performance evaluation of ResNet18 and ResNet50 models on temporal analysis of T0-T10 images using both RGB and infrared (IR) imaging modalities.

Model	Image	Bal. Accuracy	Precision	Recall	F1-Score
ResNet18	RGB T0-T10	**0.888 ± 0.041**	**0.911 ± 0.027**	**0.903 ± 0.026**	**0.905 ± 0.026**
ResNet50		0.879 ± 0.075	0.900 ± 0.051	0.890 ± 0.051	0.892 ± 0.050
ResNet18	IR T0-T10	**0.968 ± 0.020**	**0.969 ± 0.001**	**0.966 ± 0.005**	**0.966 ± 0.001**
ResNet50		0.910 ± 0.074	0.942 ± 0.050	0.938 ± 0.055	0.938 ± 0.054

Table 14: Model Performance Comparison for T0 - T10 Image Type (Rounded to Three Decimal Places) in binary classification.

Model	Image	Bal. Accuracy	Precision	Recall	F1-Score
ResNet18	RGB T0-T10	0.361 ± 0.034	0.331 ± 0.041	0.343 ± 0.055	0.343 ± 0.060
ResNet50		**0.418 ± 0.047**	**0.417 ± 0.048**	**0.421 ± 0.051**	**0.401 ± 0.052**
ResNet18	IR T0-T10	0.363 ± 0.048	0.354 ± 0.047	0.359 ± 0.052	0.345 ± 0.051
ResNet50		**0.412 ± 0.062**	**0.419 ± 0.075**	**0.414 ± 0.062**	**0.397 ± 0.053**

Table 15: Model Performance Comparison for T0 - T10 Image Type (Rounded to Three Decimal Places) in multi classification.

In binary classification (see Table 14), the balanced accuracy values ranged from 0.879 ± 0.075 to 0.968 ± 0.020. The corresponding precision values ranged from 0.900 to 0.969, the recall values from 0.890 to 0.966, and the F1-score values from 0.892 to 0.966. In multi-class classification (Table 15), the values of balanced accuracy ranged from 0.361 ± 0.034 to 0.418 ± 0.047 for RGB and from 0.363 ± 0.048 to 0.412 ± 0.062 for IR. Precision varied between 0.331 and 0.419, recall between 0.343 and 0.421, and F1-scores between 0.343 and 0.401.

Morphological Analysis (Shape-Based)

The results of the classification task based on foot shape are reported in Tables 16 and 17. These tables summarize the performance of various models, including LSNet variants (LSNet-B, LSNet-S, LSNet-T) with different configurations (CWDC, SKA), as well as ResNet architectures (ResNet18 and ResNet50), evaluated on the SHAPE image modality. Table 16 presents the binary classification outcomes, while Table 17 reports the corresponding results for the multi-class setting. The use of foot shape for classification follows the approach introduced by Li et al. [39].

Model	Image	Bal. Accuracy	Precision	Recall	F1-Score
LSNet-B		0.846 ± 0.020	0.857 ± 0.017	0.844 ± 0.017	0.847 ± 0.017
LSNet-B CWDC		0.835 ± 0.020	0.848 ± 0.016	0.825 ± 0.021	0.830 ± 0.020
LSNet-B SKA		0.854 ± 0.037	0.864 ± 0.031	0.844 ± 0.038	0.848 ± 0.037
LSNet-S		0.855 ± 0.020	0.866 ± 0.018	0.853 ± 0.023	0.856 ± 0.022
LSNet-S CWDC		0.854 ± 0.023	0.863 ± 0.019	0.844 ± 0.017	0.848 ± 0.017
LSNet-S SKA	SHAPE	0.837 ± 0.023	0.850 ± 0.020	0.828 ± 0.022	0.833 ± 0.021
LSNet-T		0.835 ± 0.016	0.850 ± 0.014	0.822 ± 0.016	0.827 ± 0.015
LSNet-T CWDC		0.841 ± 0.028	0.856 ± 0.022	0.834 ± 0.023	0.838 ± 0.023
LSNet-T SKA		0.824 ± 0.032	0.841 ± 0.028	0.822 ± 0.036	0.826 ± 0.034
ResNet18		0.913 ± 0.016	0.947 ± 0.013	0.944 ± 0.013	0.942 ± 0.013
ResNet50		**0.934 ± 0.014**	**0.951 ± 0.018**	**0.947 ± 0.021**	**0.947 ± 0.020**

Table 16: Model Performance Comparison for SHAPE Image Type (Rounded to Three Decimal Places) in binary classification.

Model	Image	Bal. Accuracy	Precision	Recall	F1-Score
LSNet-B		0.423 ± 0.048	0.419 ± 0.061	0.447 ± 0.045	0.424 ± 0.051
LSNet-B CWDC		0.444 ± 0.036	0.465 ± 0.028	0.466 ± 0.030	0.452 ± 0.019
LSNet-B SKA		0.401 ± 0.042	0.400 ± 0.062	0.425 ± 0.040	0.399 ± 0.041
LSNet-S		**0.497 ± 0.032**	**0.509 ± 0.044**	**0.513 ± 0.025**	**0.494 ± 0.027**
LSNet-S CWDC		0.432 ± 0.025	0.442 ± 0.029	0.447 ± 0.031	0.426 ± 0.027
LSNet-S SKA	SHAPE	0.433 ± 0.036	0.440 ± 0.024	0.456 ± 0.030	0.436 ± 0.028
LSNet-T		0.454 ± 0.043	0.472 ± 0.056	0.472 ± 0.042	0.453 ± 0.041
LSNet-T CWDC		0.457 ± 0.050	0.481 ± 0.059	0.478 ± 0.051	0.462 ± 0.048
LSNet-T SKA		0.476 ± 0.027	0.482 ± 0.028	0.494 ± 0.021	0.471 ± 0.016
ResNet18		0.424 ± 0.059	0.482 ± 0.067	0.444 ± 0.056	0.445 ± 0.058
ResNet50		0.415 ± 0.035	0.477 ± 0.043	0.453 ± 0.026	0.446 ± 0.035

Table 17: Model Performance Comparison for SHAPE Image Type (Rounded to Three Decimal Places) in multi classification.

In binary classification (Table 16), balanced accuracy ranged from 0.824 ± 0.032 to 0.934 ± 0.014. Precision values were between 0.841 and 0.951, recall between 0.822 and 0.947, and F1-scores from 0.826 to 0.947.
In multi-class classification (Table 17), balanced accuracy ranged from 0.401 ± 0.042 to 0.497 ± 0.032. Precision varied between 0.400 and 0.509, recall from 0.425 to 0.513, and F1-scores from 0.399 to 0.494.

3.2 Results of the Statistical Analysis

Statistical analysis constitutes a fundamental approach to understanding thermal patterns and the quantitative characterization of differences between left and right feet. The present subchapter sets forth the results obtained through the extraction of statistical features from IR images and the application of traditional machine learning algorithms for classification.

Analysis of Thermal Distributions

The analysis of thermal distributions is presented in Figures 15 and 16, showing box plots of temperature differences $|left - right|$ across different anatomical categories and regions at T0 and T10 time points respectively. The figures display the distribution of thermal asymmetry values with a reference threshold of 2.2°C indicated by the dashed red line.

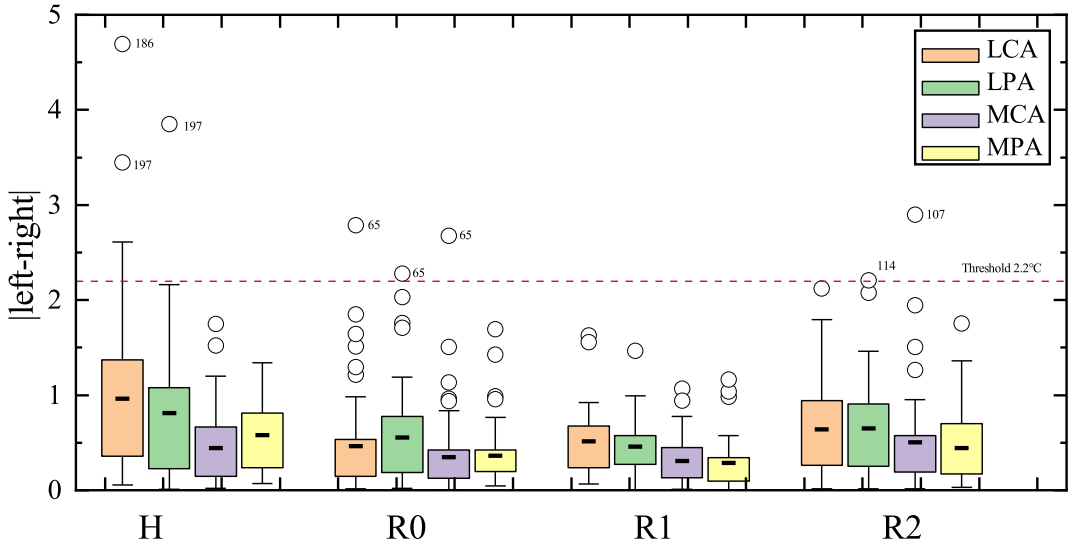


Figure 15: Box plots of absolute temperature differences ($|left - right|$) across anatomical regions (LCA, LPA, MCA, MPA) at baseline (T0). The dashed red line indicates the clinical threshold of 2.2°C.

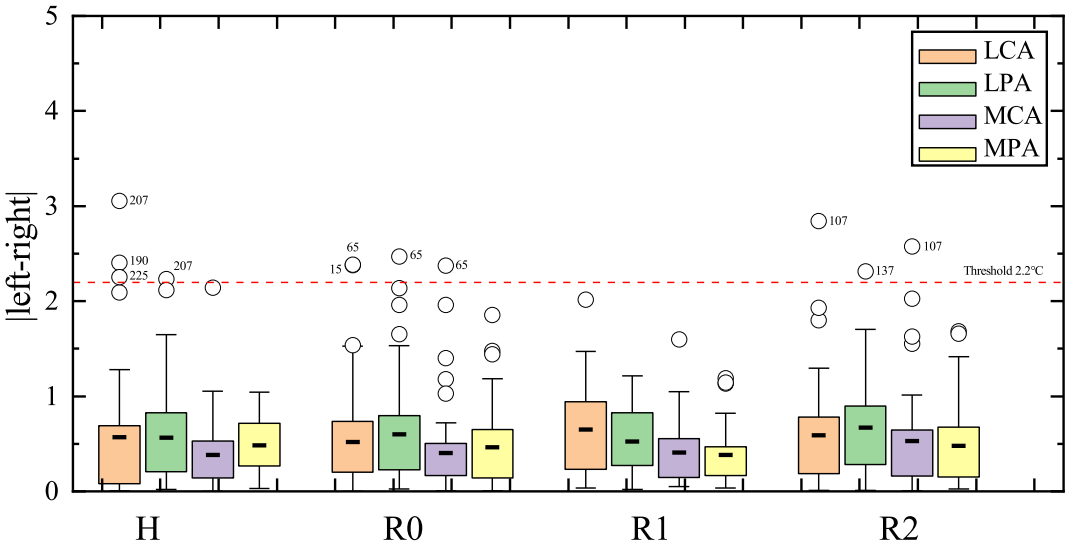


Figure 16: Box plots of absolute temperature differences ($|left - right|$) across anatomical regions (LCA, LPA, MCA, MPA) after 10 minutes (T10). The dashed red line indicates the clinical threshold of 2.2°C.

Overall, Figures 15 and 16 show that most temperature differences fall below the 2.2°C threshold. However, there are notable outliers, particularly in R0, R2 and among some healthy subjects, which exceed it.

Identification of Critical Cases

Tables 18 and 19 present the quantitative analysis of thermal asymmetries using the z-score methodology. The interpretation of z-scores was based on empirically defined thresholds, derived from the observed distribution of values in the dataset:

- 0–0.2: no difference;
- 0.2–0.5: minimal effect;
- 0.5–0.8: moderate effect;
- 0.8–1.8: significant effect;
- 1.8 or higher: very significant effect.

According to this framework, the tables show patients with very significant effect classification, displaying their category, segment, mean left-right temperature difference, z-score values, and corresponding interpretation at T0 and T10 time points respectively.

Patient	Category	Segment	LR Mean T0 (°C)	Z-score T0	Interpretation
65	R0	MCA	2.679	5.650	Very significant effect
65	R0	LCA	2.786	5.474	Very significant effect
65	R0	LPA	2.280	3.181	Very significant effect
107	R2	MCA	2.900	2.639	Very significant effect

Table 18: Mean Left-Right Temperature Difference and Z-score at T0 for Segments with Very Significant Effect

Patient	Category	Segment	LR Mean T10 (°C)	Z-score T10	Interpretation
65	R0	LCA	2.383	4.623	Very significant effect
65	R0	LPA	2.468	4.126	Very significant effect
65	R0	MCA	2.374	3.263	Very significant effect
137	R2	LPA	2.314	3.126	Very significant effect
15	R0	LCA	2.378	2.919	Very significant effect
107	R2	MCA	2.577	2.898	Very significant effect
107	R2	LCA	2.842	2.181	Very significant effect

Table 19: Mean Left-Right Temperature Difference and Z-score at T10 for Segments with Very Significant Effect

Overall, Tables 18 and 19 show that all "very significant effect" patients exceeded a z-score of 2. Patient 65 (R0) presented asymmetries across all regions (up to z = 5.65). Patients 65 and 107 were present at both T0 and T10, while new cases appeared at T10 (patients 137 and 15).

Comparison of Machine Learning Approaches

The classification results with traditional machine learning algorithms are presented in Table 20, showing the performance comparison between different algorithms (TabPFN, Random Forest, SVM) applied to different data types (Features, RGB, IR) for both binary (2 classes) and multi-class (4 classes) classification tasks.

Model	Data	Classes	Balanced Accuracy
TabPFN	Features	2	**0.984 ± 0.051**
Random Forest			0.919 ± 0.025
SVM			0.920 ± 0.066
Random Forest	RGB		0.823 ± 0.064
SVM			0.822 ± 0.053
Random Forest	IR		0.877 ± 0.047
SVM			0.842 ± 0.041
TabPFN	Features	4	**0.536 ± 0.050**
Random Forest			0.442 ± 0.034
SVM			0.465 ± 0.067
Random Forest	RGB		0.409 ± 0.031
SVM			0.446 ± 0.042
Random Forest	IR		0.446 ± 0.056
SVM			0.467 ± 0.018

Table 20: Comparison of classification performance between different machine learning algorithms and data types.

In binary classification, TabPFN achieved a balanced accuracy of 0.984 ± 0.051, Random Forest 0.919 ± 0.025, and SVM 0.920 ± 0.066. Using images, Random Forest reached 0.823 ± 0.064 with RGB and 0.877 ± 0.047 with IR, while SVM obtained 0.822 ± 0.053 with RGB and 0.842 ± 0.041 with IR.
In multi-class classification, TabPFN achieved 0.536 ± 0.050, Random Forest 0.442 ± 0.034 with features, 0.409 ± 0.031 with RGB, and 0.446 ± 0.056 with IR, while SVM reached 0.465 ± 0.067 with features, 0.446 ± 0.042 with RGB, and 0.467 ± 0.018 with IR.

3.3 Hybrid Approach: Deep Learning and Statistical Features

The integration of deep learning approaches and statistical features represents a promising strategy to combine the automatic feature extraction capability of neural networks with the interpretability and robustness of statistical features.

Model	Image	Bal. Accuracy	Precision	Recall	F1-Score
LSNet-B		0.847 ± 0.025	0.781 ± 0.030	0.847 ± 0.025	0.808 ± 0.029
LSNet-B CWDC		0.925 ± 0.022	0.947 ± 0.016	0.943 ± 0.020	0.943 ± 0.019
LSNet-B SKA		0.833 ± 0.030	0.764 ± 0.023	0.833 ± 0.030	0.780 ± 0.026
LSNet-S		0.856 ± 0.033	0.786 ± 0.042	0.856 ± 0.033	0.800 ± 0.039
LSNet-S CWDC	RGB	0.838 ± 0.035	0.770 ± 0.028	0.838 ± 0.035	0.786 ± 0.029
LSNet-S SKA	+	0.864 ± 0.057	0.797 ± 0.060	0.864 ± 0.057	0.815 ± 0.065
LSNet-T	Features	0.862 ± 0.020	0.793 ± 0.021	0.862 ± 0.020	0.823 ± 0.023
LSNet-T CWDC		0.852 ± 0.038	0.786 ± 0.047	0.852 ± 0.038	0.798 ± 0.052
LSNet-T SKA		0.869 ± 0.022	0.788 ± 0.017	0.869 ± 0.022	0.803 ± 0.020
ResNet18		**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**
ResNet50		**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**
LSNet-B		0.890 ± 0.037	0.813 ± 0.041	0.890 ± 0.037	0.833 ± 0.039
LSNet-B CWDC		0.904 ± 0.030	0.828 ± 0.043	0.904 ± 0.030	0.846 ± 0.048
LSNet-B SKA		0.893 ± 0.025	0.805 ± 0.027	0.893 ± 0.025	0.821 ± 0.035
LSNet-S		0.907 ± 0.018	0.828 ± 0.021	0.907 ± 0.018	0.848 ± 0.024
LSNet-S CWDC	IR	0.913 ± 0.006	0.830 ± 0.011	0.913 ± 0.006	0.852 ± 0.012
LSNet-S SKA	+	0.911 ± 0.041	0.833 ± 0.052	0.911 ± 0.041	0.851 ± 0.059
LSNet-T	Features	0.916 ± 0.014	0.842 ± 0.020	0.916 ± 0.014	0.867 ± 0.017
LSNet-T CWDC		0.908 ± 0.030	0.835 ± 0.024	0.908 ± 0.030	0.858 ± 0.027
LSNet-T SKA		0.896 ± 0.009	0.802 ± 0.010	0.896 ± 0.009	0.818 ± 0.013
ResNet18		**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**
ResNet50		0.993 ± 0.006	0.980 ± 0.016	0.993 ± 0.006	0.986 ± 0.011

Table 21: Performance of models with features in binary classification (values rounded to three decimal places).

This section presents results obtained through the implementation of late fusion architectures that combine representations learned from images with previously extracted statistical features. The implemented late fusion approach involves the separate training of deep learning models on images and classifiers on statistical features, followed by combining the predictions through ensemble techniques. Tables 21 and 22 report the results for binary and multi-class classification, respectively.

Model	Image	Bal. Accuracy	Precision	Recall	F1-Score
LSNet-B		0.421 ± 0.053	0.397 ± 0.059	0.421 ± 0.053	0.408 ± 0.057
LSNet-B CWDC		0.408 ± 0.033	0.383 ± 0.043	0.408 ± 0.033	0.388 ± 0.040
LSNet-B SKA		0.416 ± 0.059	0.393 ± 0.056	0.416 ± 0.059	0.391 ± 0.046
LSNet-S		0.437 ± 0.042	0.411 ± 0.047	0.437 ± 0.042	0.423 ± 0.045
LSNet-S CWDC	RGB	0.433 ± 0.042	0.408 ± 0.048	0.433 ± 0.042	0.411 ± 0.043
LSNet-S SKA	+	0.446 ± 0.059	0.442 ± 0.068	0.446 ± 0.059	0.436 ± 0.065
LSNet-T	Features	0.451 ± 0.030	0.430 ± 0.036	0.451 ± 0.030	0.439 ± 0.034
LSNet-T CWDC		0.453 ± 0.072	0.428 ± 0.095	0.453 ± 0.072	0.430 ± 0.086
LSNet-T SKA		0.423 ± 0.058	0.394 ± 0.067	0.423 ± 0.058	0.400 ± 0.061
ResNet18		0.505 ± 0.045	0.516 ± 0.075	0.505 ± 0.045	0.490 ± 0.058
ResNet50		**0.506 ± 0.034**	**0.518 ± 0.036**	**0.506 ± 0.034**	**0.490 ± 0.029**
LSNet-B		0.462 ± 0.031	0.436 ± 0.028	0.462 ± 0.031	0.444 ± 0.034
LSNet-B CWDC		0.426 ± 0.030	0.398 ± 0.031	0.426 ± 0.030	0.403 ± 0.034
LSNet-B SKA		0.479 ± 0.027	0.455 ± 0.040	0.479 ± 0.027	0.451 ± 0.029
LSNet-S		0.496 ± 0.023	0.478 ± 0.025	0.496 ± 0.023	0.482 ± 0.022
LSNet-S CWDC	IR	0.478 ± 0.025	0.450 ± 0.022	0.478 ± 0.025	0.448 ± 0.017
LSNet-S SKA	+	0.472 ± 0.052	0.442 ± 0.075	0.472 ± 0.052	0.443 ± 0.062
LSNet-T	Features	0.481 ± 0.028	0.460 ± 0.033	0.481 ± 0.028	0.463 ± 0.029
LSNet-T CWDC		0.489 ± 0.032	0.466 ± 0.037	0.489 ± 0.032	0.471 ± 0.034
LSNet-T SKA		0.484 ± 0.041	0.460 ± 0.051	0.484 ± 0.041	0.456 ± 0.050
ResNet18		**0.517 ± 0.040**	**0.533 ± 0.049**	**0.517 ± 0.040**	**0.509 ± 0.043**
ResNet50		0.494 ± 0.014	0.504 ± 0.029	0.494 ± 0.014	0.480 ± 0.022

Table 22: Performance of models with features in multi-class classification (values rounded to three decimal places).

For binary classification, balanced accuracy values ranged from 0.833 to 1.000 with RGB + features and from 0.890 to 1.000 with IR + features. Precision, recall, and F1-scores followed the same trend, with values between 0.764 and 1.000.
For multi-class classification, balanced accuracy values ranged from 0.408 to 0.506 with RGB + features and from 0.426 to 0.517 with IR + features. Precision, recall, and F1-scores showed consistent ranges between 0.383 and 0.518.

Comparative Analysis

The comparative analysis of performance differences between the use of images alone and the hybrid images+features approach is presented in Tables 23 and 24 for binary and multi-class classification, respectively. The tables show the performance metrics for each model configuration using images only versus images combined with statistical features, along with the calculated delta improvements in balanced accuracy and standard deviation changes.

Model	Data	Image	Image + Features	Delta Bal.Acc.	Delta STD
LSNet-B	RGB	0.767 ± 0.034	0.847 ± 0.025	+8.0%	-0.9%
LSNet-B CWDC	RGB	0.849 ± 0.020	0.925 ± 0.022	+7.6%	-0.2%
LSNet-B SKA	RGB	0.820 ± 0.037	0.833 ± 0.030	+1.3%	-0.7%
LSNet-S	RGB	0.801 ± 0.039	0.856 ± 0.033	+5.5%	-0.6%
LSNet-S CWDC	RGB	0.772 ± 0.025	0.838 ± 0.035	+6.6%	+1.0%
LSNet-S SKA	RGB	0.774 ± 0.023	0.864 ± 0.057	+9.0%	+3.4%
LSNet-T	RGB	0.781 ± 0.024	0.862 ± 0.020	+8.1%	-0.4%
LSNet-T CWDC	RGB	0.773 ± 0.042	0.852 ± 0.038	+7.9%	-0.4%
LSNet-T SKA	RGB	0.794 ± 0.027	0.869 ± 0.022	+7.5%	-0.5%
ResNet18	RGB	0.970 ± 0.025	1.000 ± 0.000	+3.0%	-2.5%
ResNet50	RGB	0.955 ± 0.010	1.000 ± 0.000	+4.5%	-1.0%
LSNet-B	IR	0.892 ± 0.023	0.890 ± 0.037	-0.2%	+1.4%
LSNet-B CWDC	IR	0.877 ± 0.029	0.904 ± 0.030	+2.7%	+0.1%
LSNet-B SKA	IR	0.886 ± 0.022	0.893 ± 0.025	+0.7%	+0.3%
LSNet-S	IR	0.890 ± 0.022	0.907 ± 0.018	+1.7%	-0.4%
LSNet-S CWDC	IR	0.891 ± 0.032	0.913 ± 0.006	+2.2%	-2.6%
LSNet-S SKA	IR	0.874 ± 0.037	0.911 ± 0.041	+3.7%	+0.4%
LSNet-T	IR	0.883 ± 0.017	0.916 ± 0.014	+3.3%	-0.3%
LSNet-T CWDC	IR	0.891 ± 0.036	0.908 ± 0.030	+1.7%	-0.6%
LSNet-T SKA	IR	0.870 ± 0.015	0.896 ± 0.009	+2.6%	-0.6%
ResNet18	IR	0.990 ± 0.012	1.000 ± 0.000	+1.0%	-1.2%
ResNet50	IR	0.995 ± 0.010	0.993 ± 0.006	-0.2%	-0.4%

Table 23: Comparative analysis of performance variations in the image + features approach for binary classification.

Model	Data	Image	Image + Features	Delta Bal.Acc.	Delta STD
LSNet-B	RGB	0.364 ± 0.038	0.421 ± 0.053	+5.7%	+1.5%
LSNet-B CWDC	RGB	0.312 ± 0.033	0.408 ± 0.033	+9.6%	-0.02%
LSNet-B SKA	RGB	0.369 ± 0.046	0.416 ± 0.059	+4.7%	+1.3%
LSNet-S	RGB	0.357 ± 0.060	0.437 ± 0.042	+8.0%	-1.8%
LSNet-S CWDC	RGB	0.324 ± 0.030	0.433 ± 0.042	+10.9%	+1.2%
LSNet-S SKA	RGB	0.396 ± 0.073	0.446 ± 0.059	+5.0%	-1.4%
LSNet-T	RGB	0.356 ± 0.028	0.451 ± 0.030	+9.5%	+0.2%
LSNet-T SKA	RGB	0.381 ± 0.034	0.423 ± 0.058	+4.4%	+2.4%
LSNet-T CWDC	RGB	0.395 ± 0.039	0.453 ± 0.072	+5.8%	+3.3%
ResNet18	RGB	0.442 ± 0.024	0.505 ± 0.045	+6.3%	+2.1%
ResNet50	RGB	0.431 ± 0.045	0.506 ± 0.034	+7.5%	-1.1%
LSNet-B	IR	0.428 ± 0.052	0.462 ± 0.031	+3.4%	-2.1%
LSNet-B CWDC	IR	0.452 ± 0.009	0.426 ± 0.030	-2.6%	+2.1%
LSNet-B SKA	IR	0.482 ± 0.037	0.479 ± 0.027	-0.3%	-1.0%
LSNet-S	IR	0.457 ± 0.030	0.496 ± 0.023	+3.9%	-0.7%
LSNet-S CWDC	IR	0.434 ± 0.051	0.478 ± 0.025	+4.4%	-2.6%
LSNet-S SKA	IR	0.444 ± 0.049	0.472 ± 0.052	+2.8%	+0.3%
LSNet-T	IR	0.475 ± 0.035	0.481 ± 0.028	+0.6%	-0.7%
LSNet-T CWDC	IR	0.463 ± 0.044	0.489 ± 0.032	+2.6%	-1.2%
LSNet-T SKA	IR	0.450 ± 0.038	0.484 ± 0.041	+3.4%	+0.3%
ResNet18	IR	0.476 ± 0.043	0.517 ± 0.040	+4.1%	-0.3%
ResNet50	IR	0.498 ± 0.023	0.494 ± 0.014	-0.4%	-0.9%

Table 24: Comparative analysis of performance variations in the image + features approach for multi-class classification.

The comparative analysis shows performance differences between image-only and hybrid image+features approaches.
For binary classification (Table 23), balanced accuracy deltas ranged from –0.2% to +9.0% with RGB and from –0.2% to +3.7% with IR. Standard deviation changes varied between –2.6% and +3.4%.
For multi-class classification (Table 24), balanced accuracy deltas ranged from –2.6% to +10.9% with RGB and from –0.4% to +4.4% with IR. Standard deviation changes

varied between –2.6% and +3.3%.

4 Discussion and Conclusion

In this section, we provide a critical interpretation of the experimental results presented in the previous chapter. The aim is to highlight the main findings, compare model behaviors across modalities and tasks, and discuss their implications in the context of diabetic foot assessment.
The segmentation model achieved pixel-wise accuracy and an Intersection over Union (IoU) of around 0.980, indicating a near-perfect overlap between the automatic masks produced and the manual reference masks.

Performance Comparison between RGB and IR Models

The results reported in Tables 8 and 9 show marked differences between the use of RGB and IR images in binary and multi-class classification tasks.
In the context of binary classification, the LSNet-B CWDC model performed well with RGB images, achieving a balanced accuracy of 0.840 ± 0.020. However, the other LSNet variants performed significantly poorer on RGB, while doing better on IR images, with accuracies between 0.870 and 0.890. This suggests that IR thermal information is better for some models in binary distinction.
ResNet18, on the other hand, dominated in every context, achieving a balanced accuracy of 0.970 ± 0.025 with RGB, while ResNet50 showed outstanding performance with IR, reaching 0.995 ± 0.010. Overall, IR seems to provide an advantage in binary classification, but the absolute best performance is observed with ResNet models, regardless of mode.
In the case of multi-class classification, ResNet18 maintained the lead with RGB images, while ResNet50 prevailed with IR, confirming the higher robustness of ResNet models even in complex scenarios. However, performance drops across the board for all models, indicating the inherent difficulties of the multi-class problem.

Image Fusion

As shown in Tables 12 and 13, the fusion of RGB and IR information shows promising results, especially in binary classification.
The DE-ResNet model stood out with a balanced accuracy of 0.954 ± 0.028, outperforming LF-ResNet (0.934 ± 0.015). All indicators (accuracy, recall, F1) were above 0.950, showing how well-managed fusion can make the best use of both types of information, however, the models fail to outperform the single-input models; this indicates that both images probably introduce noise that does not allow the models to perform as well as the single-input models.
In the multi-class classification, however, both models declined: DE-ResNet stopped at 0.378 ± 0.053, LF-ResNet at 0.342 ± 0.025. However, DE-ResNet maintained an advantage in terms of precision and recall, indicating a better integration capability between the two modes. These data confirm that although fusion is useful, the complexity of the multi-class task remains a significant obstacle.

Pre-Post Cold Test

Tables 14 and 15 introduce a further level of analysis: pre-post temporal variation.
In binary classification, IR images showed a clear advantage over RGB. The ResNet18 model performed excellently with IR images acquired between T0 and T10, achieving a balanced accuracy of 0.968 ± 0.020 and metrics above 0.960 on all fronts.

This suggests that thermal variations over time offer particularly relevant signals for pattern recognition.
Again, ResNet18 outperformed ResNet50, probably due to a higher generalisation capacity and lower risk of overfitting. However, in the multi-class classification, the trend is reversed: the accuracies fall between 0.360 and 0.420, and ResNet50 shows a slight superiority due to its greater representational capacity, in which case there is no difference between IR and RGB images.

Morphological Analysis (Shape-Based)

Tables 16 and 17 analyse the performance of the models in using morphological information, in particular foot shape, as a basis for classification.
In the context of binary classification, ResNet models are clearly superior to LSNet variants: ResNet50 achieves a balanced accuracy of 0.934 ± 0.014, with metrics above 0.940 on precision, recall and F1-score. Among the LSNet models, LSNet-S is the most effective with an accuracy of 0.855.
These results show that morphological features alone – in the absence of thermal or colour information – are sufficient to distinguish between two classes with good accuracy. In other words, foot shape is an informative and relevant predictor for binary classification.
The situation changes significantly in multi-class classification, where a reversal of the performance hierarchy is observed. In this more complex scenario, LSNet-S emerges as the best model, with a balanced accuracy of 0.497 ± 0.032, outperforming both ResNet18 (0.424) and ResNet50 (0.415). This result suggests that LSNet architectures, designed to enhance morphological features, are particularly well suited to capturing subtle variations in shape that are relevant for distinguishing between multiple classes.

Analysis of Thermal Distributions

The boxplots shown in Figures 15 and 16 show the trend of temperature differences $|left - right|$ between the different anatomical categories and body regions. At both acquisition times (T0 and T10), considerable homogeneity is observed in the medians, with most values below the clinical threshold of 2.2°C (indicated by the dashed red line).
However, significant outliers emerge, particularly in the R0, R2 and H categories, exceeding the critical threshold and suggesting marked thermal asymmetries, potentially indicative of physiological changes or pathological conditions.
It is relevant to note that even some clinically healthy patients present values above the threshold, questioning the reliability of a fixed diagnostic criterion such as 2.2°C. This observation motivates the introduction of a normalized statistical approach, such as the z-score, for a more precise and personalized assessment of thermal abnormalities.

Identification of Critical Cases

Tables 18 and 19 quantitatively confirm what was observed in the boxplots: the z-score approach allows a more refined identification of clinically relevant cases.
All patients classified as "very significant effect" have z-scores ¿ 2, indicating the presence of true outliers. The case of patient 65 (category R0) is emblematic: it

shows persistent thermal asymmetries in all the regions analzed, both at T0 and T10, with values up to z = 5.65, signaling a potentially pathological systemic condition. The temporal persistence of the abnormalities in patients 65 and 107 also reinforces the hypothesis of stable functional alterations over time. On the contrary, the emergence of new significant cases at T10 (e.g. patients 137 and 15) suggests that the Cold Test may reveal latent problems, not visible under normal thermal conditions. Finally, the total absence of healthy patients in cases reported with a high z-score shows that this method is more selective than the simple threshold criterion, reducing false positives and improving diagnostic accuracy.

Comparison of Machine Learning Approaches

The results shown in Table 20 highlight the effectiveness of statistical features derived from thermal measurements for automatic classification.
In binary classification, traditional algorithms such as Random Forest and SVM achieved good performance, with balanced accuracies of 0.919 ± 0.025 and 0.920 ± 0.066, respectively. The TabPFN model, however, clearly stood out with a balanced accuracy of 0.984 ± 0.051, outperforming all other approaches.
A direct comparison between images and statistical features confirms the advantage of the latter: IR-based models reached 0.919 compared to 0.877 for images, while RGB-based models reached 0.919 compared to 0.823. This indicates that statistical representations capture more discriminative information than raw pixel data.
In multi-class classification, performance drops substantially. Random Forest and SVM remained between 0.440 and 0.470, while TabPFN reached 0.536 ± 0.050, showing only a modest improvement without fully addressing the complexity of the task.

Hybrid Approach: Deep Learning and Statistical Features

To overcome the observed limitations, a hybrid approach was tested, combining the deep learning models with the addition of the extracted statistical features.
Tables 23 and 24 illustrate the performance variations obtained. The results show distinctive patterns and confirm that statistical features and neural representations complement each other, with different effects in the two classifier tasks:
LSNet: significantly improves on RGB images (up to +9%, LSNet-S SKA), indicating that statistical features fill gaps in automatic extraction. The benefit is less on IR (1-4%), where the thermal component is already well exploited.
ResNet: more variable behaviour. On RGB, integration brings clear improvements (up to 100% in some cases). On IR, however, particularly with ResNet50, a slight deterioration (-0.18%) is observed, suggesting that the addition of features could introduce noise instead of useful information.
In general, RGB models show the greatest improvements in binary classification (up to +10.89%), while in the multi-class context, the benefits are smaller or absent.
Specific improvements in multi-class (RGB):

- LSNet-S: +7.98%;
- LSNet-T: +6.49%;
- LSNet-B: +6.66%;
- ResNet18: +6.28%;

- ResNet50: +7.48%.

Overall, the addition of the features improves the stability of the models, particularly on RGB, reducing the standard deviation by up to -2.45%. This stabilisation effect is relevant in clinical contexts where model reliability is critical.

Conclusion and Future work

This study demonstrates the effectiveness of IR thermography in combination with AI approaches for the classification of DFD, with a particular focus on distinguishing between healthy and diabetic subjects. Across the full range of experimental settings, IR imaging consistently demonstrated superior performance in comparison to RGB imaging, particularly in the context of binary classification tasks. This distinction can be attributed to the enhanced sensitivity of IR imaging to physiological indicators such as local hyperthermia.

Morphological analysis based solely on segmented foot shape also yielded surprisingly strong results, both in binary and multi-class classification. This suggests that structural characteristics of the foot may carry diagnostic information of clinical relevance and merit further investigation.

The integration of statistical features including thermal asymmetry measures and texture descriptors proved especially valuable, notably in multi-class classification where image-only models frequently underperformed. In this context, a hybrid late-fusion strategy that combined deep learning predictions with statistical features consistently yielded the highest accuracy and reduced performance variability. Furthermore, the z-score-based asymmetry analysis provided a more robust and individualized alternative to the fixed 2.2°C threshold, improving the identification of clinically significant thermal anomalies and reducing false positives.

The inclusion of paired thermal data before and after cold stress (T0–T10) further enhanced model performance, suggesting that temporal thermal dynamics provide an additional discriminative dimension. This finding supports the potential of temporal thermography as a diagnostic tool and encourages future studies to incorporate multiple time points to investigate predictive patterns preceding ulcer formation.

Despite these positive results, a major limitation of the study is the lack of longitudinal follow-up. Without post-imaging clinical outcomes, the predictive validity of the proposed features and thresholds remains unverified. Thus, prospective validation is necessary to establish their clinical utility.

Importantly, one of the most critical findings of this work is the consistent difficulty encountered in multi-class classification. Regardless of the modality or model, performance rarely exceeded a balanced accuracy of 50%, even when hybrid approaches were used. This outcome suggests either an insufficient discriminative signal among the current diagnostic classes or suboptimal representation of class variability within the dataset. It highlights the need for re-examining class definitions, expanding datasets, and possibly exploring hierarchical or semi-supervised classification frameworks better suited to such ambiguous clinical settings.

In summary, this study provides strong empirical evidence in favor of a multimodal, AI-driven approach to diabetic foot assessment. Key directions for future research include shape-based modelling, extended temporal thermography, refinement of diagnostic categories, and validation through longitudinal clinical studies.

Data Avaiability Statement

The STANDUP dataset employed in this study is publicly obtainable at https://www.standupproject.eu/manager/?conf=default&route=/STANDUP_Database

Bibliography

[1] David G. Armstrong, Andrew J. M. Boulton, and Sicco A. Bus. Diabetic foot ulcers and their recurrence. *The New England Journal of Medicine*, 376(24): 2367–2375, 2017. doi: 10.1056/NEJMra1615439.

[2] David G. Armstrong, T. W. Tan, A. J. M. Boulton, and S. A. Bus. Diabetic foot ulcers: A review. *JAMA*, 330(1):62–75, 2023. doi: 10.1001/jama.2023.10578.

[3] Jr Sims, D. S., P. R. Cavanagh, and J. S. Ulbrecht. Risk factors in the diabetic foot. recognition and management. *Physical Therapy*, 68(12):1887–1902, 1988. doi: 10.1093/ptj/68.12.1887.

[4] J. Zhang, Y. Qiu, L. Peng, Q. Zhou, Z. Wang, and M. Qi. A comprehensive review of methods based on deep learning for diabetes-related foot ulcers. *Frontiers in Endocrinology*, 13:945020, 2022. doi: 10.3389/fendo.2022.945020.

[5] A. Bougrine, R. Harba, R. Canals, R. Ledee, M. Jabloun, and A. Villeneuve. Segmentation of plantar foot thermal images using prior information. *Sensors*, 22(10):3835, 2022. doi: 10.3390/s22103835.

[6] Rusab Sarmun, Muhammad E. H. Chowdhury, M. Murugappan, Ahmed Aqel, Maymouna Ezzuddin, Syed Mahfuzur Rahman, Amith Khandakar, Sanzida Akter, Rashad Alfkey, and Anwarul Hasan. Diabetic foot ulcer detection: Combining deep learning models for improved localization. *Cognitive Computation*, 16(3):1413–1431, 2024. ISSN 1866-9956.

[7] P. N. Thotad, G. R. Bharamagoudar, and B. S. Anami. Diabetic foot ulcer detection using deep learning approaches. *Sensors International*, 4:100210, 2023.

[8] C. S. Chowdary, D. Ruthwik, K. N. Sai, S. Rishik, and N. Sampath. A comparative analysis of deep learning models for detection of diabetic foot ulcer using foot thermography images. In *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–8, 2024. doi: 10.1109/ICCCNT61001.2024.10724795.

[9] N. Al-Garaawi, R. Ebsim, A. F. H. Alharan, and M. H. Yap. Diabetic foot ulcer classification using mapped binary patterns and convolutional neural networks. *Computers in Biology and Medicine*, 140:105055, 2022. doi: 10.1016/j.compbiomed.2021.105055.

[10] R. Niri, H. Douzi, Y. Lucas, and S. Treuillet. A superpixel-wise fully convolutional neural network approach for diabetic foot ulcer tissue classification. In A. et al. Del Bimbo, editor, *Pattern Recognition. ICPR International Workshops and Challenges. ICPR 2021. Lecture Notes in Computer Science*, volume 12661, pages 266–278. Springer, Cham, 2021. doi: 10.1007/978-3-030-68763-2_23.

[11] C. A. Salazar and M. L. Zequera Díaz. Thermography as a diagnostic tool for early detection of diabetic foot ulceration risk: A review. In *8th Latin American Conference on Biomedical Engineering and 42nd National Conference on Biomedical Engineering - Proceedings of CLAIB-CNIB 2019 (IFMBE Proceedings)*, volume 75, pages 1233–1252. Springer, 2020. doi: 10.1007/978-3-030-30648-9_161.

[12] J. Reyes-Luévano, J. A. Guerrero-Viramontes, J. R. Romo-Andrade, and M. Funes-Gallanzi. Dfu_virnet: A novel visible-infrared cnn to improve diabetic foot ulcer classification and early detection of ulcer risk zones. *Biomedical Signal Processing and Control*, 86:105341, 2023.

[13] Khairul Munadi, Khairun Saddami, Maulisa Oktiana, Roslidar Roslidar, Kahlil Muchtar, Melinda Melinda, Rusdha Muharar, Maimun Syukri, Taufik Fuadi Abidin, and Fitri Arnia. A deep learning method for early detection of diabetic foot using decision fusion and thermal images. *Applied Sciences*, 12(15), 2022. ISSN 2076-3417. doi: 10.3390/app12157524. URL https://www.mdpi.com/2076-3417/12/15/7524.

[14] J. Katual and A. Kaul. Analysis of thermal images with parallel convolutional deep neural network for diabetic foot detection. In *2022 IEEE 3rd Global Conference for Advancement in Technology (GCAT)*, pages 1–5, Bangalore, India, 2022. doi: 10.1109/GCAT55367.2022.9972064.

[15] Keiron O'shea and Ryan Nash. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*, 2015.

[16] A. Ajit, K. Acharya, and A. Samanta. A review of convolutional neural networks. In *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*, pages 1–5, 2020. doi: 10.1109/ic-ETITE47903.2020.049.

[17] Abien Fred Agarap. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*, 2018.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[19] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou. A survey of convolutional neural networks: Analysis, applications, and prospects. *IEEE Transactions on Neural Networks and Learning Systems*, 33(12):6999–7019, 2022. doi: 10.1109/TNNLS.2021.3084827.

[20] Kunihiko Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36(4):193–202, 1980.

[21] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

[22] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, volume 25, 2012.

[23] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[24] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[25] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[26] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[27] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*. PMLR, 2019.

[28] D. Bouallal, A. Bougrine, R. Harba, R. Canals, H. Douzi, L. Vilcahuaman, and H. Arbanil. Standup database of plantar foot thermal and rgb images for early ulcer detection. *Open Research Europe*, 2:77, 2022. doi: 10.12688/openreseurope.14416.2.

[29] K. KC, Z. Yin, D. Li, and Z. Wu. Impacts of background removal on convolutional neural networks for plant disease classification in-situ. *Agriculture*, 11(9):827, 2021. doi: 10.3390/agriculture11090827. URL https://www.mdpi.com/2077-0472/11/9/827.

[30] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, G. Cottrell, P. Dollár, S. Belongie, B. Hariharan, P. Perona, M. Hebert, L. Van Der Maaten, J. Malik, and R. Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[31] A. Wang, H. Chen, Z. Lin, J. Han, and G. Ding. Lsnet: See large, focus small. *arXiv preprint arXiv:2503.23135*, 2025. URL https://arxiv.org/abs/2503.23135.

[32] D. Bouallal, H. Douzi, and R. Harba. Diabetic foot thermal image segmentation using double encoder-resunet (de-resunet). *Journal of Medical Engineering & Technology*, 46(5):378–392, 2022. doi: 10.1080/03091902.2022.2103701.

[33] S. Y. Boulahia, A. Amamra, M. R. Madi, and S. Daikh. Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition. *Machine Vision and Applications*, 32(6):121, 2021. doi: 10.1007/s00138-021-01232-4.

[34] John D. Kelleher, Brian Mac Namee, and Aoife D'Arcy. *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies*. MIT Press, 2020.

[35] N. Arteaga-Marrero, A. Hernández-Guedes, J. Ortega-Rodríguez, and J. Ruiz-Alzola. State-of-the-art features for early-stage detection of diabetic foot ulcers based on thermograms. *Biomedicines*, 11(12):3209, 2023. doi: 10.3390/biomedicines11123209.

[36] L. A. Lavery, K. R. Higgins, D. R. Lanctot, G. P. Constantinides, R. G. Zamorano, K. A. Athanasiou, and C. M. Agrawal. Preventing diabetic foot ulcer recurrence in high-risk patients: use of temperature monitoring as a self-assessment tool. *Diabetes Care*, 30(1):14–20, 2007. doi: 10.2337/dc06-1600.

[37] J. Featherston, A. M. Wijlens, and J. J. van Netten. Is a left-to-right ¿ 2.2° c difference a valid measurement to predict diabetic foot ulceration in people with diabetes and a history of diabetic foot ulceration? *The International Journal of Lower Extremity Wounds*, 23(3):428–435, 2024. doi: 10.1177/15347346231215513.

[38] N. Hollmann, S. Müller, K. Eggensperger, and F. Hutter. Tabpfn: A transformer that solves small tabular classification problems in a second. *arXiv preprint arXiv:2207.01848*, 2022. URL https://arxiv.org/abs/2207.01848.

[39] Pl. Li, Qf. Xiao, Kl. Yick, et al. A novel deep learning approach to classify 3d foot types of diabetic patients. *Scientific Reports*, 15:13819, 2025. doi: 10.1038/s41598-025-98471-5.