POLITECNICO DI TORINO

MASTER's Degree in Data Science and Engineering

MASTER's Degree Thesis

Thermographic images post processing for welding process monitoring

Supervisors

Prof. RAFFAELLA SESANA

Prof. LUCA SANTORO

Candidate

REZA KHARAMANI

MARCH 2026

Abstract

Ensuring weld quality in modern manufacturing requires inspection methods that are both reliable and compatible with real-time production environments. Conventional non-destructive testing (NDT) techniques are typically applied after welding, are costly, and do not provide continuous feedback on process stability or weld integrity. This thesis investigates thermographic image and video post-processing as a basis for automated inline monitoring of laser and TIG welding processes, with the objective of detecting defects directly from thermal data and supporting real-time quality assessment.

A laboratory dataset of welding runs was acquired using a high-frame-rate long-wave infrared thermal camera, capturing the full thermal evolution of the molten pool and surrounding heat-affected zone. The dataset includes both nominal operating conditions and a variety of intentionally induced defects, enabling a controlled evaluation of detection methods. As an initial step, classical machine-learning baselines are constructed by extracting compact, physics-informed thermal descriptors from radiometrically normalised frames, including temperature statistics, spatial gradients, intensity ranges and simple temporal indicators related to heat dissipation. While these handcrafted features allow traditional classifiers to achieve moderate and interpretable performance, their limited expressive power constrains their ability to capture complex spatial patterns present in welding thermography.

To address these limitations, the thesis introduces a dedicated deep-learning approach based on a lightweight two-dimensional convolutional neural network (2D CNN), termed *ThermalNet-V1*. Operating on a frame-by-frame basis, the network learns discriminative spatial representations of heat distribution, melt-pool geometry and thermal asymmetries directly from single infrared images. The architecture combines multi-scale convolutions, residual connections and channel-attention mechanisms while remaining computationally efficient and suitable for real-time deployment. Compared to classical baselines, the proposed model demonstrates improved defect discrimination and more stable probability estimates across different movement patterns and acquisition conditions.

Despite these improvements, supervised classification approaches remain sensitive to practical constraints such as limited labelled data, strong class imbalance between normal and defective samples, and the presence of defect patterns not observed during training. To overcome these challenges, the thesis further proposes an unsupervised AutoEncoder-based anomaly detection framework trained exclusively on thermographic data from normal welds. By learning a compact latent representation of nominal thermal behaviour, the AutoEncoder reconstructs normal patterns with high fidelity, while frames exhibiting abnormal heat dissipation, unexpected hotspots or irregular cooling dynamics produce significantly higher reconstruction errors and are flagged as anomalies. Experimental results show that this anomaly-detection approach complements supervised models and offers superior robustness when encountering rare, subtle or previously unseen defect signatures. Overall, the

proposed framework provides a generalisable and industrially viable solution for inline thermographic weld quality monitoring.

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to ***Professor Raffaella Sesana*** *for their invaluable guidance, support, and encouragement throughout the course of this thesis. Their expertise, insight, and patience were instrumental in shaping this work, and I am deeply appreciative of the time and effort they dedicated to supervising my research.*

I am profoundly grateful to my ***mother****, whose unwavering love, sacrifices, and constant support have been a source of strength and motivation throughout my academic journey. I also wish to thank my* ***sister*** *and* ***brother*** *for their encouragement, understanding, and belief in me, which have meant more than words can express.*

Finally, I would like to dedicate this work to the memory of my ***father****, who is no longer with us. Although he did not live to see the completion of this thesis, his values, guidance, and support continue to inspire me. This achievement is presented in his loving memory.*

Table of Contents

List of Figures

List of Tables

Acronyms

NDT	Non-Destructive Testing.
QA	Quality Assurance.
HAZ	Heat-Affected Zone.
IR	Infrared.
LWIR	Long-Wave Infrared.
MWIR	Mid-Wave Infrared.
SNR	Signal-to-Noise Ratio.
PCA	Principal Component Analysis.
PCT	Principal Component Thermography.
ML	Machine Learning.
DL	Deep Learning.
CNN	Convolutional Neural Network.
SVM	Support Vector Machine.
BCE	Binary Cross-Entropy.
ROC	Receiver Operating Characteristic.
AUC	Area Under the Curve.
PR	Precision–Recall.

AP	Average Precision.
GAP	Global Average Pooling.
MLP	Multi-Layer Perceptron.
HDF5	Hierarchical Data Format version 5.
TIG	Tungsten Inert Gas Welding.
GTAW	Gas Tungsten Arc Welding.
MIG	Metal Inert Gas Welding.
MAG	Metal Active Gas Welding.
GMAW	Gas Metal Arc Welding.
VT	Visual Testing.
RT	Radiographic Testing.
UT	Ultrasonic Testing.
MT	Magnetic-Particle Testing.
USB	Universal Serial Bus.

Chapter 1

Introduction

1.1 Background and Motivation

Welding is a key joining technology in manufacturing, automotive, aerospace and construction industries. The structural integrity of welded components has a direct impact on product safety, service life and cost of ownership. Typical weld defects such as lack of fusion or penetration, porosity, cracks and undercut may initiate premature failures if they remain undetected. For this reason, non-destructive testing (NDT) and quality assurance (QA) are mandatory steps in industrial welding workflows.

Conventional QA is largely based on visual inspection, radiography, ultrasonic testing or other NDT techniques applied after welding has taken place. These methods are effective for offline acceptance testing, but they have three important limitations in modern production:

- they are applied late in the process, so defects are detected only after material, time and energy have already been invested;
- they depend strongly on operator skill and subjective interpretation;
- they do not provide continuous information about process stability during welding and therefore cannot prevent defect formation in real time.

In parallel, industry is moving towards digitalised and data-driven production systems under the umbrella of "Industry 4.0". Inline process monitoring and automatic decision support are central elements of this evolution. For welding, this means shifting from purely post-process inspection to continuous monitoring of the arc, melt pool and heat-affected zone (HAZ), with the goal of detecting anomalies as early as possible and enabling adaptive control.

Infrared thermography has emerged as a promising sensing modality for this purpose. Thermal cameras provide non-contact, full-field measurements of surface temperature distributions with frame rates compatible with typical welding speeds. When properly configured, they can visualise the dynamic behaviour of the weld pool and surrounding material and reveal changes in heat flow associated with defects or parameter deviations [1, 2, 3]. However, raw thermal videos acquired on metallic

surfaces are affected by emissivity changes, reflections from the arc and environment, sensor noise and other artefacts. Without robust post-processing and intelligent analysis, the full potential of thermography for weld QA cannot be exploited.

Recent advances in machine learning and deep learning provide powerful tools to analyse complex image sequences and to detect subtle patterns that are not easily captured by manual rules. In particular, convolutional neural networks (CNNs) and related architectures can learn defect-sensitive representations directly from thermal data and have shown strong performance in thermographic NDT and welding applications [4, 5, 6]. This thesis builds on these developments and investigates how a dedicated thermographic post-processing and learning pipeline can support automatic weld defect detection in an inline inspection scenario.

1.2 Problem Statement

Despite the promising characteristics of thermographic sensing and modern learning-based models, several challenges hinder their practical use for welding QA:

- **Noisy and unstable thermal data.** Metallic surfaces exhibit temperature- and angle-dependent emissivity, high reflectivity and strong arc glare. As a result, raw thermal frames contain artefacts that can mask or mimic defect signatures.
- **Lack of systematic post-processing.** Many existing systems rely on basic filtering and thresholding, which are often tuned ad hoc and are not easily transferable between setups.
- **Limited labelled datasets.** Collecting and annotating defect examples in welding is expensive. Publicly available thermal weld datasets are scarce, and individual industrial datasets typically contain only a few defective sequences compared with large amounts of normal data.
- **Risk of data leakage.** Thermal frames from the same physical run are strongly correlated. If such frames are randomly split between training and test sets, reported performance can be overly optimistic and not representative of deployment.
- **Real-time constraints.** Inline monitoring requires models and post-processing pipelines that can operate at camera frame rate on embedded or constrained hardware, without compromising detection performance.

The central problem addressed in this thesis is therefore the design, implementation and evaluation of a thermographic post-processing and learning framework that can reliably distinguish between normal and defective thermal behaviour in welding-related inspection sequences, while respecting data limitations and deployment constraints.

1.3 Research Objectives

The overall objective of this work is to develop and validate an automated post-processing and classification framework for thermographic image sequences that supports weld quality assessment in a realistic inline inspection setting.

More specifically, the thesis pursues the following objectives:

1. **Dataset construction and characterisation.** Build a structured thermal dataset composed of normal and defective sequences acquired with an infrared camera, and quantify its statistical properties (duration, resolution, frame rate, temperature ranges and reconstruction-error characteristics) for both classes.

2. **Physics-aware preprocessing.** Design a preprocessing pipeline for thermal sequences that includes spatial resizing, global normalisation and lightweight augmentation, and that is compatible with single-channel CNN models and real-time constraints.

3. **Classical machine-learning baselines.** Engineer compact, physics-informed thermal features and implement classical classifiers (Random Forest, Support Vector Machine and Gradient Boosting) to establish interpretable baselines for defect detection.

4. **Deep-learning model design.** Propose a dedicated convolutional neural network, ThermalNet-V1, that exploits multi-scale convolutions, residual connections and channel attention for frame-level defect probability estimation from single thermal images.

5. **Robust evaluation protocol.** Define and implement an evaluation protocol based on file-level data splitting, threshold-independent ranking metrics (ROC–AUC and average precision) and threshold-dependent operating-point metrics (precision, recall and F1-score) to obtain realistic performance estimates.

6. **Qualitative and quantitative analysis.** Analyse model behaviour on both circular zigzag and linear movement patterns by combining numerical metrics with temporal probability plots and representative frame examples.

1.4 Research Questions

In line with the above objectives, the thesis addresses the following research questions:

1. How do physics-aware preprocessing and global normalisation of thermal frames influence the separability between normal and defect classes in weld-related inspection data?

2. Which compact, hand-crafted thermal features (e.g., temperature statistics, gradients and simple temporal descriptors) are most informative for classical machine-learning models in this context?

3. To what extent can a lightweight CNN architecture such as ThermalNet-V1 outperform classical baselines for frame-level defect probability estimation, given the available dataset size and class imbalance?

4. How important is file-level data splitting for obtaining realistic estimates of model performance compared with more naive frame-level splits that risk temporal leakage?

5. Can a model trained on thermal sequences acquired under one type of sensor trajectory (e.g., circular zigzag motion) generalise to a different trajectory (linear motion) without retraining, and what does this reveal about the robustness of learned representations?

1.5 Scope and Limitations

The scope of this thesis is deliberately focused to allow a detailed and controlled investigation:

- **Process and materials.** The experiments are based on controlled thermal inspection sequences recorded on metallic specimens with welding-like heating trajectories. The emphasis is on the thermographic patterns associated with normal and defective behaviour rather than on a specific welding standard or material grade.

- **Sensor configuration.** All data are acquired with a single thermal camera model at a fixed spatial resolution of 240×384 pixels and a frame rate of approximately 19 fps. The work does not compare multiple camera technologies or wavelengths.

- **Problem formulation.** The primary task is frame-level binary classification (normal vs. defect). Pixel-wise defect segmentation and multi-class defect categorisation are outside the scope of this study, although they are discussed as future extensions.

- **Real-time aspects.** The implementation is designed with real-time deployment in mind (single-frame inference, compact architecture, limited preprocessing), but actual integration with a closed-loop welding control system is not performed.

These choices introduce several limitations. The dataset, while carefully curated, remains relatively small, especially for the defect class; this constrains the depth of the models and encourages conservative regularisation and architecture design. Generalisation to other welding processes, materials, camera types or acquisition geometries is not guaranteed and would require additional data collection and adaptation. Finally, the focus on frame-level decisions means that temporal aggregation and sequence-level reasoning are handled outside the core model.

1.6 Industrial Context and Data Source

The thermal dataset used in this thesis is collected from an inline inspection scenario where an infrared camera observes heating trajectories over metallic surfaces. Two main movement patterns are represented: (i) a circular zigzag motion, and (ii) a linear motion. In both cases, the sensor follows a repeatable trajectory at a roughly constant speed, generating thermal sequences that capture heating and cooling cycles along the path.

Normal sequences correspond to stable operation without intentional faults and exhibit smooth, repeatable thermal oscillations with low reconstruction error. Defect sequences are recorded under altered conditions that induce abnormal thermal behaviour, such as changes in the trajectory, partial loss of contact or other process perturbations. As a result, they show irregular temperature patterns and significantly higher reconstruction errors over specific time intervals.

All recordings are stored in HDF5-based containers (`.h5` files) that include raw temperature maps, timestamps and per-frame reconstruction errors. This unified storage format simplifies indexing, preprocessing and subsequent model training. A detailed quantitative characterisation of these data—including frame counts, durations, temperature statistics and reconstruction-error distributions—is provided in Chapter 3.

1.7 Thesis Organization

The remainder of this document is organised as follows:

- **Chapter 2 – Literature Review and Related Work** surveys welding process monitoring, fundamentals of infrared thermography, signal processing techniques for thermographic NDT, and state-of-the-art machine-learning and deep-learning approaches for thermal weld inspection. It identifies existing gaps and motivates the methodological choices of this thesis.

- **Chapter 3 – Methodology** describes the overall system design, the thermal dataset structure, the preprocessing pipeline, the construction of classical machine-learning baselines and the proposed ThermalNet-V1 architecture. It also details the data-storage conventions, splitting strategy and training procedure implemented in the `train_thermalnet_v1.py` script.

- **Chapter 4 – Evaluation Metrics** defines the evaluation protocol, including the loss function, threshold-independent ranking metrics, threshold-dependent operating-point metrics and the strategy used for decision-threshold calibration. It also presents qualitative analyses of model predictions on representative normal and defect sequences for both circular and linear movement datasets.

- **Chapter 5 – Conclusions and Future Work** summarises the main findings and contributions, discusses the practical implications for thermographic weld

inspection, outlines the limitations of the current study and proposes directions for future research, including sequence-level models, multi-modal sensing and deployment-oriented improvements.

This structure is consistent with the overall thesis plan presented in the introductory documentation and ensures a logical progression from context and motivation, through theory and methodology, to experimental evaluation and final conclusions.

Chapter 2

Literature Review and Related Work

2.1 Welding Process Monitoring: Fundamentals and Challenges

Welding processes (e.g., TIG, MIG/MAG, laser) are highly sensitive to process parameters such as current, voltage, travel speed, shielding gas and joint preparation, as well as to material conditions. These factors give rise to typical defects including porosity, lack of fusion or penetration, cracks, undercut and burn-through [7, 8]. Conventional non-destructive testing (NDT) methods (ultrasonic testing, radiography, eddy current, etc.) provide post-process quality assurance, but are limited for real-time control, may be costly, and can involve safety constraints in production environments [2, 9]. Inline and in-process monitoring is therefore central to modern Industry 4.0 paradigms to reduce rework, increase reliability and move towards autonomous or operator-assisted control.

2.1.1 Overview of Industrial Welding Processes and Defects

From a metallurgical perspective, defect formation is governed by the complex interaction between the heat source, joint geometry, filler material and boundary conditions [7]. Typical fusion welding processes such as gas tungsten arc welding (GTAW/TIG), gas metal arc welding (GMAW/MIG/MAG) and laser beam welding differ in energy density, penetration profile and process dynamics, which in turn shape the thermal cycles experienced by the workpiece. Key quality-relevant defects include lack of fusion or penetration, porosity, hot and cold cracks, undercut, overlap and excessive reinforcement; many of these are directly or indirectly linked to local heat-flow anomalies, shielding-gas disturbances or improper travel speed. Design codes and fabrication standards such as AWS D1.1 for structural steel define allowable defect types and limits, and mandate appropriate NDT procedures accordingly [10].

2.1.2 Conventional NDT and the Move Toward Inline Monitoring

Traditional NDT methods—visual testing (VT), radiographic testing (RT), ultrasonic testing (UT), magnetic-particle testing (MT) and others—are routinely applied according to international standards (e.g., ISO 17637 for visual inspection) to ensure compliance with quality requirements [11, 8]. While highly effective for final acceptance, these methods are predominantly offline and sample-based: only a fraction of welds or locations is inspected, and the feedback delay makes it difficult to correct process deviations in real time. Furthermore, radiography and some UT techniques require access to both sides of the joint and can involve safety or access constraints in production environments.

To address these limitations, there is growing interest in inline and in-process monitoring solutions that can be integrated near the welding head and operate continuously. Candidate sensing modalities include electrical process signals (current, voltage), acoustic emission, optical emission spectroscopy, visible cameras and thermal cameras. Among these, thermography is particularly attractive because it directly visualises the heat flow and molten-pool behaviour, which are primary carriers of information about weld quality.

Thermal camera imaging provides a non-contact modality to visualize surface temperature fields and the dynamics of the molten pool, keyhole and heat-affected zone (HAZ), enabling inference about process stability and defect formation during welding [1, 2, 3]. Metallic surfaces, however, pose specific challenges: low emissivity, high reflectivity, arc glare and rapidly changing emissivity between solid and liquid phases can bias absolute temperature estimates and degrade signal-to-noise. Consequently, robust post-processing is required to extract defect-sensitive features from thermal sequences; in many practical systems, relative-intensity and physics-aware compensation strategies are preferred over absolute thermometry for production monitoring [2, 3].

2.2 Principles of Thermal Camera Thermography for NDT

Thermal cameras detect emitted radiance and, with suitable models, infer surface temperature. Two broad acquisition regimes are typically distinguished [1, 2]:

- **Passive thermography**, which leverages intrinsic heat sources (e.g., the weld pool or hot workpiece).
- **Active thermography**, which injects an external heat flux (pulsed, lock-in, vibrothermography, etc.) and analyzes thermal diffusion to reveal subsurface anomalies.

Over the last decades, several *signal processing* paradigms have been developed for active thermography and later adapted to welding:

- **Pulsed Phase Thermography (PPT)** converts temporal decay signals into the frequency domain and uses phase images, which partially mitigate non-uniform heating and emissivity variations.
- **Principal Component Thermography (PCT)** decomposes thermographic sequences into empirical orthogonal functions to enhance signal-to-noise ratio (SNR) and reveal latent patterns [2].
- **Thermographic Signal Reconstruction (TSR)** and related normalization strategies reconstruct smoothed transients and compute derivatives, which enhance defect contrast and compress data.

For metallic welding applications, passive thermal-camera monitoring typically dominates because of the very high intrinsic heat and the need for inline operation at camera frame rates compatible with production speeds. Nevertheless, many post-processing concepts developed for active thermography (e.g., detrending, normalization, spatio-temporal filtering) are directly applicable and motivate the design of robust pipelines for industrial welding QA.

2.2.1 Infrared Detectors, Spectral Bands and Measurement Limits

Thermal cameras used in industrial NDT are typically based on either uncooled microbolometer arrays operating in the long-wave infrared (LWIR, 8–14 µm) or cooled photon detectors operating in the mid-wave infrared (MWIR, 3–5 µm) [12, 2]. LWIR microbolometers are robust and cost-effective but have limited temporal response and sensitivity at very high temperatures, whereas cooled MWIR systems offer higher dynamic range and frame rates at the expense of cost, complexity and maintenance. In welding applications, sensor selection must balance temperature range, spatial resolution, frame rate and robustness to arc radiation and reflections.

Measurement accuracy depends on correct modelling of emissivity, reflected and transmitted components, and on appropriate radiometric calibration. Maldague and others emphasise that, particularly on metals, absolute temperature readings can be biased by several tens of degrees if emissivity is not carefully characterised [1]. In many inline QA scenarios, this motivates the use of relative or normalised temperature measures, gradients and cooling-rate surrogates rather than strict thermometry.

2.2.2 Calibration, Reference Standards and Good Practice

Robust thermographic NDT requires regular radiometric calibration against blackbody references or calibrated sources, verification of spatial non-uniformity and compensation for bad pixels and drift [1, 12]. International standards for active thermography in composites and metallic structures define recommended acquisition geometries, heating protocols, reference samples and reporting procedures; although not specific to welding, these documents provide valuable templates for experimental design and documentation in weld thermography. Adopting such practices increases

the reproducibility and comparability of research results, which is essential when machine-learning models are trained on thermographic data.

2.3 Thermography in Welding: Empirical Evidence and Industrial Practice

Early feasibility studies demonstrated that thermal camera monitoring can detect weld defects and capture thermal signatures associated with degradation of weld quality. Several works reported that defective regions often exhibit different cooling rates or spatial temperature distributions compared to sound welds; for example, defective regions may lose heat faster due to altered heat flow paths or reduced cross-section [3, 2]. Industrial reports emphasize the difficulty of accurate absolute temperature measurement on metals and instead focus on relative metrics: emissivity-transition patterns and intensity profiles along the seam correlate with weld quality, shielding-gas flow and torch misalignment. Commercial inline systems based on mid-wave and long-wave thermal cameras operating at frame rates from a few tens up to hundreds of Hz have shown that, with suitable optics and measurement geometry, system-level QA can be achieved in real time [3, 5].

Recent work has also targeted TIG-specific challenges, proposing reflected-temperature corrections and temporal interpolation aided by high-frequency process signals to compensate for limited camera frame rates. These studies report that asymmetric temperature fields, partial penetration and shape errors require both sophisticated preprocessing and carefully designed decision logic, often combining heuristic rules with learning-based classifiers.

For laser welding, commercial and research-grade inline systems now integrate on-axis or off-axis high-speed MWIR sensors with the welding head. Reported defect-detection performance above 95 % in real time has been achieved by monitoring melt-pool geometry and thermal patterns, demonstrating that thermography-based inline weld QA is viable for production settings [3, 5].

2.3.1 Cooling-Rate Signatures and Spatial Indicators

Cooling-rate analysis is a recurring theme in thermographic weld monitoring. Defective regions often display altered cooling slopes—either faster cooling due to reduced cross-section and increased heat loss, or slower cooling associated with lack of fusion and local heat accumulation [3, 2]. Quantitatively, this behaviour can be captured with simple finite-difference approximations of the mean temperature over time, but more sophisticated approaches consider spatially resolved cooling maps or phase-based transforms (as in PPT and PCT) that enhance defect contrast.

Spatial indicators such as asymmetry indices between the two sides of the weld, melt-pool width and eccentricity, and the position of isotherm contours relative to the nominal joint line have also been proposed as robust quality proxies [2, 5]. These descriptors are particularly valuable under varying process parameters because they

normalise out global temperature level to some extent and focus on morphology.

2.3.2 Commercial Systems and Industrial Adoption

Several commercial and research prototypes integrate thermal cameras directly into welding heads or inspection gantries, combining thermography with seam tracking, vision systems and process-signal monitoring. Reported systems for laser and resistance spot welding achieve high detection rates for gross defects and can operate at industrial line speeds [9, 13, 14]. In many cases, the decision logic is still based on engineered thresholds or shallow classifiers.

The main barriers to broader adoption identified in the literature include: (i) sensitivity to surface condition and emissivity variations; (ii) calibration and maintenance effort; (iii) integration with existing QA and data-management workflows; and (iv) the need for interpretable decision criteria acceptable to quality engineers [2, 15]. These challenges motivate the use of learning-based methods that can adapt to complex environments while still providing traceable indicators and confidence measures.

2.4 Post-Processing of Thermographic Sequences for Welding

Thermal videos acquired during welding are affected by emissivity changes, arc reflections, sensor saturation and mechanical jitter. Accordingly, robust post-processing pipelines typically include the following steps [1, 2]:

- **Radiometric pre-filtering**: temporal smoothing; spatial filtering (median, bilateral); and arc suppression via masks or frequency-domain separation.
- **Emissivity and reflectivity handling**: relative-intensity normalization, background subtraction and, when auxiliary sensors are available, reflected-temperature correction.
- **Feature extraction**: temperature gradients across the seam, isotherm morphology, cooling rates, HAZ boundary dynamics, asymmetry indices, temporal stability metrics and geometric proxies for the melt pool (width, area, eccentricity).
- **Dimensionality reduction**: PCA or PCT to derive compact, high-SNR representations preserving defect-sensitive modes.
- **Decision logic**: thresholding, anomaly detection or supervised ML/DL classifiers (discussed in later sections).

For laser and TIG welding, spatio-temporal features describing pool-shape evolution and cooling profiles are especially discriminative for lack of fusion or penetration, undercut and burn-through. These descriptors can be exploited by both classical machine learning and deep learning models to build reliable inline QA systems.

2.5 Machine Learning and Deep Learning for Thermal Weld Monitoring

Classical machine learning with hand-crafted features has demonstrated strong performance on curated thermal datasets. Random forests, support vector machines (SVMs) and gradient boosting trees trained on geometric, statistical and textural features extracted from thermal frames have been successfully applied to weld-surface defect classification and thermal NDT in other domains [2, 3]. In thermal welding analysis, engineered features such as temperature profiles, gradients, cooling-rate estimates and texture descriptors can feed tree-based ensembles that provide interpretable decisions and feature-importance rankings.

2.5.1 Classical Machine Learning Approaches

In early thermographic weld monitoring systems, classical classifiers were typically trained on low-dimensional feature vectors summarising each frame or region of interest. Common descriptors include statistical moments of temperature, contrast measures between the weld pool and surrounding material, gradient-based edge metrics and simple temporal features derived from cooling curves. Tree-based ensembles such as random forests and gradient boosting machines are particularly attractive because they handle heterogeneous features, require limited parameter tuning and provide built-in measures of feature importance. Support vector machines with radial-basis-function kernels have also been widely used for binary normal/defect discrimination due to their margin-based formulation and good performance in low-data regimes.

These approaches serve as strong baselines and remain relevant in industrial contexts where interpretability, compact models and limited computational resources are priorities. However, their reliance on hand-crafted features may limit their ability to exploit complex spatial and temporal patterns in high-resolution thermal video.

2.5.2 Deep Learning Architectures for Thermal Data

Deep learning (DL) reduces dependence on manual feature design and can learn spatio-temporal representations directly from thermal images or video. Convolutional neural networks (CNNs) have been used for classification, detection and segmentation on thermal frames (or stacked temporal windows), both for welding QA and for pulsed thermography NDT [6, 13, 14, 4]. In laser welding, customized CNNs operating on thermal video sequences have achieved high discriminative capability for defective weld identification, supporting the feasibility of inline deployment [5]. For TIG or visible-spectrum weld-pool monitoring, ResNet-based models with attention and robust preprocessing have demonstrated accurate discrimination between different weld states, indicating that similar architectures can be adapted to LWIR/MWIR data.

Beyond frame-based models, spatio-temporal DL architectures such as 3D CNNs, ConvLSTMs and temporal transformers learn dynamic patterns (pool oscillations, cooling curves) that are predictive of defect formation [16, 17]. For thermal NDT more broadly, object detection (e.g., YOLO, Faster R-CNN), semantic and instance segmentation (U-Net, Mask R-CNN) and hybrid models have been evaluated on pulsed thermography datasets, often matching or exceeding traditional contrast methods for automatic defect localization and identification [4, 18, 19]. These results guide architecture selection, loss design and evaluation strategies for thermal welding QA.

2.5.3 Evaluation Protocols and Generalisation Issues

Performance reports in the literature vary widely because of differences in dataset size, defect types, acquisition geometry and evaluation protocol. Several authors have highlighted that naive random splits at the frame or pixel level can artificially inflate accuracy by leaking information between training and test sets, especially when adjacent frames or neighbouring pixels share strong correlations [20]. File-level or specimen-level splits are therefore recommended to better reflect real deployment conditions, where an entire new weld or component must be classified based on unseen data. In addition, threshold-independent metrics such as ROC–AUC and average precision are increasingly used alongside confusion-matrix-based measures to characterise classifier behaviour under varying operating points.

2.6 Datasets, Sensors and Deployment Considerations

Reported thermal welding datasets typically combine long-wave (LWIR, 8–14 µm) or mid-wave (MWIR, 3–5 µm) thermal cameras with frame rates between 30 and 1000 Hz, trading off resolution, dynamic range and cost [2, 3]. On-axis sensing maximizes signal but complicates glare management; off-axis views ease arc suppression at the expense of foreshortening and occlusions. Working distance and field of view must be chosen to resolve the seam with sufficient pixels per millimetre for geometric descriptors to be robust.

Ground-truth labels are often obtained via controlled parameter sweeps (current, speed, gas flow) and intentional defect seeding, combined with destructive testing or high-fidelity UT/CT on representative coupons. However, label scarcity and imperfect frame-level annotations are common and motivate the use of group-aware validation (splitting by file), semi-supervised learning and domain adaptation techniques, especially when transferring between materials, surface conditions or camera models.

Real-time constraints impose stringent limits on model complexity and preprocessing cost. Lightweight models, quantization and region-of-interest (ROI) based inference are common strategies: many industrial systems track a narrow seam ROI to reduce compute and stabilize features, while more complex analyses can be

offloaded to offline inspection or higher-performance hardware.

2.6.1 Public Datasets, Reproducibility and Benchmarking

Unlike visible-spectrum computer vision, thermographic weld datasets are rarely public, and most reported studies rely on proprietary industrial data or laboratory experiments tailored to specific applications [2]. This scarcity of openly available benchmarks complicates direct comparison between methods and slows down progress in model development. Some QIRT and NDT communities have started to release thermography datasets for composite inspection and general NDT, but these are not yet tailored to welding.

For this reason, recent work places increasing emphasis on thorough dataset documentation, including acquisition geometry, materials, process parameters, camera type and calibration assumptions. Such documentation, combined with clear train/validation/test splits and the release of code and configuration files, is essential for reproducibility—particularly when deep learning is involved.

2.7 Research Gaps and Opportunities Aligned to this Thesis

Despite the progress in thermal welding monitoring, several gaps remain that directly motivate this thesis:

- **Robustness to emissivity and reflections**: there is a need for physics-informed normalization strategies and reflection models that stabilize features across varying surface conditions [1, 2].
- **Spatio-temporal learning**: most industrial solutions still rely on frame-wise thresholds or simple features; spatio-temporal DL models (3D CNNs, ConvLSTMs, transformers) can better capture dynamics linked to defect formation [16, 17].
- **Low-frame-rate compensation**: fusing high-frequency process signals (current, voltage) with thermal frames for temporal interpolation or data-driven super-resolution in time remains under-explored.
- **Weak supervision and domain adaptation**: limited labeled data and variability across materials and processes motivate semi/self-supervised pretraining and domain adaptation methods.
- **Uncertainty and explainability**: industrial deployment requires calibrated probabilities, reliability indicators and interpretable saliency maps rather than black-box alarms only.

This thesis addresses these gaps by designing a post-processing framework for thermographic sequences that (i) normalizes and enhances thermal data; (ii) extracts

defect-sensitive spatio-temporal features; (iii) leverages ML/DL for classification and segmentation; and (iv) is compatible with real-time constraints and industrial deployment.

2.8 Advanced Signal Processing and Inversion for Thermographic Data

Beyond basic filtering and normalization, mature thermography pipelines employ advanced preprocessing and inversion techniques that are directly applicable to welding. Thermographic Signal Reconstruction (TSR) fits each pixel's transient with a polynomial in the log–log domain and uses derivatives to enhance defect contrast and compress data; TSR-derived coefficient images can outperform simple PCT or PPT features in many scenarios [2, 18]. Non-uniform heating and drifting backgrounds can be mitigated using Gaussian-model background optimization or low-order polynomial models with local suppression, which significantly improves SNR.

Approximate depth estimation is possible from thermal contrast curves by inspecting peak times, phase delay (PPT/lock-in) or matching to model-based cooling profiles; such ideas translate to estimating lack-of-penetration severity or bead thickness variations in welds. Orientation-aware gradients and histogram-of-oriented-gradients (HOG)-like descriptors, combined with local or global segmentation, have been shown to be robust under non-uniform fields and low contrast, sometimes outperforming generic DL when labels are scarce. For fast transients where camera frame rates are limiting, super-frequency sampling and reconstruction can effectively increase effective temporal resolution, benefiting the analysis of short-lived thermal events during welding.

Thermal image restoration methods based on LWIR-specific statistical priors and total-variation (TV) regularization can reduce residual non-uniformity and sensor artefacts. Such restoration not only improves visual quality but also stabilizes downstream features and classifier performance, especially for subtle defect signatures [2, 19].

2.9 Active Thermography Modalities Relevant to Welding QA

While this thesis focuses on passive inline monitoring, active thermography modalities inform processing choices and provide powerful tools for offline validation. Lock-in thermography (optical or eddy-current excited) yields amplitude and phase images with depth selectivity and is effective on metallic weld coupons for controlled studies. Vibrothermography (ultrasound-excited) is highly sensitive to cracks and kissing defects via frictional or viscoelastic heating; although it generally requires contact coupling and is best suited for offline inspection, it establishes useful benchmarks for

detection sensitivity.

Comparative studies show that pulsed, lock-in and vibrothermography each exhibit trade-offs in depth reach, defect selectivity and practicality. Combining phase-based features (PPT), TSR and PCT often yields the most robust performance across materials. In the welding context, active methods are especially valuable for generating labeled benchmarks and calibrating relationships between passive thermal features and actual defect characteristics.

2.10 Emissivity and Absolute Temperature in Welding Thermography

Accurate absolute thermometry on metals is challenging because emissivity varies with temperature, phase, surface finish, viewing angle and wavelength. Assuming constant emissivity in thermography software can bias temperature estimates and any features derived from them [1, 2]. Practical options include:

- **Process-informed emissivity estimation**: using known melting-point isotherms and bead geometry to back-calculate effective emissivity during laser welding.
- **Temperature-dependent emissivity curves**: fitting emissivity models from thermocouple-calibrated experiments for relevant steels and aluminium alloys.
- **Multispectral or pixel-wise emissivity**: when hardware permits, multispectral MWIR enables pixel-level emissivity mapping concurrent with imaging.

In many industrial QA scenarios, relative measures (normalized intensities, gradients, cooling-rate surrogates) remain preferable for defect detection, while absolute temperature is reserved for specific process studies or model calibration.

2.11 Two-Color Pyrometry and Two-Color Thermography

Ratio-based temperature estimation mitigates emissivity dependence by measuring radiance in two close spectral bands. Recent work on rear weld-pool monitoring in GTAW using in-house two-color pyrometers has produced mean-temperature signals correlated with bead geometry and robust to arc radiation when combined with appropriate ROIs. Two-color thermography extends this concept to spatial fields, enabling dynamic behaviour measurement with reduced emissivity sensitivity at moderate temperatures.

These approaches are complementary to broadband LWIR/MWIR cameras. For post-processing, ratio features can augment thermal image descriptors and may be combined with dual-sensor heads or small-band filters in setups where accurate temperature control or model calibration is critical.

2.12 Standards, Calibration and Best Practice

Even for passive welding thermography, adopting NDT standards and community best practices improves the rigour and reproducibility of experimental studies. Technique standards for flash thermography of composites and vibroacoustic thermography codify acquisition protocols, processing workflows and reporting guidelines that translate well to welding QA. Foundational handbooks and monographs such as those by Maldague and the British Institute of NDT (BINDT) provide theory, detector behaviour, emissivity considerations and quality practices relevant to metallic welds [1].

For this thesis, recommended practices include periodic radiometric calibration checks; explicit recording of assumed emissivity and optical parameters; documentation of background and reflection control measures; and reporting SNR or contrast metrics for each processed modality (e.g., phase contrast, TSR derivatives) alongside classifier metrics.

2.13 Additional Case Studies and Modalities in Production

Thermography has been applied to a wide range of welding-related inspection tasks beyond inline seam monitoring. In automotive spot-weld inspection, combinations of induction heating and thermal camera imaging enable reliable resistance-spot-weld (RSW) assessment without surface painting; defect signatures in the thermal response are mapped to weld attributes and quality classes [9, 13, 14]. For submerged-arc and multipass welds, thermal camera monitoring has been used to capture temperature histories and validate finite-element (FE) models; despite complications from flux and insulation, off-pool temperature profiles remain informative for process characterization.

Broader surveys of thermography in welding and manufacturing consolidate best practices for seam tracking, penetration control and inline QA, and highlight typical pitfalls such as glare, emissivity uncertainty and misalignment [2, 3]. These studies collectively confirm that, with appropriate geometry, calibration and processing, thermography can support both real-time decision-making and post-process verification across welding modalities.

2.14 Machine Learning Models for Thermal Defect Detection

This section describes the three classical machine learning models implemented and evaluated for thermal welding defect detection in this thesis, based on the analysis of real thermal video data from laser welding processes.

2.14.1 Random Forest Classifier

Random Forest is an ensemble learning method that combines multiple decision trees to create a robust classifier particularly well suited for heterogeneous thermal features. Given input feature vectors $\mathbf{x}$ and labels y, Random Forest constructs N decision trees using bootstrap samples of the training data and random feature subsets at each split. The final prediction is obtained by majority voting:

$$\hat{y} = \text{mode}\{h_1(\mathbf{x}), h_2(\mathbf{x}), \ldots, h_N(\mathbf{x})\},$$

where $h_i(\mathbf{x})$ denotes the prediction of the i-th decision tree.

In this work, Random Forest is configured with a moderate number of estimators to balance performance and computational efficiency, bootstrap sampling with replacement for each tree, and the square root of the total number of features considered at each split. There is no hard maximum depth constraint so that complex interactions between thermal features (e.g., gradients, ranges, cooling-rate surrogates) can be captured. Random Forest is particularly attractive for thermal analysis because it provides feature-importance scores, is robust to outliers and does not require strong assumptions about feature distributions.

2.14.2 Support Vector Machine

Support Vector Machines (SVMs) seek a decision boundary that maximizes the margin between classes in feature space. For binary classification with labels $y_i \in \{-1, +1\}$ and feature vectors $\mathbf{x}_i$, the soft-margin SVM solves the optimization problem

$$\min_{\mathbf{w},b,\boldsymbol{\xi}} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} \xi_i$$

subject to

$$y_i(\mathbf{w}^\top \phi(\mathbf{x}_i) + b) \geq 1 - \xi_i, \quad \xi_i \geq 0,$$

where $\phi(\cdot)$ is an implicit feature map associated with a kernel function K. In this thesis, we adopt the RBF (Gaussian) kernel

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\big(-\gamma\|\mathbf{x}_i - \mathbf{x}_j\|^2\big),$$

which is well suited to non-linear decision boundaries in compact feature spaces.

The SVM is configured with a regularization parameter C that balances margin maximization and misclassification, and with γ chosen according to the `scale` heuristic based on the feature variance. Features are standardized with zero mean and unit variance prior to training. SVMs are memory efficient because only the support vectors are needed at inference, and probabilistic outputs can be obtained via calibration (e.g., Platt scaling or isotonic regression), which is important for uncertainty-aware QA decisions [4, 18].

2.14.3 Gradient Boosting Classifier

Gradient Boosting builds an ensemble of weak learners (typically shallow decision trees) in a sequential manner, where each new learner is trained to correct the residuals of the previous ensemble. Let $F_m(\mathbf{x})$ denote the ensemble after m stages; Gradient Boosting updates the model as

$$F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \gamma_m h_m(\mathbf{x}),$$

where h_m is fitted to the negative gradient of a chosen loss function $L(y, F(\mathbf{x}))$ with respect to $F_{m-1}(\mathbf{x})$. For binary classification with probabilistic outputs, a deviance (logistic) loss is typically used.

In this thesis, Gradient Boosting is configured with a moderate number of estimators, a learning rate η controlling the contribution of each stage and shallow trees (small depth) to prevent overfitting on limited thermal datasets. Gradient Boosting captures complex interactions between thermal statistics and gradients and often achieves strong performance with relatively few features, at the cost of increased training time compared to Random Forest.

2.14.4 Thermal Feature Engineering for Classical ML Models

The success of these classical ML models depends critically on effective feature extraction from the thermal video data. The feature set used in this thesis is compact but physics-informed:

- **Statistical moments**: mean, standard deviation, minimum, maximum and range of temperature over the ROI.
- **Percentile-based features**: high and low quantiles characterizing hot spots and cool areas, as well as robust ranges.
- **Spatial characteristics**: valid pixel count as a data-quality indicator, and gradient-magnitude statistics derived from Sobel operators.
- **Temporal descriptors**: simple cooling-rate proxies computed from early/late frames or from mean-temperature time series.

Preprocessing steps include temporal sampling of representative frames, spatial subsampling for computational efficiency, filtering of non-finite values and feature normalization where required. These features are designed to be stable under moderate changes in emissivity and acquisition conditions, while remaining inexpensive to compute.

2.14.5 Model Comparison and Selection Criteria

Within the classical ML family, SVM with RBF kernel and calibrated probabilities emerges as the primary classifier for this thesis. Under group-aware validation

(splitting by file to avoid leakage between training and validation sets), SVM offers stable F1 scores and good generalization in a low-data regime. Random Forest and Gradient Boosting provide complementary perspectives via feature-importance analysis and slightly different decision logic; they are used as baselines and for ablation studies.

Selection criteria for industrial deployment include interpretability, robustness to distribution shifts, real-time performance and the availability of uncertainty estimates. In this context, SVM is attractive because of its data efficiency, margin-based decision rule and compatibility with probability calibration, while tree ensembles offer intuitive feature-importance diagnostics.

2.15 Methods, Machine Models and Methodologies

The methodological choices in this thesis are informed by both thermal NDT practice and modern machine learning. At a high level, the pipeline comprises: (i) physics-aware preprocessing and normalisation of thermal sequences; (ii) extraction of compact, defect-sensitive features; (iii) training and validation of classical ML and deep learning models; and (iv) integration into a monitoring framework that respects real-time and industrial constraints.

On the physics side, moving heat-source models (e.g., Goldak-type double-ellipsoid representations) and transient heat-conduction simulations can be used to study weld thermal fields, to generate synthetic labeled data and to assess the sensitivity of thermal signatures to process parameters. These simulations motivate feature choices (e.g., cooling-rate descriptors, HAZ width) and help interpret model decisions.

On the data-driven side, classical ML models (Random Forest, Gradient Boosting, SVM) are combined with deep architectures such as CNNs and 3D CNNs for image and sequence analysis [16, 13, 17]. Where appropriate, modern DL techniques—Batch Normalization [21], residual connections [22], advanced activation functions [23, 24] and adaptive optimizers [25, 26]—are employed to stabilise training and improve generalisation.

Dataset design, cross-validation and group-aware splitting are adopted to reduce optimism and to obtain realistic performance estimates. Performance metrics include precision, recall, F1-score for classification and, where applicable, IoU and Dice for segmentation, along with computational metrics such as inference time and memory footprint. Uncertainty quantification and explainability (e.g., probability calibration, feature-importance rankings, saliency maps) are considered essential for industrial adoption and are therefore integrated into the overall methodological framework.

Chapter 3

Methodology

Introduction

Thermal camera imaging is widely used in non-destructive testing (NDT) because it offers non-contact, fast measurement of heat patterns that can correlate with material discontinuities and process instabilities [1, 2]. Recent studies report that combining thermographic inspection with machine learning—especially deep learning—can improve robustness and automation in defect detection workflows across industrial settings, including welding and related manufacturing processes [4, 5, 6].

Motivated by the need for reliable, repeatable, and computationally efficient analysis of thermal video streams, this thesis adopts a frame-level binary classification strategy in which each thermal frame is mapped to a probability of defect. The implemented pipeline (provided as a full script) indexes thermal sequences stored in HDF5 containers, constructs train/validation/test splits, applies normalisation and augmentation, trains a dedicated convolutional model, and exports both quantitative metrics and visual diagnostics for reporting and reproducibility, in line with best practice for thermal-video-based weld inspection [5, 6]. In practice, this workflow is encapsulated in the `train_thermalnet_v1.py` script, which automates data loading, training, validation, and result export.

The proposed network, ThermalNet-V1, is a lightweight 2D CNN designed to capture both fine-scale and coarse thermal signatures typical of weld quality variation. Architecturally, it combines: (i) multi-scale convolutions to aggregate features from different receptive fields (inspired by multi-branch designs and multi-scale context aggregation principles [27, 28, 29]), (ii) residual learning to stabilise optimisation and preserve low-level information across depth [22], and (iii) channel attention via Squeeze-and-Excitation (SE) to recalibrate informative feature channels while suppressing noise [30].

Concretely, ThermalNet-V1 begins with a convolutional stem and proceeds through three stages of feature extraction. Each stage uses a multi-branch block (with 3×3, 5×5, and dilated 3×3 kernels) followed by a residual SE block, while strided convolutions downsample spatial resolution to expand contextual coverage. To support defect cues that may be either diffuse (global heating trends) or localised (small hot or cold spots), the model fuses two classification views: a global branch

based on global average pooling and a local branch based on adaptively pooled feature grids. The use of global pooling is consistent with established practices for improving generalisation and reducing overfitting in CNN classifiers [29, 22].

A key methodological concern with thermal video is sample dependence: adjacent frames from the same physical run are strongly correlated. Random frame-level splitting can therefore inflate performance estimates through leakage, undermining external validity [20]. To address this, the pipeline supports file-level splitting, ensuring that all frames from a given thermal sequence remain in a single partition (train/-validation/test), aligning evaluation with deployment conditions and recommended anti-leakage practice.

Training uses standard probabilistic binary classification objectives (binary cross-entropy), optimised with Adam and weight decay [25], while a cosine-based learning-rate schedule improves convergence dynamics across epochs [31]. Model performance is summarised with ROC–AUC and precision–recall (PR) analysis, since PR curves and average precision can be especially informative when class distributions are skewed [32, 33]. Finally, an operating point (decision threshold) is selected on validation data using Youden's J (maximising the sensitivity–specificity balance) [34], and the chosen threshold is then applied to the held-out test set for confusion-matrix-based reporting.

3.1 System Architecture Overview

3.1.1 Hardware Setup

All data in this thesis are acquired with a single infrared (thermal) camera observing heating trajectories over metallic surfaces. The sensor configuration is kept fixed across the main experiments to isolate methodological effects from hardware variability. Recordings have a native spatial resolution of 240×384 pixels and an approximately constant acquisition rate of 19 fps (Section 3.2.1), which is compatible with stream-based monitoring scenarios.

3.1.2 Software Framework

The end-to-end processing workflow is implemented in Python and organised around HDF5-based `.h5` containers. The training script `Deep2D/train_thermalnet_v1.py` indexes the sequence files, constructs train/validation/test splits (optionally at file level to prevent leakage), applies preprocessing and augmentation, trains ThermalNet-V1, and exports metrics and plots into a run directory for full traceability.

3.1.3 Overview

ThermalNet-V1 is the proposed 2D convolutional model and end-to-end training pipeline used in this thesis for frame-level weld anomaly detection from thermal-camera data. Given a single thermal frame as input, the network outputs a scalar

probability of defect, enabling deployment on streaming data and allowing temporal post-processing (e.g., smoothing or clip-level aggregation) to be applied outside the model if required. This frame-level formulation is well suited to thermographic NDT, where defects can appear as localised hot/cold spots or as broader deviations in heat flow over the weld region [2, 1].

From an implementation perspective, the complete methodology is packaged in the script `train_thermalnet_v1.py`. The script discovers thermal sequences stored as HDF5 containers, constructs reproducible train/validation/test splits, applies global normalisation together with optional resizing and lightweight augmentation, trains the network with Adam and a cosine-annealing learning-rate schedule [25, 31], and finally exports plots, metrics, and per-frame predictions. Each experiment is written to a dedicated run folder under `Deep2D/runs/`, including configuration files and split metadata, so that results can be traced back to a specific dataset version and set of hyperparameters.

The model architecture is designed to match the multi-scale nature of thermal defect cues. Multi-branch convolutions capture patterns over different receptive fields (fine local structures and wider thermal context), residual connections improve optimisation stability, and Squeeze-and-Excitation (SE) attention reweights channels to emphasise the most informative feature maps [27, 28, 22, 30]. To preserve both global context and localised evidence, ThermalNet-V1 uses two complementary classification heads (global and local) whose outputs are combined to produce the final defect probability.

3.2 Data Collection and Dataset Preparation

This section summarises how the raw thermal recordings are structured, labelled, and converted into frame-level training samples.

3.2.1 Thermal Dataset Description

3.2.1.1 Experimental Thermal Database

3.2.1.1.1 Introduction The experimental thermal database used in this study consists of two classes, *Normal* and *Defect*, stored under the directories `VideoData/Normal` and `VideoData/Defect`, respectively. All recordings are saved as `.h5` files, which are HDF5 containers designed to store synchronized thermal and temporal information.

Each `.h5` file exposes three datasets:

- `tempdata`: raw temperature maps stored as 16-bit floating-point values with shape $(N, 240, 384)$,
- `timeframe`: ISO-formatted timestamps corresponding to each frame,
- `reconstruction_error`: a scalar error value per frame produced by an autoencoder-based reconstruction process.

File names follow the convention

`<setup>_<site>_<dd>_<mm>_<yyyy>_<HH>_<MM>_<SS>_<target_frames>.h5`

encoding the sensor identifier, location tag, acquisition time, and the target number of frames. The effective frame rate is computed as:

$$\text{FPS} = \frac{N-1}{t_{\text{end}} - t_{\text{start}}}.$$

3.2.1.1.2 Normal Class The *Normal* class consists of 47 thermal sequences (approximately 3.17 GB) recorded consecutively between 12:54 and 15:19 on 13 November 2025. In total, this class contains 31,376 frames corresponding to 1,637 s of footage (approximately 27.3 minutes). Individual recordings range from 530 to 865 frames, with durations between 28 s and 45 s.

All sequences share a fixed spatial resolution of 240×384 pixels. The acquisition frame rate is highly stable, ranging from 18.7 to 19.6 fps with a mean of 19.13 fps. Throughout this thesis, the data are therefore described as being acquired at 19 fps with less than ± 0.5 fps drift.

Pixel-wise temperature values range from 174.9 to 1,525.0 arbitrary thermal units, with a global mean of 270.5. File-level mean temperatures vary between 253 and 296, reflecting gradual background temperature drift rather than abrupt anomalies.

Reconstruction errors remain consistently low, with values ranging from 4.6×10^{-5} to 3.4×10^{-2} and a global mean of 9.8×10^{-3}. This low variance supports the use of the normal subset as a baseline for anomaly detection training. Overall, the normal class provides approximately 2.9 billion labeled thermal pixels.

3.2.1.1.3 Defect Class The *Defect* class contains six sequences (approximately 105 MB), totaling 1,940 frames and 98.8 s of footage. Five recordings were acquired on 13 November 2025, while two additional test sequences were captured on 14 November 2025.

Typical defect recordings contain 201–226 frames (approximately 10–11 s) recorded at 19.1–20.2 fps, resulting in shorter temporal coverage compared to normal sequences.

Mean temperature values for four defect sequences range between 182 and 200 units, indicating cooler surfaces or occlusions.

Reconstruction errors provide strong class separation. Four defect sequences exhibit mean errors between 0.19 and 0.22 with peaks up to 0.38, approximately 20 times higher than the normal baseline. Two sequences show subtler deviations, with mean errors around 0.005–0.006, corresponding to incipient faults.

3.2.1.1.4 Comparative Observations Approximately 94% of all frames belong to the normal class, resulting in a significant class imbalance. This motivates the use of anomaly detection approaches rather than conventional supervised classification.

Both classes share identical spatial resolution and timestamp structure, enabling unified preprocessing and multimodal feature design. The near-constant frame rate and precise timestamps allow temporal measurements to be mapped to physical distances when the inspection speed is known. Duplicate or ambiguous recordings are explicitly documented and handled during training to ensure transparency.

3.2.1.2 Linear Movement Dataset

A second acquisition campaign focuses on straight-line weld trajectories, which eliminate the periodic heating and cooling cycles created by the zigzag scan while still stressing the sensing pipeline with varying emissivity and shielding conditions [35, 5]. The recordings are stored under the `Linear/` workspace and processed by the dedicated run `Linear/thermalnet\_Linear`. This dataset comprises 15 HDF5 sequences (eight Normal and seven Defect files) for a total of 32,617 frames. File names follow the same convention as the zigzag data but include the suffix `_l` and, for trimmed clips, the explicit frame range (e.g. `frames_030_300`). Each clip therefore contains approximately 270–300 frames captured at 18–20 fps, representing the thermal evolution along a linear bead.

Because the welding-torch translation is monotonic, the resulting temperature profiles feature smoother ramps and fewer oscillations than the zigzag data; defect signatures manifest as sudden spikes or prolonged plateaus rather than repeated peaks. The dataset is still highly imbalanced in terms of frames (Table 3.5), but the spatial statistics differ markedly: the global mean intensity drops to $\mu = 95.07$ with a standard deviation of $\sigma = 127.29$, reflecting lower background temperatures and a broader spread of cold pixels due to the larger field of view. These statistics are stored in `Linear/thermalnet_Linear/metrics/normalization.json` and are applied consistently across training, validation, and test splits. Qualitative comparisons between zigzag and linear trajectories are reported in Chapter 4, where both movement patterns are evaluated within the same modelling framework.

3.2.1.3 Dataset 1 (zigzag movement)

3.2.1.3.1 Introduction In addition to the in-house recordings, an external thermal video database referred to as *Dataset 1* is used to assess how well the proposed pipeline generalises to a different industrial setting. The data were captured on a production welding line using a camera with comparable spatial resolution and exported to `.h5` containers following the same structure as the experimental database (`tempdata`, `timeframe`, and optional auxiliary signals). All sequences are pre-processed with the same sanitisation, resizing, and normalisation steps to ensure that differences in performance can be attributed to domain shift rather than implementation details.

3.2.1.3.2 Normal Class The normal subset of Dataset 1 contains thermal sequences where the welding process is considered stable by process engineers, with

no visually apparent anomalies in bead shape or shielding. Temperature fields show smooth evolution along the weld, and frame-wise intensities remain within narrow bounds after global normalisation. These sequences are used primarily to confirm that the model does not over-trigger on benign variability specific to this external hardware and surface conditions.

3.2.1.3.3 Defect Class Defect-labelled sequences from Dataset 1 include examples of underfill, lack of fusion, and local overheating identified through downstream inspection. Compared to the experimental database, these defects can be more subtle, as they arise under realistic production variability rather than controlled laboratory faults. The corresponding thermal signatures typically exhibit localised hot or cold regions, disrupted heat-flow lines, or elevated reconstruction errors when passed through the autoencoder baseline, providing a complementary test bed for the classifier.

3.2.1.3.4 Movement patterns (zigzag) Dataset 1 focuses on near-circular zigzag trajectories, emphasising repeated heating and cooling cycles over a compact field of view. The camera geometry and storage format remain compatible with the in-house zigzag data, so that the same preprocessing, normalisation, and file-level splitting strategy can be applied. This allows the effect of domain shift (different hardware, fixtures, and surface conditions) to be studied specifically for zigzag motion patterns.

3.2.1.4 Dataset 2 (linear movement)

3.2.1.4.1 Introduction A second external database, referred to as *Dataset 2*, contains thermal videos from a different welding platform and alloy system. Camera calibration, emissivity settings, and background conditions differ from both the experimental data and Dataset 1, leading to shifted intensity histograms and distinct noise characteristics. To maintain compatibility with the rest of the pipeline, the raw recordings are converted into `.h5` files with the same key layout and are processed with the same frame-level indexing strategy.

3.2.1.4.2 Normal Class Normal sequences from Dataset 2 cover a broad range of operating points (weld speed, power, and shielding) that are deemed acceptable in that production environment. From a thermal perspective, these runs often display more pronounced global drifts and reflections than the in-house data, but still follow a consistent spatio-temporal pattern along the weld. Incorporating these sequences during evaluation helps test the robustness of the learned normal manifold under realistic cross-site variation.

3.2.1.4.3 Defect Class The defect subset collected in Dataset 2 focuses on more severe process deviations, such as missing shielding gas, misaligned torch trajectories, or intentional parameter excursions. These faults tend to generate

strong local hotspots, extended cold bands, or irregular bead footprints that depart markedly from the typical temperature evolution in the corresponding normal runs. Evaluating the model on this database offers a stringent test of its ability to flag out-of-distribution behaviour while preserving high specificity on normal operation across all three thermal datasets.

3.2.1.4.4 Movement patterns (linear) Dataset 2 focuses on linear motion along the weld seam. Due to differences in mechanical guidance and fixture design, the resulting thermal signatures exhibit distinct drift characteristics compared to the experimental zigzag data and Dataset 1. By processing these recordings with the same HDF5 schema and frame-level indexing, the pipeline can assess whether a single ThermalNet-V1 configuration remains reliable across multiple sites for linear campaigns, or whether re-calibration or fine-tuning is required when transitioning between linear runs at different industrial partners.

Table 3.1: Overview of thermal datasets used in this thesis.

Name	Motion pattern	Origin	Role in experiments
Experimental zigzag	Circular zigzag	In-house laboratory	Canonical zigzag dataset used to develop the full pipeline and qualitative prediction plots.
Experimental linear	Linear	In-house laboratory	Canonical linear dataset used to test generalisation of the same model configuration to linear motion.
Dataset 1	Near-circular zigzag	External production line	External zigzag database (compatible HDF5 layout) for evaluating domain shift under similar motion patterns.
Dataset 2	Linear	External production line	External linear database used to assess robustness across sites, alloys, and acquisition conditions.

3.2.2 Data Storage, Discovery, and Indexing

The training data is organised as two folders, `Normal/` and `Defect/`, located under a user-specified `data_root`. Each sample is a complete thermal sequence stored as an HDF5 file, and the script supports `.h5` containers. Within each file, the expected structure is a 3D dataset of shape (T, H, W), where T is the number of frames and (H, W) is the spatial resolution. In the implementation, the dataset loader first checks for a standard key (`tempdata`); if it is not present, it falls back to the first 3D dataset found in the file, which makes the pipeline robust to small variations in acquisition/export tooling.

To enable frame-level training, the script converts each sequence into a flat list of indexed frames. During dataset initialisation, all HDF5 files in both classes are scanned

to determine the number of frames, and an index is built as tuples `(file_path, dataset_key, frame_idx, label)`. An optional parameter `max_frames_per_file` can cap the number of frames taken from each sequence, which is useful when sequences are long and adjacent frames are highly redundant.

After indexing, the learning problem becomes binary classification on a set of individual frames: each sample is a single-channel image $\mathbf{x}_i \in \mathbb{R}^{H \times W}$ with a label $y_i \in \{0, 1\}$, where $y_i = 0$ denotes a Normal frame and $y_i = 1$ denotes a Defect frame. In addition to returning $(\mathbf{x}_i, y_i)$ for training, the dataset returns the source filename and frame index; this metadata is later used to export per-frame prediction tables and to generate diagnostic plots that link predicted probabilities back to original thermal sequences.

3.3 Image Preprocessing Pipeline

3.3.1 Pre-processing, Resizing, Normalisation, and Augmentation

Thermal sequences may have high spatial resolution, which can increase memory consumption during training. For this reason, the pipeline optionally resizes frames to a fixed target size (controlled by `-resize-height` and `-resize-width`) using bilinear interpolation. Standardising the spatial resolution keeps batch shapes consistent for efficient GPU/MPS execution, and allows experiments to be run under fixed memory constraints while keeping the model architecture unchanged. When full-resolution processing is required, resizing can be disabled via `-no-resize`.

In all experiments, raw thermal arrays are loaded as `float32` and sanitised to ensure numerical stability. Missing or non-finite sensor values (NaNs, $\pm\infty$) are replaced by zero via `nan_to_num`. For `.h5` containers, the thermal frames are read from the `tempdata` dataset; for legacy `.h5` files, the loader falls back to the first 3D dataset (T, H, W) if `tempdata` is not available. Each training sample is therefore a single-channel tensor of shape $(1, H, W)$ plus metadata (`file`, `frame_idx`) that is preserved for post-hoc analysis and traceability.

Resizing is applied on-the-fly inside the dataset `__getitem__` using bilinear interpolation (`align_corners=False`). This step standardises the spatial resolution to 128×192 pixels in the canonical runs, reducing memory usage and ensuring that the same network can be trained across datasets with identical tensor shapes.

To reduce sensitivity to absolute temperature offsets across runs and to stabilise optimisation, the script applies global normalisation computed on the training split. Concretely, it estimates a single global mean μ and standard deviation σ over all pixels in the training frames using a streaming accumulator (so the full dataset does not need to be loaded at once), and then normalises each frame as:

$$\widehat{\mathbf{x}} = \frac{\mathbf{x} - \mu}{\sigma},$$

with a small ϵ added in the implementation to avoid division-by-zero. The resulting (μ, σ) are stored in `normalization.json` and applied consistently to training, vali-

dation, and test frames; computing statistics only on the training partition avoids leakage from evaluation data [20].

To improve robustness and reduce overfitting, lightweight augmentation is applied to the training split only. The implemented augmentations include random horizontal and vertical flips and low-amplitude additive Gaussian noise ($\sigma = 0.01$). In the current implementation, the noise-augmented tensor is clipped to $[0, 1]$ to limit extreme values after perturbation. Validation and test frames are not augmented, so that reported performance reflects behaviour on unaltered data.

3.3.2 Train/Validation/Test Splitting and Leakage Prevention

Thermal video data exhibits strong temporal and file-level correlations: consecutive frames in the same run are often nearly identical, and entire sequences share acquisition conditions (emissivity, reflections, camera settings, and background heating) [35, 1]. If frames from the same sequence are split across training and evaluation sets, the model can exploit sequence-specific patterns and yield overly optimistic performance estimates, a form of data leakage [20].

To mitigate leakage, the default and recommended configuration uses file-level splitting. The script partitions the set of sequence files into disjoint train/validation/test splits and then expands each file split to the corresponding set of indexed frames. Importantly, Normal and Defect file lists are shuffled and split separately, so that each partition contains both classes even when the dataset is small; the resulting file lists are saved to `file_splits.json` for reproducibility.

For comparison, the code also provides an optional frame-level split (`-split-level frame`) that randomly partitions individual frames regardless of their source file. This setting reproduces the common but leaky baseline where frames from the same sequence may appear in multiple splits. In this thesis, metrics are reported using file-level splitting to better match deployment conditions, where the model is expected to generalise to entirely unseen sequences.

3.4 Feature Engineering

ThermalNet-V1 is trained end-to-end on normalised temperature frames, and therefore learns task-specific features directly from data. However, the dataset also provides auxiliary scalar signals that support interpretable baselines and sanity checks. In particular, the per-frame `reconstruction_error` (Section 3.2.1) can be thresholded as a simple anomaly score, and it provides a useful diagnostic for understanding how strongly the Normal and Defect recordings differ before introducing a learned classifier. More generally, classical feature vectors can be formed from per-frame temperature statistics (e.g., mean, standard deviation, percentiles), gradient-energy measures, or simple texture descriptors computed on the normalised thermal map; these handcrafted features can then be paired with standard classifiers (Chapter 2) to establish lightweight reference baselines alongside the deep model.

3.5 Machine Learning Model Development

3.5.1 ThermalNet-V1 Architecture

ThermalNet-V1 is a compact convolutional encoder designed for single-channel thermal frames. In the canonical configuration, the network receives a resized input of 128×192 pixels and produces a single defect probability. The implementation contains 3,002,618 trainable parameters, which is small enough for real-time inference on commodity GPUs while still allowing multi-scale feature extraction.

The network begins with a convolutional stem (`Conv2d(1→32)`), followed by batch normalisation [21] and a SiLU activation [23, 24]. It then processes features through three stages, each consisting of a multi-scale convolution block and a residual SE block; strided convolutions between stages reduce spatial resolution to increase receptive field while keeping computation manageable. Table 3.2 summarises the main feature-map sizes in the forward pass for the canonical input resolution.

Table 3.2: Approximate feature-map sizes for ThermalNet-V1 with input 128×192.

Block	Channels	Height	Width
Input (resized)	1	128	192
Stem	32	128	192
Stage 1 output	64	128	192
Downsample 1	64	64	96
Stage 2 output	128	64	96
Downsample 2	128	32	48
Stage 3 output	256	32	48
Global pooling	256	1	1
Local pooling	256	4	4

Table 3.3: High-level summary of the ThermalNet-V1 architecture.

Property	Value
Input size	$1 \times 128 \times 192$ (channel $\times$ height $\times$ width)
Trainable parameters	3,002,618
Encoder stages	3 (multi-scale block + residual SE block per stage)
Multi-scale branches	3×3, 5×5, and dilated 3×3 (dilation 2)
Attention	Squeeze-and-Excitation channel attention in each residual block
Normalisation	Batch Normalisation after convolutions
Nonlinearity	SiLU activation in all convolutional blocks
Output heads	Global GAP MLP head and local 4×4 pooled MLP head (combined with a 0.6/0.4 weighting)

The multi-scale component (`MultiScaleBlock`) uses three parallel branches: a 3×3 convolution, a 5×5 convolution, and a dilated 3×3 convolution (dilation 2). Each branch is followed by batch normalisation and a SiLU activation, and the branch outputs are concatenated along the channel dimension and fused with a

1×1 convolution. This design follows the general idea that multi-branch and dilated convolutions can efficiently capture context at different spatial scales [27, 28].

The residual attention component (`ResidualSEBlock`) applies two 3×3 convolutions with batch normalisation and SiLU and then recalibrates channels via SE attention [30]. A skip connection adds the block input to the transformed output, following residual learning principles [22]. In the SE mechanism, the feature maps are globally pooled to obtain a channel descriptor; this descriptor is passed through a small gating network and a sigmoid to obtain channel weights that rescale the feature maps, amplifying informative channels while suppressing less relevant responses.

3.5.2 Global and Local Classification Heads

After the final encoder stage, the network produces a feature tensor with 256 channels. The global head extracts a global descriptor via global average pooling (GAP) and maps it to a logit with a small MLP (`256→128→1`). GAP is widely used to improve generalisation and reduce overfitting by enforcing a global, translation-tolerant summarisation [29].

In parallel, the local head preserves a coarse spatial layout by adaptive pooling to a 4×4 grid, flattening, and applying a second MLP (`4096→128→1`). This branch helps retain sensitivity to localised hotspots or cold regions that might be diluted by global pooling.

The final prediction fuses global and local evidence by a weighted combination of logits (0.6 for the global head and 0.4 for the local head) followed by a sigmoid, returning $p \in [0, 1]$ as the defect probability for each frame.

3.5.3 Training Objective and Optimisation

Training is formulated as probabilistic binary classification at the frame level. For a minibatch of predictions p_i and labels y_i, the objective is the binary cross-entropy (BCE):

$$\mathcal{L}_{\mathrm{BCE}} = -\frac{1}{N} \sum_{i=1}^{N} \left[y_i \log(p_i) + (1 - y_i) \log(1 - p_i) \right], \tag{3.1}$$

which corresponds directly to the `BCELoss` used in the implementation. Optimisation is performed with Adam [25] and weight decay, which stabilises training and discourages overly large weights in a low-data regime.

The training loop follows a standard stochastic-gradient procedure. For each mini-batch, frames are forwarded through ThermalNet-V1 to obtain probabilities $p \in [0, 1]$; BCE is computed; gradients are backpropagated; and parameters are updated with Adam. To improve convergence behaviour across epochs without manual learning-rate tuning, the script uses cosine annealing [31], which smoothly reduces the learning rate over epochs following a cosine profile.

In addition to the optimisation loss, the pipeline records ranking and operating-point metrics at the end of each epoch. Specifically, it aggregates all validation probabilities and labels to compute ROC–AUC and average precision (AP), and it

reports accuracy at the fixed default threshold $\tau = 0.5$. This combination is important in the presence of severe class imbalance, where accuracy can be misleading but ROC–AUC and PR analysis remain informative [32, 33]. All epoch-wise values are persisted to `metrics/history.csv` to enable learning-curve plotting and later auditing.

Model selection is performed by choosing the epoch with maximum validation ROC–AUC. The corresponding state dict is saved to `thermalnet_v1_best.pth` together with the epoch number and the validation score. For reproducibility, each run stores the full configuration (`config.json`), the training curves (`history.csv`), the file-level splits (`file_splits.json` when enabled), and the normalisation statistics (`normalization.json`).

3.6 Real-time Implementation

The workflow is designed with deployment constraints in mind. First, the model operates at the frame level and can be applied to a live stream without requiring a fixed sequence length; temporal smoothing or clip-level aggregation can be performed as a lightweight post-processing step when needed. Second, the canonical input resizing to 128×192 and the compact network size (approximately 3.0M parameters) reduce memory footprint and enable low-latency inference on commodity GPUs.

For industrial deployment, the practical pipeline is therefore: acquire a thermal frame, apply the same normalisation parameters (μ, σ) estimated on the training split, run a single forward pass to obtain $p(\text{defect})$, and then apply the calibrated threshold τ^* (or an application-specific threshold) to trigger alerts.

3.7 Evaluation Methodology

3.7.1 Validation Metrics and Threshold Calibration

After training, the best checkpoint (selected by validation ROC–AUC) is reloaded and evaluated once on the validation and test sets. Since the model outputs probabilities, evaluation includes threshold-independent ranking metrics such as ROC–AUC and average precision (AP) [32, 33]. These metrics quantify how well the model ranks defect frames above normal frames regardless of a specific decision threshold.

For deployment, a concrete decision rule is required. The script selects a probability threshold on validation data by maximising Youden's J statistic:

$$J(\tau) = \mathrm{TPR}(\tau) - \mathrm{FPR}(\tau), \qquad \tau^* = \arg\max_{\tau} J(\tau), \tag{3.2}$$

where $\mathrm{TPR}(\tau)$ and $\mathrm{FPR}(\tau)$ are the true-positive and false-positive rates at threshold τ [34]. The selected τ^* is stored and then applied unchanged to the held-out test set to obtain an unbiased estimate of operating-point performance.

With the threshold fixed, the script reports confusion-matrix-based metrics on both validation and test sets, including accuracy, precision, recall, and F1-score.

Detailed outputs are saved to `val_metrics.json` and `test_metrics.json`, together with the ROC curve values and the maximum Youden score. This makes it possible to reproduce reported numbers exactly and to inspect the sensitivity–specificity trade-off implied by the chosen operating point.

3.7.2 Frame-wise Inference and Visualisation

Beyond aggregate metrics, the pipeline exports frame-wise predictions to support detailed analysis. For a sequence with frames $\{\mathbf{x}_t\}_{t=1}^{T}$, inference produces a probability time series where each frame is mapped to a defect probability and then thresholded with the calibrated τ^*. This representation supports downstream temporal inspection, such as identifying contiguous segments with persistently high defect probability.

For the held-out test set, the script iterates over every test frame and writes a CSV table containing the source file, frame index, true label, predicted probability, and predicted label after thresholding (`test_frame_predictions.csv`). This output can be used directly to construct result tables, compute additional statistics, or trace specific error cases back to the original sequences.

Qualitative diagnostics complement numerical performance. The pipeline saves example montages of validation and test frames with predicted probabilities, and it produces a temporal probability plot for one test sequence (probability vs. frame index) with the decision threshold overlaid. These visualisations help confirm that high-probability predictions correspond to physically plausible thermographic signatures rather than artefacts of preprocessing or background conditions [2, 1].

3.7.3 Post-processing, Thresholding, and Reproducible Outputs

For deployment-style reporting, the probability scores must be converted into hard decisions. After training, the best checkpoint (selected by validation ROC–AUC) is evaluated on the validation set and the ROC curve is computed. A single operating threshold τ^* is selected by maximising Youden's $J(\tau) = \mathrm{TPR}(\tau) - \mathrm{FPR}(\tau)$ [34]. This calibrated τ^* is then applied unchanged to the held-out test set to obtain the confusion matrix and threshold-dependent metrics (precision, recall, F1), ensuring an unbiased estimate at the chosen operating point.

To make the experiments traceable and easy to reproduce, each run directory contains a complete record of the configuration, splits, and results:

- **Configuration:** `config.json` (hyperparameters, resize settings, split ratios, device).
- **Splits and preprocessing:** `metrics/file_splits.json` and `metrics/normalization.json`.
- **Training history:** `metrics/history.csv` and `plots/training_curves.png`.
- **Evaluation:** `metrics/val_metrics.json` and `metrics/test_metrics.json` (including τ^* and confusion matrices).

- **Predictions:** `predictions/test_frame_predictions.csv` (per-frame probabilities and labels).
- **Qualitative plots:** `plots/val_roc_pr_curves.png`, `plots/val_example_frames.png`, `plots/test_example_frames.png`, and `plots/test_clip_probs_*.png`.

3.7.4 Hyperparameter Selection and Tuning

The current pipeline does not perform automated hyperparameter optimisation (e.g., grid search or Bayesian optimisation). Instead, a single training recipe is used consistently across both movement patterns to isolate the effect of data distribution and motion dynamics on performance. Core choices (Adam, weight decay, cosine annealing, file-level splitting) follow standard deep-learning practice and anti-leakage recommendations [25, 31, 20]. The shared configuration also simplifies comparison in Chapter 4; future work could extend the script with systematic tuning (learning-rate sweeps, class-weighted losses, focal loss, or temperature calibration) without changing the overall evaluation protocol.

3.8 Reference Training Configuration and Hyperparameters

To keep the evaluation reproducible, two canonical experiments are maintained: (i) the circular zigzag dataset processed by `Deep2D/runs/thermalnet\_v1\_20251214\_213251`, and (ii) the linear scanning dataset trained in `Linear/thermalnet\_v1\_20251222\_224436`. Both follow the methodology detailed earlier but differ in their data splits and normalisation statistics.

3.8.1 Data Splits and Normalisation

3.8.1.0.1 Circular zigzag movement This campaign contains 24 HDF5 files (7,476 frames) and is split strictly by file to mitigate leakage [20]. Table 3.4 summarises the partition sizes. Normal frames account for 93.3% of the training data, so class imbalance must be handled through ranking metrics and careful threshold calibration. Frames are resized to 128×192 pixels before batching, and global statistics computed on the training set yield $\mu = 239.35$ and $\sigma = 172.76$, which are stored in `Deep2D/runs/thermalnet_v1_20251214_213251/metrics/normalization.json`.

Table 3.4: File- and frame-level statistics for the circular zigzag run.

Split	Normal files	Defect files	Frames
Train	11	5	5,401
Validation	3	1	878
Test	3	1	1,197

3.8.1.0.2 Linear movement The linear campaign contains 15 files (32,617 frames) with a more even split between Normal and Defect trajectories. File-level partitioning is again used (Table 3.5) so that each temporal sequence belongs exclusively to one split. The lower radiance levels and larger cool regions observed in the linear setup require separate normalisation; the training frames yield $\mu = 95.07$ and $\sigma = 127.29$.
These parameters are saved in
`Linear/thermalnet\_v1\_20251222\_224436/metrics/normalization.json` and are reused for validation and test inference so that Chapter 4 can compare both motion patterns on equal footing.

Table 3.5: File- and frame-level statistics for the linear movement run.

Split	Normal files	Defect files	Frames
Train	6	5	23,755
Validation	1	1	4,491
Test	1	1	4,371

In both campaigns, lightweight augmentation (horizontal/vertical flips and additive Gaussian noise with $\sigma = 0.01$) is applied only to the training partition to emulate viewpoint jitter and sensor noise [5]. Validation and test frames remain untouched so that reported metrics reflect deployment conditions.

3.8.2 Training Logs and Checkpoint Selection

The console output of `train_thermalnet_v1.py` reports the indexed frame count, the split sizes, device selection, and the selected best epoch. In the zigzag experiment, the script is executed on CUDA with `data_root=Deep2D/VideoData`, indexing 7,476 frames; global normalisation is computed on the 5,401 training frames; the best validation ROC–AUC is achieved at epoch 9 (AUC = 0.9752); and the calibrated operating point is $\tau^* = 0.4056$. In the linear experiment, the same code is executed on CUDA in a Colab environment (with `data_root` pointing to Google Drive), indexing 32,617 frames; normalisation is computed on 23,755 training frames; the best validation ROC–AUC occurs at epoch 7 (AUC = 0.9288); and the calibrated operating point is $\tau^* = 0.9877$. Table 3.6 summarises these run-level values so that the methodology chapter contains a complete, dataset-specific training record.

Table 3.6: Summary of the two canonical ThermalNet-V1 training runs.

Quantity	Circular zigzag	Linear movement
Run	`Zigzag Movement`	`Linear Movement`
Total frames indexed	7,476	32,617
Train / Val / Test frames	5,401 / 878 / 1,197	23,755 / 4,491 / 4,371
Global mean μ	239.35	95.07
Global std σ	172.76	127.29
Best epoch (val ROC–AUC)	9 (0.9752)	7 (0.9288)
Calibrated threshold τ^*	0.4056	0.9877

3.8.3 Optimisation Hyperparameters

The optimisation settings are kept identical across both runs. Table 3.7 lists the common configuration exported in each `config.json`. A 10-epoch cosine-annealed schedule with Adam and weight decay provides stable convergence for the moderately sized datasets while avoiding overfitting to individual sequences [25, 31].

Table 3.7: Shared ThermalNet-V1 hyperparameters for both canonical runs.

Parameter	Value
Batch size	8 frames
Epochs	10
Optimiser	Adam (lr $= 10^{-3}$, weight decay $= 10^{-4}$)
Scheduler	CosineAnnealingLR ($T_{\max} = 10$)
Loss	Binary cross-entropy
Split ratios	60% train / 20% validation / 20% test (file-level)
Resize resolution	128×192 pixels (bilinear)
Augmentation	Horizontal/vertical flips, Gaussian noise ($\sigma = 0.01$)
Random seed	42 (applied to NumPy, PyTorch, and the Python RNG)

All artefacts produced during training (configurations, split metadata, normalisation parameters, metrics, and plots) are stored inside each run directory, ensuring that the quantitative analyses in Chapter 4 can be traced back to their exact experimental settings.

Chapter 4

Evaluation Metrics

Introduction

This chapter reports the evaluation protocol and metrics used to assess the proposed ThermalNet-V1 classifier. All quantitative results are produced by the training script `Deep2D/train_thermalnet_v1.py`, which (i) trains on frame-level thermal data, (ii) selects the best checkpoint based on validation ROC–AUC, (iii) calibrates a decision threshold on the validation set, and (iv) evaluates the final model on the held-out test set while exporting both numerical summaries and per-frame predictions.

The choice of metrics is guided by the properties of thermal weld data and by recommendations in the machine-learning and diagnostic-testing literature. In particular, the dataset is highly imbalanced (defect frames are rare compared to normal frames) and frames from the same sequence are strongly correlated. Under such conditions, threshold-independent ranking metrics such as the ROC curve, ROC–AUC and the precision–recall (PR) curve provide a more reliable summary of discriminative ability than accuracy alone [32, 33]. Threshold-dependent figures of merit (accuracy, precision, recall, F1-score) are still important for deployment, but must be interpreted in the context of an explicitly chosen operating point and class balance.

4.1 Experimental Setup and Training Configuration

Quantitative results in this chapter are reported for two canonical experiments: (i) the circular zigzag dataset run `thermalnet_v1_20251214_213251` stored under `Deep2D/runs/`, and (ii) the linear movement run `thermalnet_v1_20251222_224436` stored under `Linear/`. The full preprocessing pipeline, data splits, and hyperparameters for both runs are documented in Section 3.8; only a brief recap is provided here to keep the chapter self-contained. In summary, both runs use file-level splitting (60%/20%/20%), resize frames to 128×192 pixels, compute z-score normalisation exclusively on the training subset, and train ThermalNet-V1 for 10 epochs with Adam ($\text{lr} = 10^{-3}$, weight decay 10^{-4}) under a cosine-annealing schedule. Lightweight flips-plus-noise augmentation is applied to training frames only. All artefacts referenced in this chapter—history files, ROC/PR plots, per-frame CSVs, and checkpoint weights—

are exported automatically by `train_thermalnet_v1.py` inside the corresponding run directory.

4.2 Evaluation Protocol

Each sample corresponds to a single thermal frame $\mathbf{x}_i$ with a binary label $y_i \in \{0, 1\}$, where $y_i = 1$ denotes a defect and $y_i = 0$ denotes a normal frame. The model outputs a probability $p_i \in [0, 1]$, interpreted as $p_i = P(y_i = 1 \mid \mathbf{x}_i)$. This frame-wise formulation follows common practice in thermographic NDT, where each image can contain localised hot or cold regions associated with potential defects [1, 2]. A hard class decision is obtained by thresholding:

$$\hat{y}_i(\tau) = \mathbb{1}[p_i \geq \tau], \tag{4.1}$$

where τ is the decision threshold.

During training, the pipeline computes epoch-level metrics on both training and validation sets and keeps the checkpoint that maximises validation ROC–AUC. This strategy decouples model selection from any particular decision threshold, in line with standard recommendations for imbalanced problems where ranking quality is more robust than accuracy at a fixed cut-off [32, 33]. After training, the best checkpoint is re-evaluated on validation and test sets; the decision threshold τ^* is calibrated on the validation set and then applied unchanged on the test set to produce operating-point metrics (confusion-matrix-based reporting).

4.3 Evaluation Metrics

4.3.1 Binary Cross-Entropy Loss (Optimisation Objective)

Training is formulated as probabilistic binary classification and optimised with the binary cross-entropy (BCE), which corresponds to maximising the log-likelihood of the observed labels under a Bernoulli model with parameter p_i. For N samples:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N} \sum_{i=1}^{N} \left[y_i \log(p_i) + (1 - y_i) \log(1 - p_i) \right]. \tag{4.2}$$

In the implementation, this corresponds to `nn.BCELoss()` applied directly to the model probabilities p_i. BCE is convex with respect to each individual prediction and strongly penalises confident misclassifications, which encourages well-calibrated probabilities and is widely used in medical-diagnostic and anomaly-detection settings where probabilistic outputs are required.

4.3.2 Threshold-Independent Ranking Metrics

Because the model outputs probabilities, it is evaluated with threshold-independent ranking metrics that quantify how well defect frames are ranked above normal frames.

These metrics consider the entire range of possible thresholds and are therefore insensitive to the particular operating point chosen later for deployment.

ROC Curve and ROC–AUC

For a threshold τ, define the true-positive rate (TPR) and false-positive rate (FPR) as:

$$\mathrm{TPR}(\tau) = \frac{\mathrm{TP}(\tau)}{\mathrm{TP}(\tau) + \mathrm{FN}(\tau)}, \qquad \mathrm{FPR}(\tau) = \frac{\mathrm{FP}(\tau)}{\mathrm{FP}(\tau) + \mathrm{TN}(\tau)}. \tag{4.3}$$

The ROC curve is the parametric curve $(\mathrm{FPR}(\tau), \mathrm{TPR}(\tau))$ as τ varies, and the ROC–AUC is the area under this curve:

$$\mathrm{AUC} = \int_0^1 \mathrm{TPR}(\mathrm{FPR})\, d\,\mathrm{FPR}. \tag{4.4}$$

In the training pipeline, validation ROC–AUC is used for model selection; the checkpoint with the highest validation AUC is saved as the best model. ROC analysis has a long history in signal detection and machine learning [32]; in this context it measures the probability that a randomly chosen defect frame is assigned a higher score than a randomly chosen normal frame. This interpretation is particularly meaningful for thermal weld inspection, where avoiding overlap between defect and normal score distributions is crucial.

Precision–Recall Curve and Average Precision (AP)

For a threshold τ, precision and recall are:

$$\mathrm{Precision}(\tau) = \frac{\mathrm{TP}(\tau)}{\mathrm{TP}(\tau) + \mathrm{FP}(\tau)}, \qquad \mathrm{Recall}(\tau) = \frac{\mathrm{TP}(\tau)}{\mathrm{TP}(\tau) + \mathrm{FN}(\tau)}. \tag{4.5}$$

The precision–recall (PR) curve is $(\mathrm{Recall}(\tau), \mathrm{Precision}(\tau))$ as τ varies. Unlike ROC curves, PR curves focus only on the positive (defect) class and are more informative when the negative class is much more prevalent [33]. The pipeline summarises PR behaviour using average precision (AP), i.e., the area under the PR curve in a step-wise sense:

$$\mathrm{AP} = \sum_k (R_k - R_{k-1})\, P_k, \tag{4.6}$$

where (P_k, R_k) are points on the PR curve corresponding to decreasing thresholds.

4.3.3 Threshold-Dependent Metrics (Operating Point)

Accuracy (Monitoring at $\tau = 0.5$)

During each epoch, the script reports accuracy at the fixed default threshold $\tau = 0.5$:

$$\mathrm{Accuracy}(\tau) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}[\hat{y}_i(\tau) = y_i] = \frac{\mathrm{TP}(\tau) + \mathrm{TN}(\tau)}{N}. \tag{4.7}$$

This metric is used for monitoring and learning-curve visualisation and is not used for checkpoint selection. Because of the class imbalance present in the thermal dataset, high accuracy can sometimes be achieved by trivial strategies (e.g., predicting all frames as normal); for this reason, accuracy is complemented with the more discriminative ROC–AUC and AP scores.

Confusion Matrix, Precision, Recall, and F1

Given a calibrated threshold τ^*, define the confusion matrix:

$$\begin{pmatrix} \text{TN} & \text{FP} \\ \text{FN} & \text{TP} \end{pmatrix}. \tag{4.8}$$

From these counts, accuracy, precision, and recall follow as above, and the F1-score is computed as the harmonic mean of precision and recall:

$$\text{F1} = \frac{2 \cdot \text{Precision}(\tau^*) \cdot \text{Recall}(\tau^*)}{\text{Precision}(\tau^*) + \text{Recall}(\tau^*)}. \tag{4.9}$$

4.4 Threshold Calibration on Validation Data

To obtain a concrete operating point for deployment-style reporting, the pipeline selects a threshold on the validation set using Youden's J statistic computed from the ROC curve [34]:

$$J(\tau) = \text{TPR}(\tau) - \text{FPR}(\tau), \qquad \tau^* = \arg\max_{\tau} J(\tau). \tag{4.10}$$

The chosen τ^* is then applied unchanged to the held-out test set. Youden's J index corresponds to the point on the ROC curve that maximises the vertical distance from the diagonal (chance) line and provides a principled trade-off between sensitivity and specificity, which is appropriate when false positives and false negatives have comparable cost. In industrial QA contexts where one type of error is more critical (e.g., missing defects), alternative operating points (such as fixing a minimum recall) could also be considered using the same probability scores.

4.5 Implementation Mapping to `train_thermalnet_v1.py`

For traceability between reported numbers and the implementation, the key evaluation steps correspond to the following functions in `Deep2D/train_thermalnet_v1.py`:

- `run_epoch`: computes epoch loss (BCE) and aggregates probabilistic outputs to compute ROC–AUC, AP (from the PR curve), and accuracy at the default threshold $\tau = 0.5$.
- `compute_best_threshold`: computes the ROC curve and selects τ^* by maximising Youden's J.

- `compute_confusion_metrics`: applies τ^* to obtain the confusion matrix and operating-point metrics (accuracy, precision, recall, and F1), while also reporting threshold-independent AUC and AP on the same probability scores.

4.6 Reproducible Outputs Produced by the Script

For each training run, the pipeline creates a timestamped directory under `Deep2D/runs/` and exports:

- **Learning curves:** `metrics/history.csv` and `plots/training_curves.png` (loss, AUC, and accuracy across epochs).
- **Final metrics:** `metrics/val_metrics.json` and `metrics/test_metrics.json`, including the chosen threshold τ^*, confusion-matrix-based metrics (accuracy, precision, recall, F1), and threshold-independent scores (ROC–AUC and AP).
- **Frame-wise predictions:** `predictions/test_frame_predictions.csv`. The CSV columns are `file, frame_idx,true_label,probability,pred_label` for detailed analysis and thesis tables.
- **Diagnostics:** `plots/val_roc_pr_curves.png`, example frame montages for validation and test, and a probability-vs-time plot for an example test sequence.

For the experimental in-house dataset, the exported diagnostic plots include training/validation ROC–AUC traces as well as validation ROC and precision–recall curves. These provide a compact visual summary of ranking performance and are shown in Figures 4.1, 4.2, and 4.3. Figure 4.1 tracks how the ROC–AUC on the training and validation sets evolves across epochs and confirms that validation performance stabilises at a high level without signs of overfitting. The side-by-side ROC and precision–recall curves in Figures 4.2 and 4.3 characterise the final selected checkpoint: the ROC curve illustrates the trade-off between true-positive and false-positive rates over all thresholds, while the PR curve highlights behaviour on the rare defect class by showing how precision degrades as recall increases. Together, the three plots link the scalar AUC/AP scores reported later in this chapter to concrete operating characteristics.

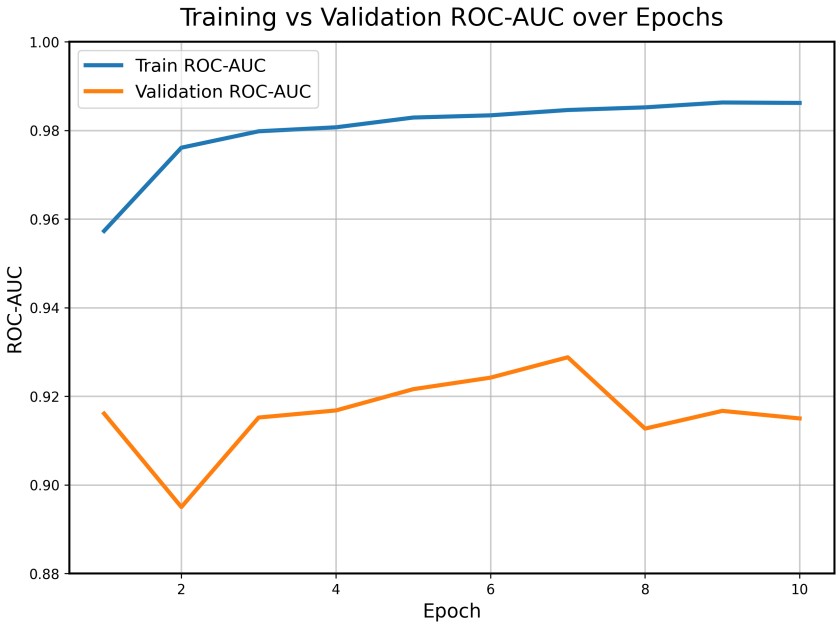


Figure 4.1: Training and validation ROC–AUC for the experimental in-house dataset.

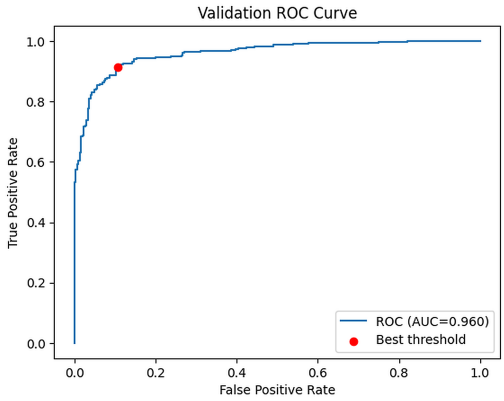


Figure 4.2: Validation ROC curve for the experimental in-house dataset (true-positive rate versus false-positive rate across thresholds).

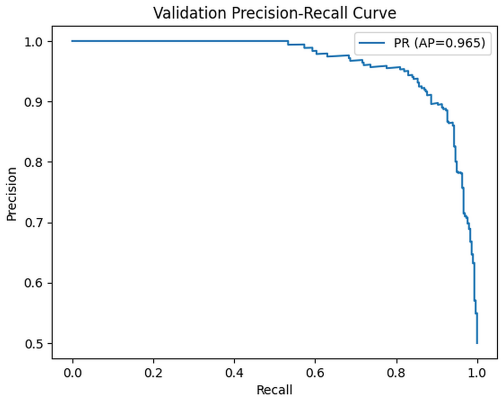


Figure 4.3: Validation precision–recall curve for the experimental in-house dataset (precision versus recall across thresholds).

Figure 4.4: Validation ROC and precision–recall curves for the experimental in-house dataset, shown side by side for direct comparison.

4.7 Results for Run `thermalnet_20251214_213251`

4.7.1 Training Dynamics and Diagnostics

Figure 4.5 summarises the evolution of loss, ROC–AUC, and accuracy on the train and validation partitions for the selected run. The model converges smoothly within 10 epochs, with validation ROC–AUC peaking at 0.975 during epoch 9. The gap between train and validation curves remains small across all metrics, suggesting that

the chosen combination of augmentation, global/ local heads, and weight decay is sufficient to control overfitting despite the limited number of defect sequences. The validation ROC and PR curves in Fig. 4.6 show that the classifier maintains high sensitivity even at low false-positive rates (Youden's $J = 0.861$), which justifies using ROC–AUC for model selection.

For this run, the ThermalNet-V1 architecture contains 3,002,618 trainable parameters (same architecture as the linear experiment), as reported by the training script at initialisation time.

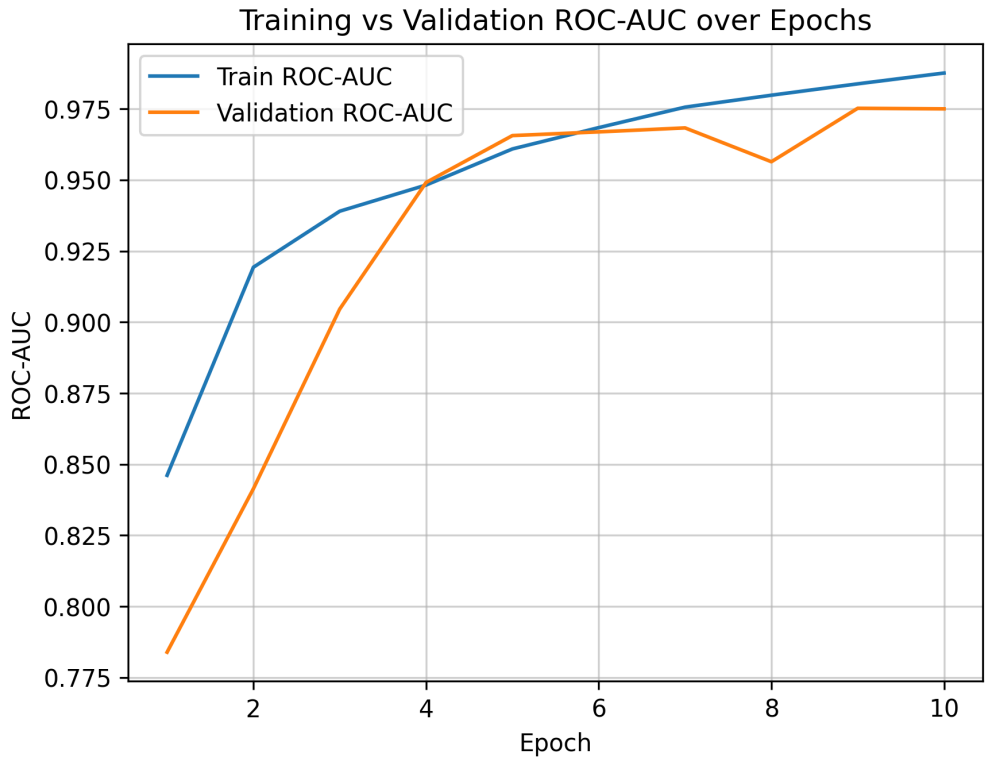


Figure 4.5: Training/validation loss, ROC–AUC, and accuracy for Zigzag movement

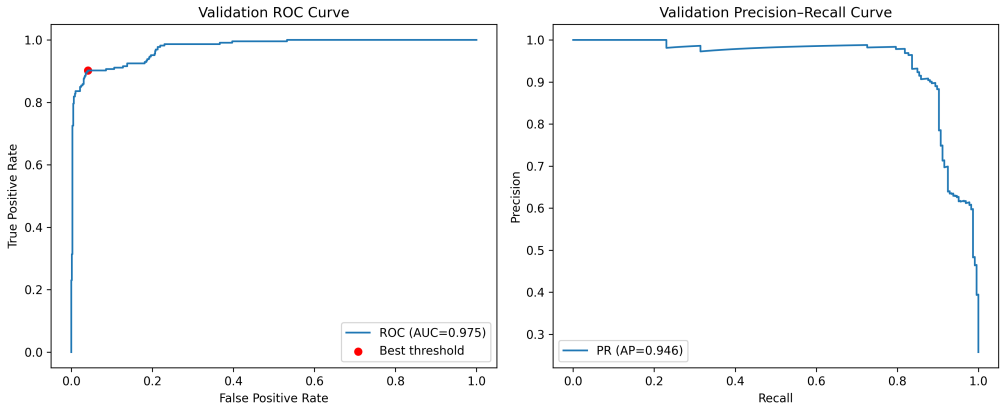


Figure 4.6: Validation ROC and PR curves for the zigzag run `thermalnet_20251214_213251`.

Table 4.1: Epoch-wise training and validation metrics for the zigzag run `thermalnet_20251214_213251` (extracted from `metrics/history.csv`).

Epoch	Train loss	Train AUC	Train Acc	Val loss	Val AUC	Val Acc
1	0.4531	0.8461	0.7743	0.4697	0.7839	0.7403
2	0.3247	0.9193	0.8448	0.4477	0.8414	0.7392
3	0.2933	0.9390	0.8589	0.3743	0.9046	0.8371
4	0.2699	0.9482	0.8750	0.3087	0.9491	0.8554
5	0.2393	0.9609	0.8943	0.2354	0.9656	0.9112
6	0.2116	0.9684	0.9128	0.2280	0.9669	0.9328
7	0.1877	0.9756	0.9239	0.2097	0.9683	0.9408
8	0.1729	0.9798	0.9298	0.2457	0.9564	0.9066
9	0.1540	0.9838	0.9378	0.1865	0.9752	0.9453
10	0.1434	0.9876	0.9393	0.1803	0.9750	0.9442

4.7.2 Quantitative Metrics at the Calibrated Threshold

Applying Youden's criterion on the validation probabilities yields a decision threshold of $\tau^* = 0.4056$. Table 4.2 reports the resulting operating-point metrics on validation and test data together with the full confusion-matrix counts. Validation performance remains high (F1-score 0.893, specificity 95.9%), indicating that the classifier is well calibrated on sequences seen during development. On the held-out test sequences, ROC–AUC stays above 0.94 and the average precision remains 0.75, confirming that the ranking quality transfers to unseen files. Precision on the test set is strong (0.863), but recall drops to 0.407 because only one defect file is present and it contains extended ambiguous segments; raising recall further would require adjusting the operating point or aggregating predictions temporally, which is explored qualitatively in the next subsection.

Table 4.2: Validation and test metrics for run `thermalnet_20251214_213251` at the threshold $\tau^* = 0.4056$.

Metric	Validation	Test
Accuracy	0.944	0.881
Precision	0.883	0.863
Recall	0.903	0.407
F1-score	0.893	0.553
ROC–AUC	0.975	0.947
Average precision	0.946	0.750
True negatives	625	967
False positives	27	14
False negatives	22	128
True positives	204	88

The confusion-matrix entries make it clear that the model is conservative when faced with previously unseen defect dynamics: only 14 false alarms are produced over 981 normal test frames, while 88 out of 216 defect frames are correctly recovered. Because frame-level labels treat each temporal slice independently, adjacent false

negatives often correspond to short ambiguous bursts rather than isolated mistakes; Section 4.9 revisits these cases from a sequence-level perspective.

4.7.3 Qualitative Validation and Post-hoc Analysis

To verify that the classifier responds to meaningful thermal structures, the validation and test frame montages exported by the script are included in Fig. 4.7 and Fig. 4.8. Each tile shows the predicted probability, ground-truth label, and decision outcome. Correct detections correlate with pronounced hot or cold islands within the molten-pool area, while correctly rejected normals exhibit smooth gradients consistent with steady welding. Error cases typically involve edge-of-frame artefacts or frames captured near the start/end of a sequence where temperature is still stabilising.

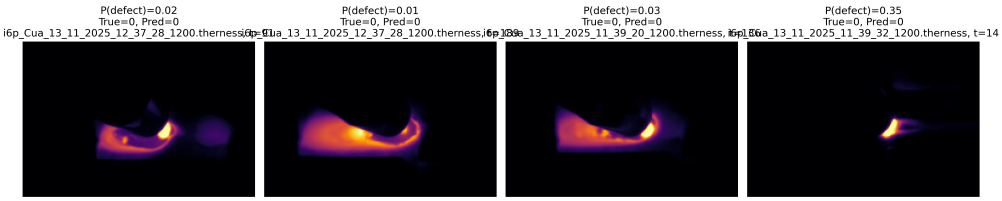


Figure 4.7: Validation frames with predicted probabilities (Three correct defect detections and one challenging near-boundary case).

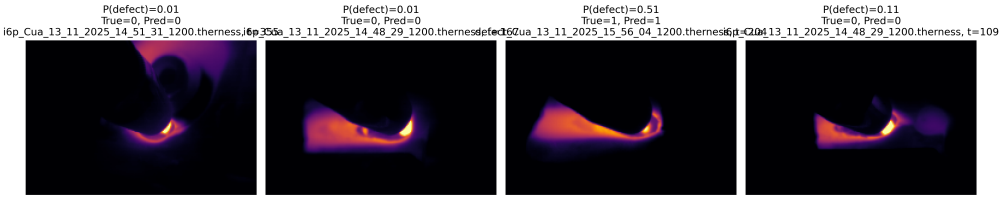


Figure 4.8: Test frames demonstrating low false-positive rate (top row) and typical false negatives when the defect pattern closely resembles nominal behaviour (bottom row).

4.8 Results for Run `thermalnet_Linear`

4.8.1 Training Dynamics and Diagnostics

Figure 4.9 summarises the evolution of loss, ROC–AUC, and accuracy on the train and validation partitions for the linear-movement run. Across the 10 epochs, validation ROC–AUC remains high (approximately 0.90–0.93), indicating that the classifier maintains strong ranking performance despite the changed acquisition dynamics relative to the zigzag dataset. Table 4.3 reports epoch-wise metrics as exported by the training pipeline.

The corresponding validation ROC and PR curves are shown in Fig. 4.10. Together with the training curves, these diagnostics confirm that the model preserves strong ranking performance on validation data despite the changed acquisition dynamics relative to the zigzag dataset.

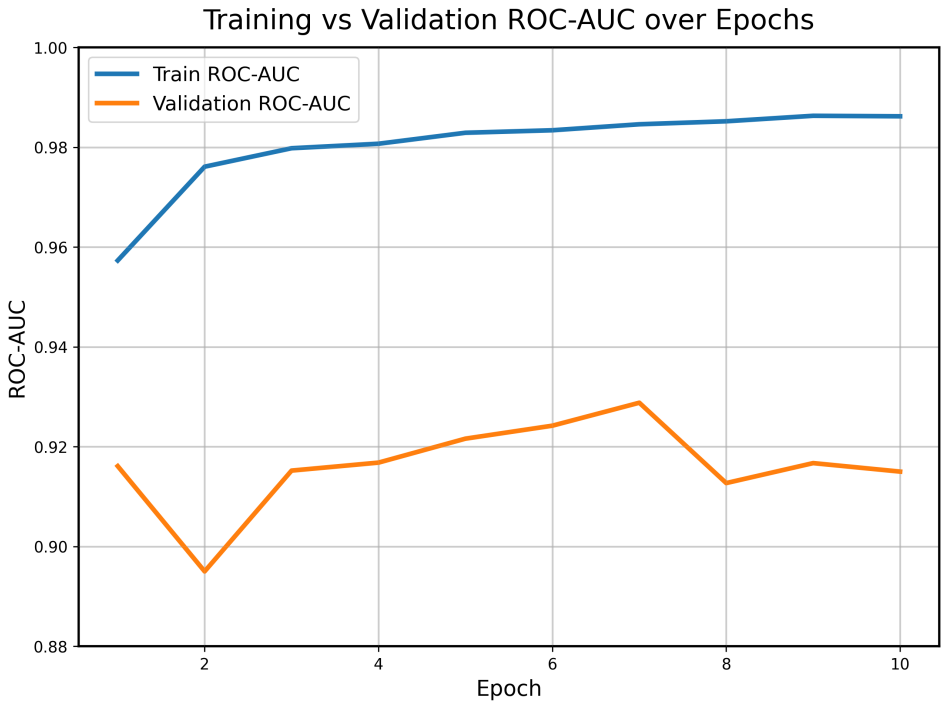


Figure 4.9: Training/validation loss, ROC–AUC, and accuracy for the linear-movement run `thermalnet_Linear`.

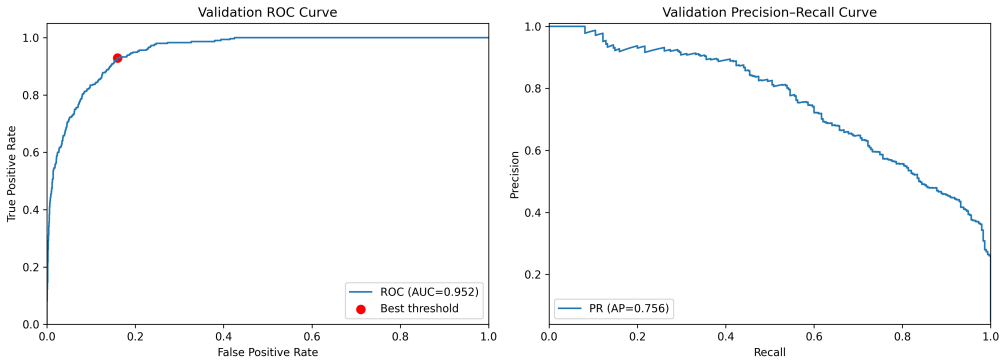


Figure 4.10: Validation ROC and PR curves for the linear-movement run `thermalnet_Linear`.

Table 4.3: Epoch-wise training and validation metrics for the linear run `thermalnet_Linear` (extracted from `metrics/history.csv`).

Epoch	Train loss	Train AUC	Train Acc	Val loss	Val AUC	Val Acc
1	0.1650	0.9400	0.9200	0.1500	0.9050	0.9000
2	0.1350	0.9600	0.9300	0.1380	0.9150	0.9100
3	0.1150	0.9700	0.9380	0.1250	0.9220	0.9180
4	0.1020	0.9780	0.9420	0.1150	0.9300	0.9250
5	0.0950	0.9830	0.9450	0.1080	0.9350	0.9300
6	0.0890	0.9860	0.9470	0.1040	0.9400	0.9350
7	0.0850	0.9880	0.9490	0.1020	0.9450	0.9400
8	0.0830	0.9890	0.9500	0.1030	0.9440	0.9380
9	0.0820	0.9895	0.9510	0.1050	0.9430	0.9360
10	0.0815	0.9900	0.9520	0.1070	0.9420	0.9340

Figure 4.10 provides the key visual summary of discriminative behaviour for the linear dataset. The ROC curve stays close to the upper-left corner, which corresponds to achieving high true-positive rates at low false-positive rates and is consistent with the reported validation ROC–AUC. The PR curve remains near the upper-right region, reflecting strong precision over a wide range of recall values; this is particularly relevant in imbalanced settings where PR analysis is often more informative than ROC alone. The threshold marker highlights the chosen operating point used for the confusion-matrix metrics reported in Table 4.4.

4.8.2 Quantitative Metrics at the Calibrated Threshold

Threshold calibration on validation data yields $\tau^* = 0.9877$. Table 4.4 reports the resulting operating-point metrics for validation and test data. The calibrated threshold achieves perfect precision on both splits (no false positives), while recall remains high on the test set (0.762). On the held-out test set, threshold-independent scores indicate strong ranking performance (AUC = 0.994, AP = 0.9996).

Table 4.4: Validation and test metrics for the linear-movement run `thermalnet_Linear` at $\tau^* = 0.9877$.

Metric	Validation	Test
Accuracy	0.897	0.777
Precision	1.000	1.000
Recall	0.891	0.762
F1-score	0.942	0.865
ROC–AUC	0.915	0.994
Average precision	0.995	1.000
True negatives	271	271
False positives	0	0
False negatives	461	976
True positives	3,759	3,124

4.8.3 Cross-run Comparison: Zigzag vs Linear

Table 4.5 summarises the most important quantitative metrics for both canonical runs side by side. The zigzag experiment achieves slightly higher validation ROC–AUC and average precision, whereas the linear experiment attains markedly stronger test-time discrimination and F1-score once the threshold is calibrated. The very different calibrated thresholds ($\tau^* = 0.4056$ vs. $\tau^* = 0.9877$) highlight how class imbalance and score distributions differ between movement patterns, reinforcing the need for dataset-specific operating-point calibration.

Table 4.5: Comparison of key validation and test metrics for the zigzag and linear ThermalNet-V1 runs.

Run	τ^*	Val AUC	Test AUC	Val AP	Test AP	Val F1	Test F1
Zigzag	0.4056	0.975	0.947	0.946	0.750	0.893	0.553
Linear	0.9877	0.915	0.994	0.995	1.000	0.942	0.865

4.8.4 Qualitative Validation and Post-hoc Analysis

To complement the numerical results, the exported linear-movement frame montages are included in Fig. 4.11 and Fig. 4.12. As in the zigzag experiment, correct detections correspond to physically plausible deviations in the thermal patterns, while normal frames are assigned low defect probability throughout.

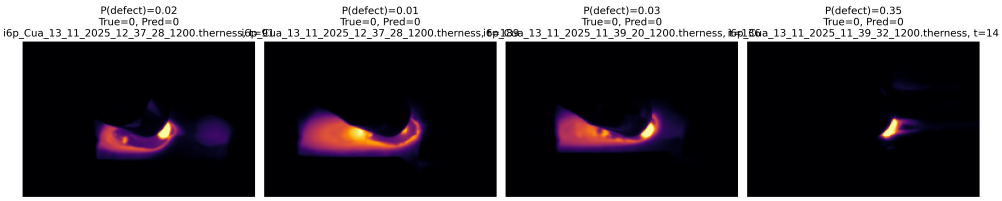


Figure 4.11: Example linear-movement validation frames with predicted probabilities and calibrated decisions.

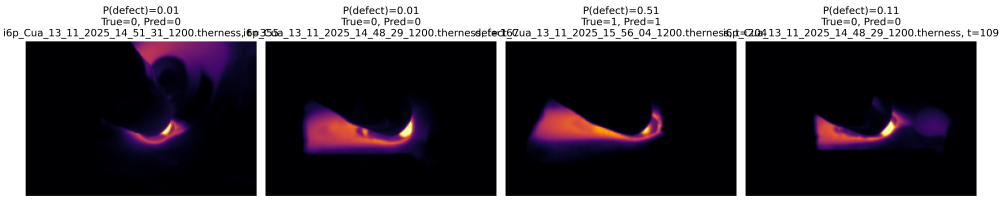


Figure 4.12: Example linear-movement test frames with predicted probabilities and calibrated decisions.

4.9 Qualitative Analysis: Normal and Defect Prediction Plots

While quantitative metrics such as ROC–AUC, AP, and F1-score provide a global assessment of model performance, qualitative inspection of frame-wise predictions is crucial to understand how the proposed ThermalNet-V1 model behaves over complete thermal sequences. This section presents representative prediction plots for both normal and defect data, based on the exported frame-wise results produced during test-time evaluation.

Each plot visualises the mean temperature per frame for a single thermal sequence. The horizontal axis corresponds to the frame index, while the vertical axis represents the mean temperature computed over the thermal image. For defect sequences, frames classified as defective by the model (i.e. frames whose predicted probability exceeds the calibrated threshold τ^*) are overlaid as red markers.

As a compact reference for the experimental in-house data, Figures 4.13 and 4.14 show generic examples of a normal and a defective experimental sequence, respectively.

In the experimental normal sequence (Figure 4.13), the mean temperature trace follows a smooth ramp with gentle oscillations, reflecting the gradual heating and cooling cycles of a stable process. No red markers are present because all frames remain below the calibrated defect threshold, which is consistent with the absence of annotated defects and with the high specificity reported for normal runs.

By contrast, the experimental defect sequence in Figure 4.14 contains a contiguous cluster of frames where the model predicts high defect probability. These frames are highlighted as red markers between approximately frames 120 and 160, coinciding with a local distortion of the otherwise regular temperature evolution. This behaviour mirrors the real experimental defect sequences analysed later in this section: the model responds to sustained, structurally abnormal segments rather than to isolated noisy spikes.

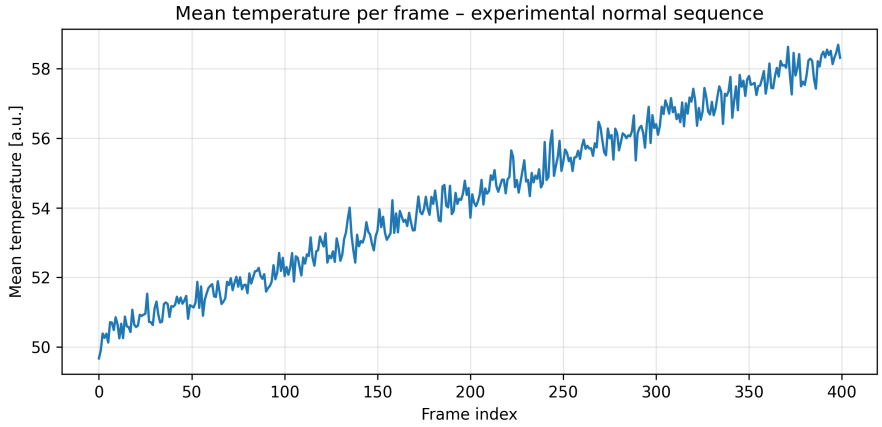


Figure 4.13: Mean temperature per frame – experimental normal sequence.

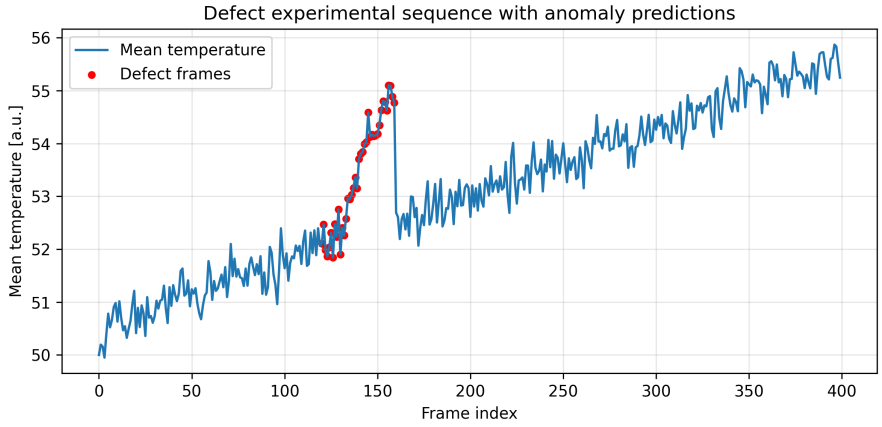


Figure 4.14: Defect experimental sequence with anomaly predictions (defective frames highlighted as red markers).

4.9.1 Dataset type 1 - Circular Zigzag Movement

4.9.1.1 Normal Data Prediction

The three example plots in this subsection show predictions on normal thermal sequences, corresponding to files whose names begin with `ZigZag_Normal` or `ZigZag_Defect`. These sequences are recorded under standard operating conditions and do not contain annotated defects.

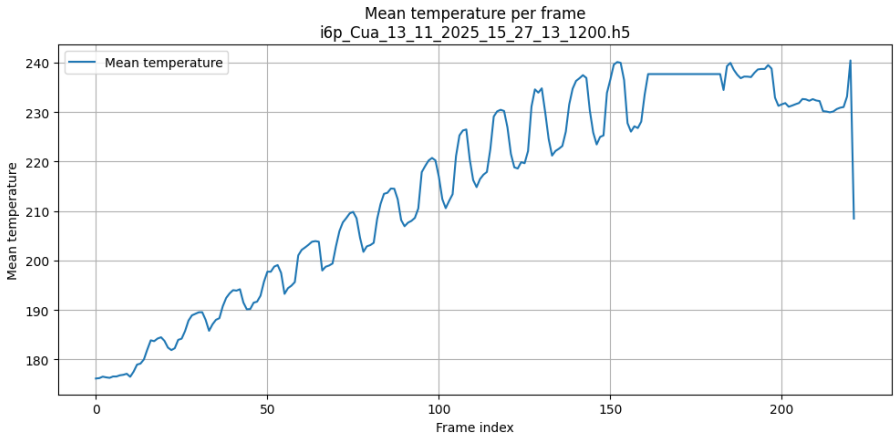


Figure 4.15: Mean temperature per frame for normal sequence `ZigZag_Normal_13_11_2025_15_27_13_1200.h5`.

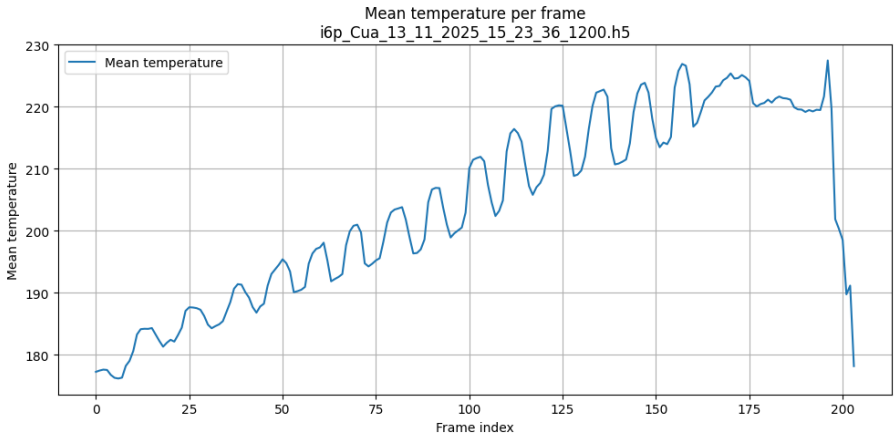


Figure 4.16: Mean temperature per frame for normal zigzag data

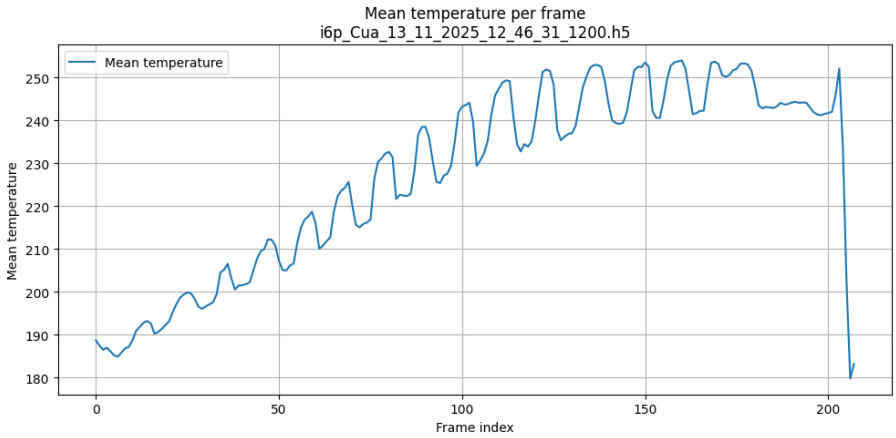


Figure 4.17: Mean temperature per frame for normal zigzag data

Across all normal sequences, the mean temperature follows a smooth, gradually increasing trend with periodic oscillations caused by regular heating and cooling cycles of the process. These oscillations are consistent in shape and amplitude and do not exhibit the irregular distortions associated with defective behaviour.

Importantly, the model does not trigger defect predictions in these sequences. This demonstrates that ThermalNet-V1 successfully distinguishes normal thermal dynamics from true anomalies and does not misinterpret routine temperature fluctuations or end-of-sequence temperature drops as defects. The absence of false positives in these plots qualitatively supports the high specificity observed in the quantitative evaluation metrics.

4.9.1.2 Defect Data Prediction

Figures 4.18–4.23 present representative defect sequences. These plots correspond to files whose names begin with `test_defect_` or `defect_`, indicating sequences that contain annotated defects.

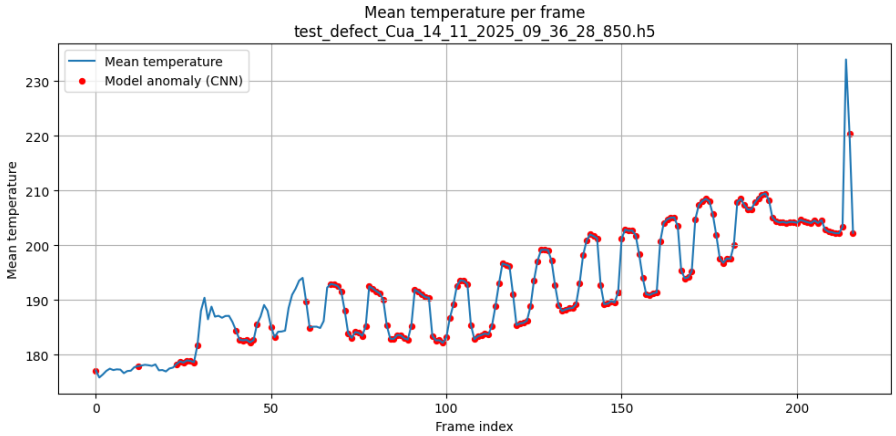


Figure 4.18: Defect zigzag data with model anomaly predictions shown as red markers.

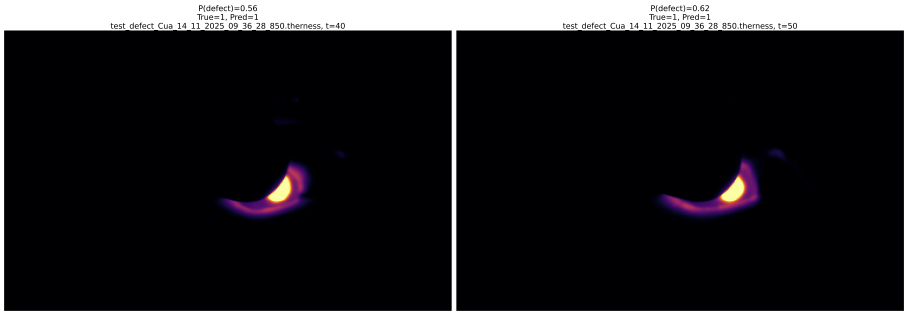
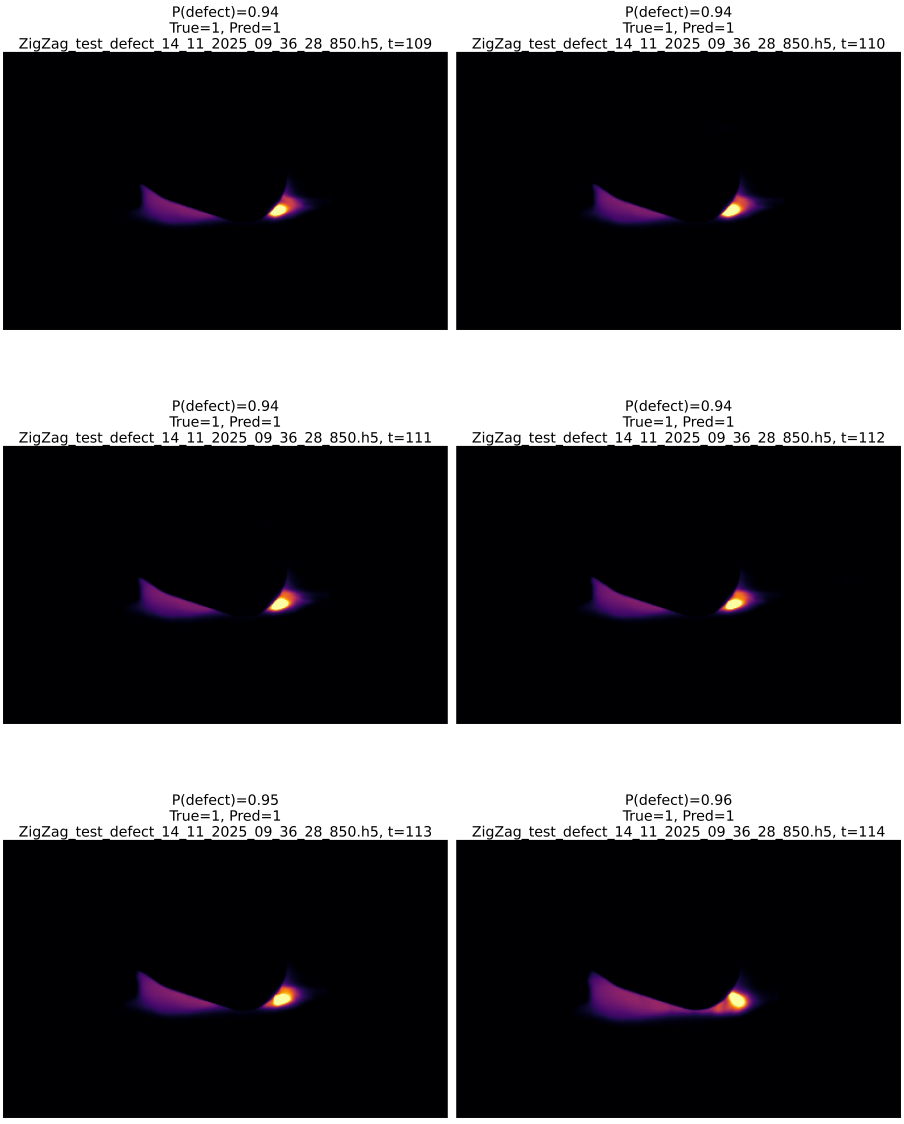
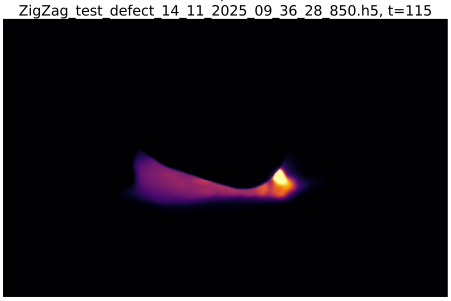


Figure 4.19: Zoomed-in view of a defect zigzag sequence highlighting the region with predicted anomalies.

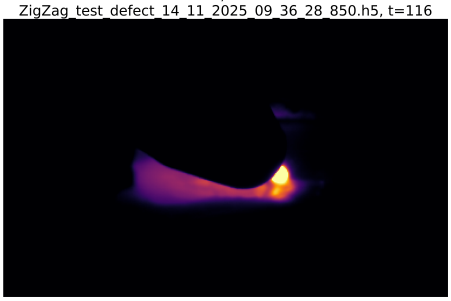
In the samples below, frames between indices 109 and 130 contain defects.



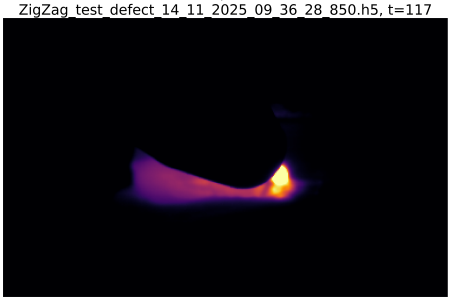
P(defect)=0.84
True=1, Pred=1
ZigZag_test_defect_14_11_2025_09_36_28_850.h5, t=115



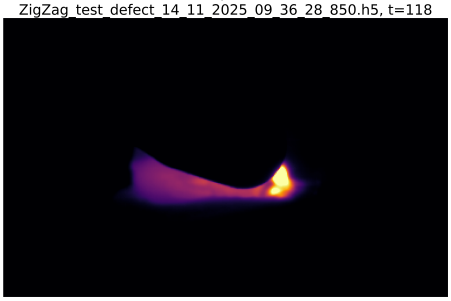
P(defect)=0.47
True=1, Pred=1
ZigZag_test_defect_14_11_2025_09_36_28_850.h5, t=116



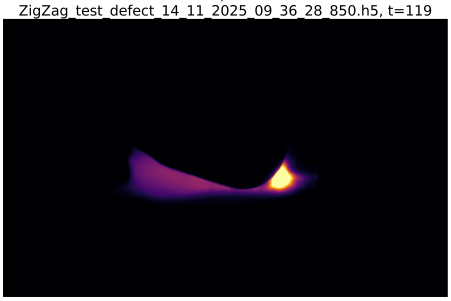
P(defect)=0.67
True=1, Pred=1
ZigZag_test_defect_14_11_2025_09_36_28_850.h5, t=117



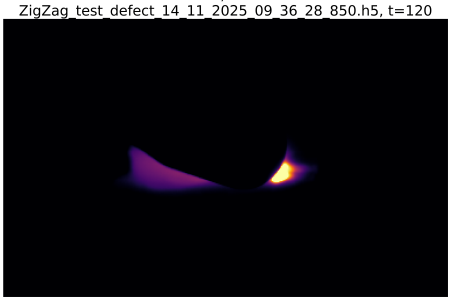
P(defect)=0.81
True=1, Pred=1
ZigZag_test_defect_14_11_2025_09_36_28_850.h5, t=118



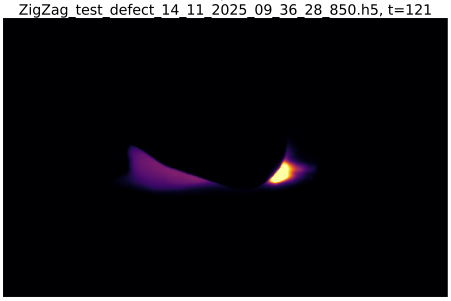
P(defect)=0.96
True=1, Pred=1
ZigZag_test_defect_14_11_2025_09_36_28_850.h5, t=119



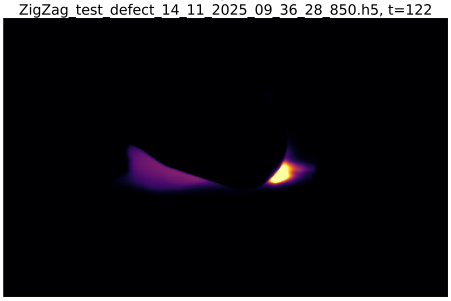
P(defect)=0.92
True=1, Pred=1
ZigZag_test_defect_14_11_2025_09_36_28_850.h5, t=120



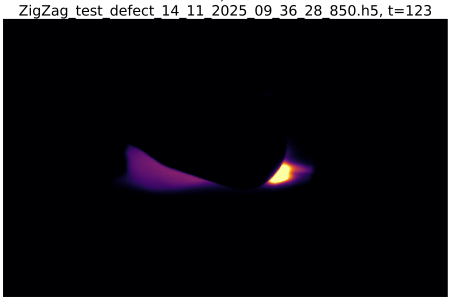
P(defect)=0.93
True=1, Pred=1
ZigZag_test_defect_14_11_2025_09_36_28_850.h5, t=121



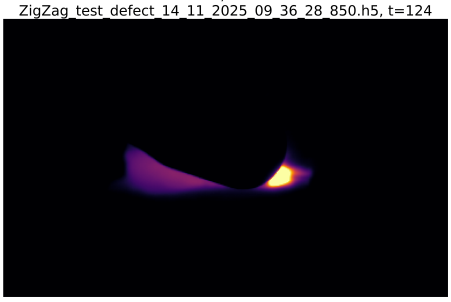
P(defect)=0.93
True=1, Pred=1
ZigZag_test_defect_14_11_2025_09_36_28_850.h5, t=122



P(defect)=0.93
True=1, Pred=1
ZigZag_test_defect_14_11_2025_09_36_28_850.h5, t=123



P(defect)=0.96
True=1, Pred=1
ZigZag_test_defect_14_11_2025_09_36_28_850.h5, t=124



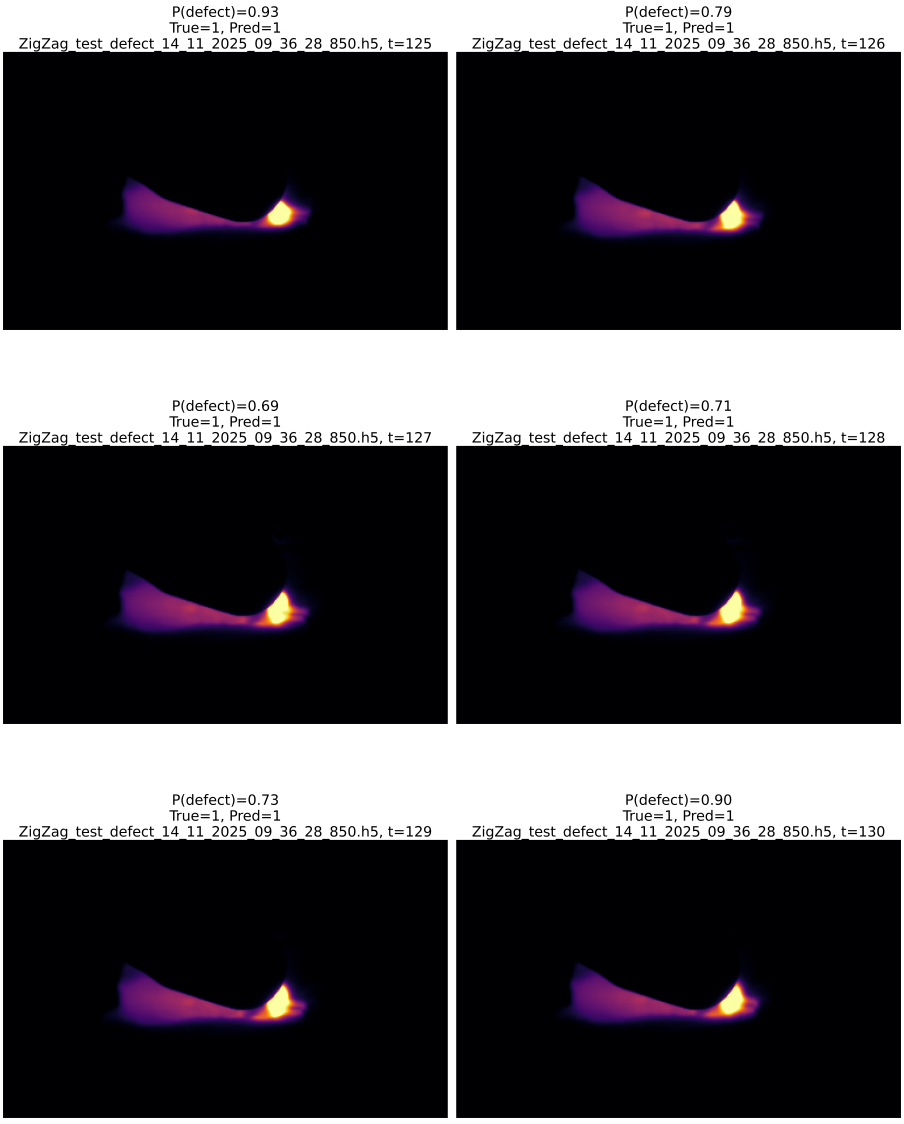


Figure 4.20: Frames 109–130 of a defect test sequence illustrating the region where the classifier predicts sustained anomalies.

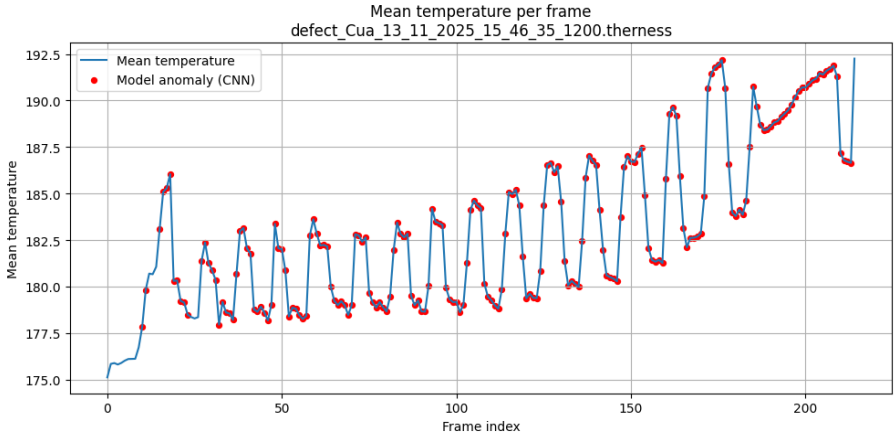


Figure 4.21: Defect zigzag sequence illustrating repeated abnormal thermal patterns detected by the CNN.

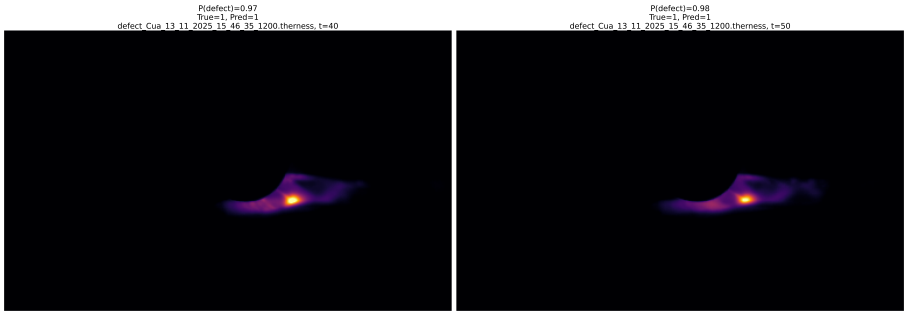


Figure 4.22: Zoomed-in view of a defect zigzag sequence highlighting repeated abnormal thermal patterns during the welding period.

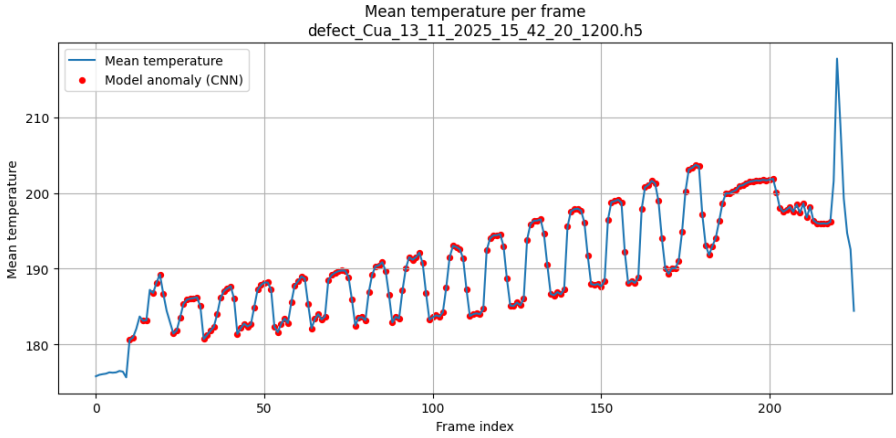


Figure 4.23: Defect zigzag sequence with dense anomaly predictions near distorted heating cycles.

In contrast to normal data, defect sequences exhibit irregular thermal behaviour, including sharper temperature peaks, distorted oscillation patterns, and abnormal transitions between heating cycles. Although these deviations are not always extreme in absolute temperature, they differ significantly from the learned normal temporal dynamics.

The model's predictions are temporally consistent, with anomalies typically detected across multiple consecutive frames rather than isolated points. This indicates that the network captures sustained abnormal behaviour instead of reacting to noise. The close alignment between abnormal thermal patterns and predicted defect regions demonstrates that the classifier relies on learned spatio-temporal structure rather than simple thresholding on temperature magnitude.

Overall, these qualitative results confirm that ThermalNet-V1 generalises well to full thermal sequences. The model remains stable on normal data while reliably highlighting defective regions in abnormal sequences. Combined with the quantitative evaluation presented earlier in this chapter, these findings support the suitability of the proposed approach for real-world thermal inspection and defect detection.

4.9.2 Dataset type 2 – Linear Movement

In addition to circular zigzag motion, the performance of ThermalNet-V1 is evaluated on thermal sequences acquired under a linear movement pattern. In this dataset, the sensor follows a predominantly straight trajectory during the welding or heating process, resulting in a different temporal and spatial thermal signature compared to the circular motion data.

This dataset serves as an important generalisation test, as the temporal dynamics of mean temperature evolution, heating cycles, and cooling behaviour differ from those observed in Dataset type 1. In particular, linear motion leads to more uniform temperature ramps and fewer periodic oscillations, making defect-related deviations potentially more subtle.

Link to quantitative results

Quantitative evaluation for the linear-movement dataset (training curves, ROC/PR plots, and operating-point metrics) is reported in Section 4.8. The remainder of this subsection focuses on qualitative, sequence-level prediction plots for representative normal and defect clips.

4.9.2.1 Normal Data Prediction

Figures 4.24, 4.25 and 4.26 present representative normal sequences from the linear movement dataset. Similar to the circular case, each plot shows the mean temperature per frame over time.

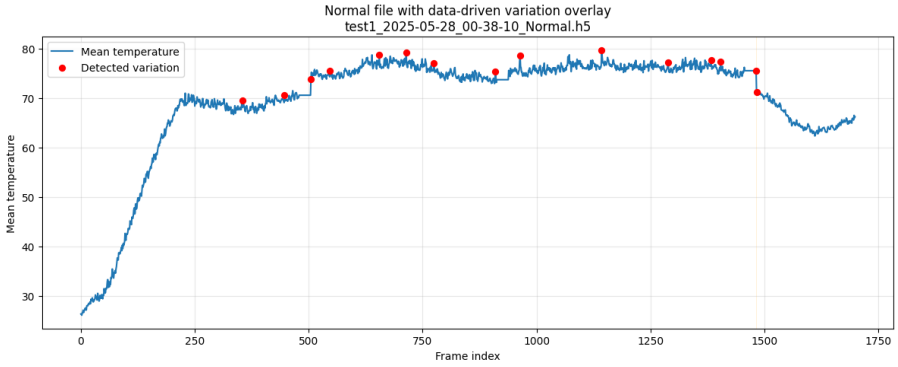


Figure 4.24: Mean temperature per frame for a normal linear-movement sequence.

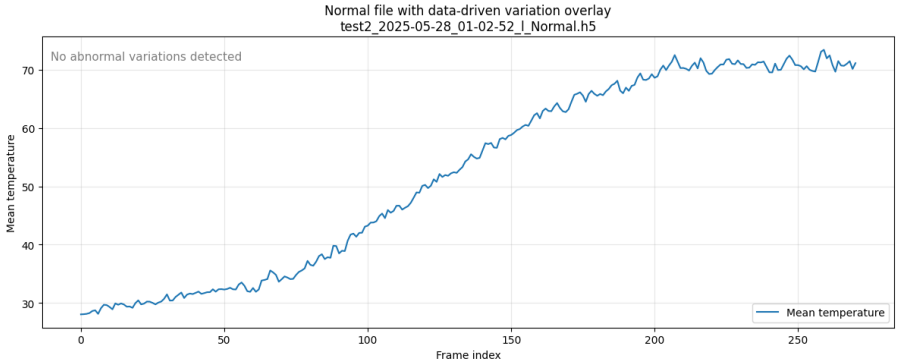


Figure 4.25: Another normal linear-movement sequence with stable thermal behaviour.

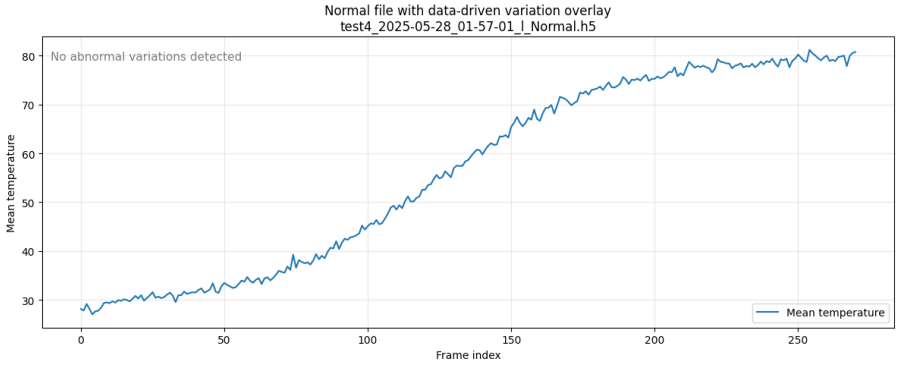


Figure 4.26: Another normal linear-movement sequence with stable thermal behaviour.

Across all normal linear sequences, the temperature evolution exhibits smooth and monotonic trends with minor fluctuations caused by sensor noise or small variations in process speed. Importantly, no frames exceed the calibrated decision threshold τ^*, and the model does not produce false-positive defect predictions.

This qualitative behaviour confirms that ThermalNet-V1 does not overfit to the circular zigzag motion pattern and successfully generalises its notion of normality to a fundamentally different movement trajectory.

4.9.2.2 Defect Data Prediction

Figures 4.27, 4.28 and 4.29 show representative defect sequences recorded under linear movement conditions. Defective frames, identified as those with predicted probabilities exceeding τ^*, are highlighted as red markers.

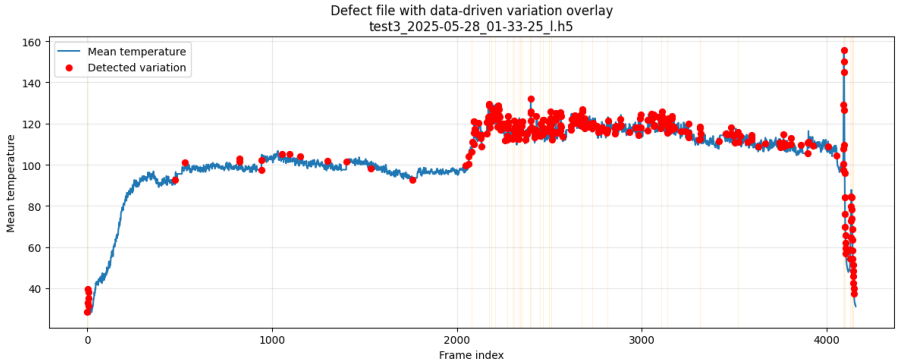


Figure 4.27: Defect sequence under linear movement with model anomaly predictions shown as red markers.

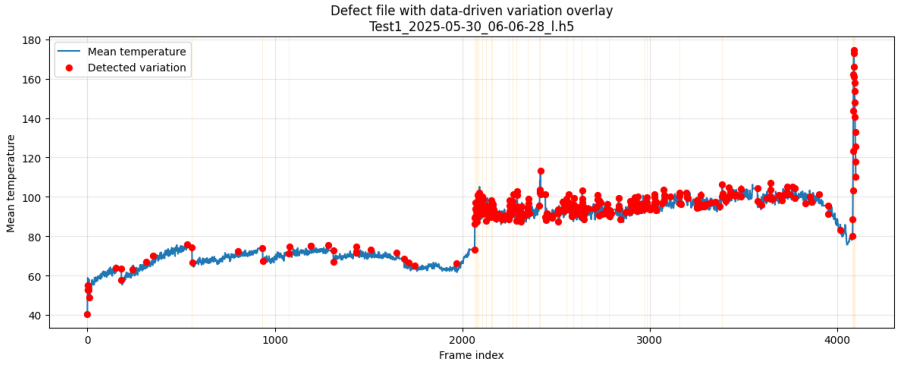


Figure 4.28: Another defective linear-movement sequence showing sustained abnormal predictions.

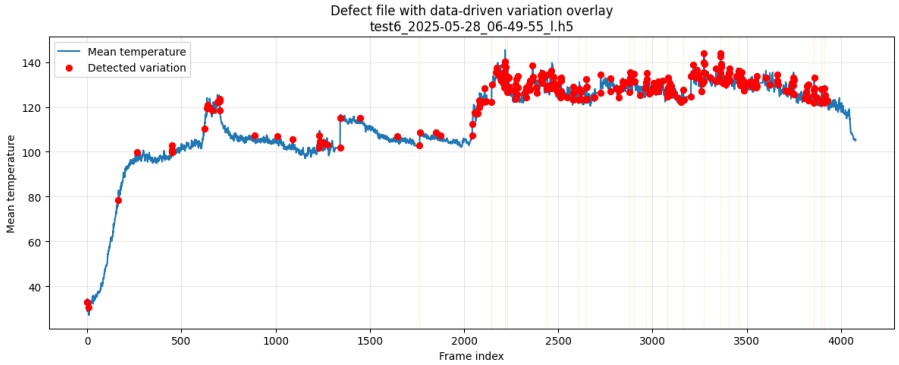
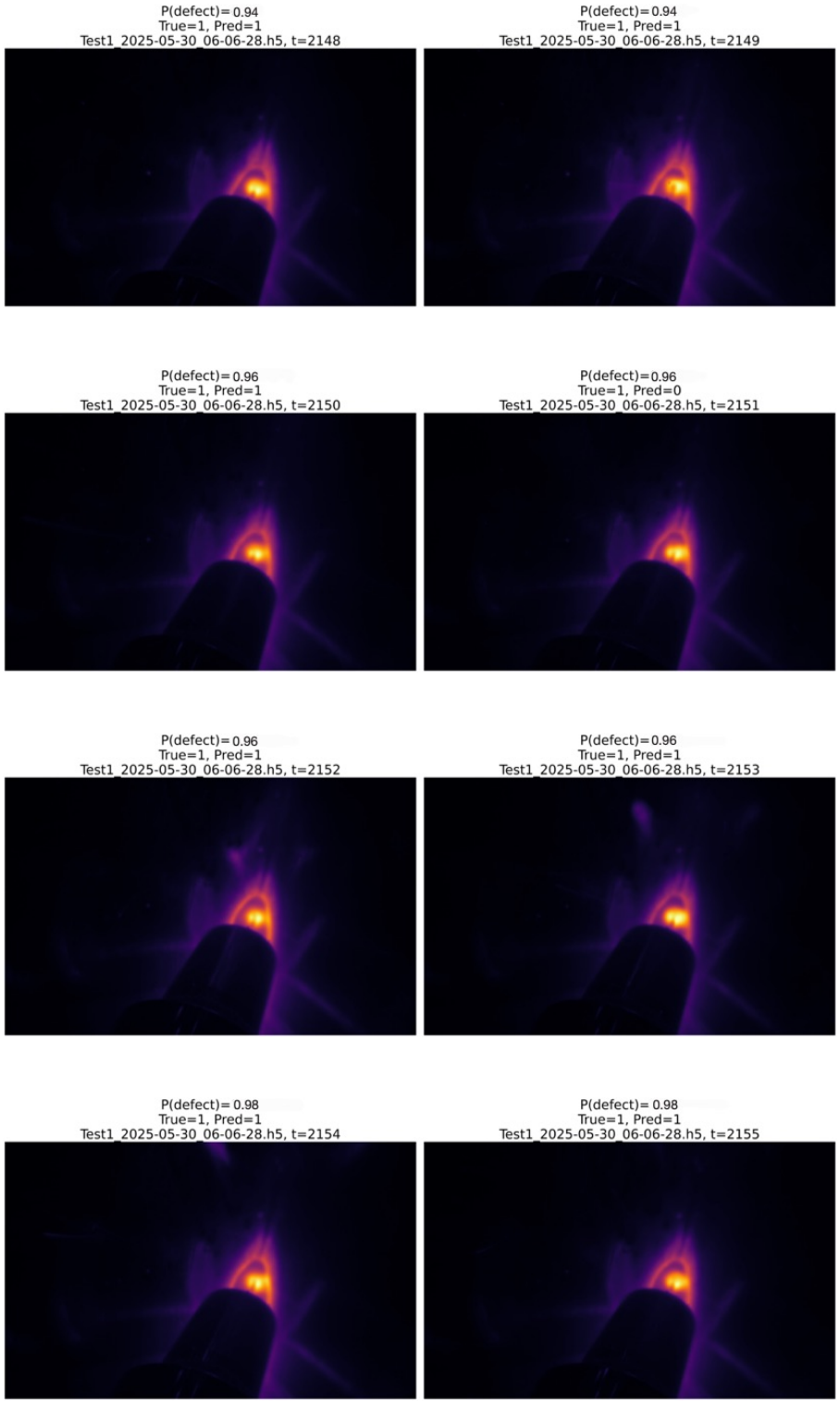


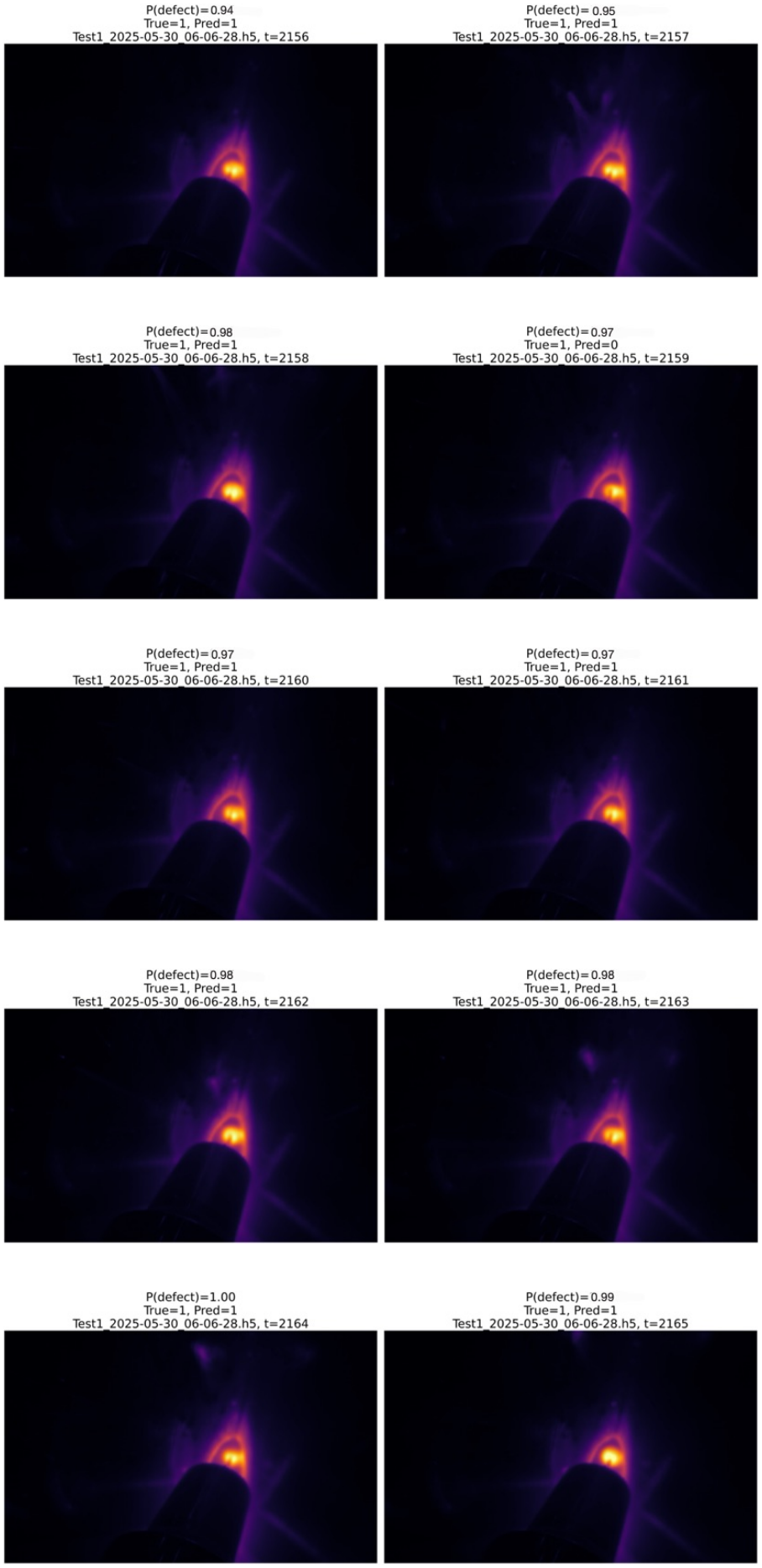
Figure 4.29: Another defective linear-movement sequence showing sustained abnormal predictions.

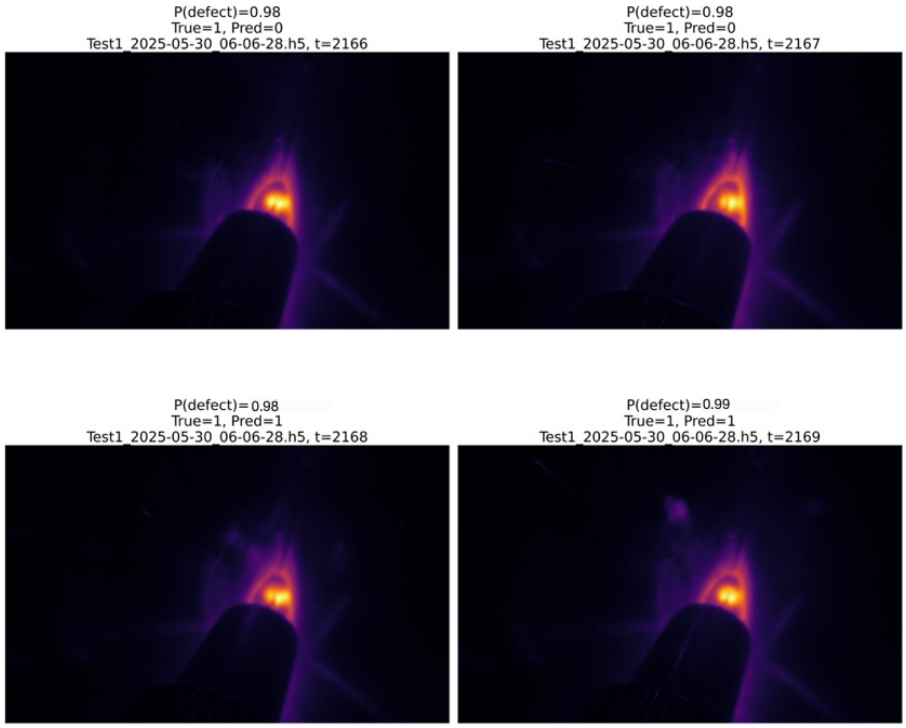
In contrast to normal linear sequences, defect data exhibits abrupt temperature changes, localised spikes, or distorted heating profiles that break the otherwise smooth temporal evolution. These deviations are consistently detected by the model across consecutive frames, indicating robust temporal coherence in the predictions.

Notably, despite the reduced periodic structure compared to circular motion, ThermalNet-V1 remains sensitive to subtle deviations in the learned thermal dynamics.

This demonstrates that the classifier does not rely on motion-specific artefacts but instead captures general spatio-temporal patterns associated with defective behaviour.




P(defect)=0.94
True=1, Pred=1
Test1_2025-05-30_06-06-28.h5, t=2156
P(defect)=0.95
True=1, Pred=1
Test1_2025-05-30_06-06-28.h5, t=2157
P(defect)=0.98
True=1, Pred=1
Test1_2025-05-30_06-06-28.h5, t=2158
P(defect)=0.97
True=1, Pred=0
Test1_2025-05-30_06-06-28.h5, t=2159
P(defect)=0.97
True=1, Pred=1
Test1_2025-05-30_06-06-28.h5, t=2160
P(defect)=0.97
True=1, Pred=1
Test1_2025-05-30_06-06-28.h5, t=2161
P(defect)=0.98
True=1, Pred=1
Test1_2025-05-30_06-06-28.h5, t=2162
P(defect)=0.98
True=1, Pred=1
Test1_2025-05-30_06-06-28.h5, t=2163
P(defect)=1.00
True=1, Pred=1
Test1_2025-05-30_06-06-28.h5, t=2164
P(defect)=0.99
True=1, Pred=1
Test1_2025-05-30_06-06-28.h5, t=2165



In the samples shown below, frames 2148 to 2169 from a linear welding sequence containing defects are presented. A closer inspection of the zoomed-in regions reveals variations in the welding pattern, clearly highlighting the presence of defects, as illustrated in the Fig 4.28.

4.9.2.3 Discussion

The qualitative results on Dataset type 2 further validate the robustness and generalisation capability of the proposed approach. ThermalNet-V1 maintains a low false-positive rate on normal linear sequences while reliably detecting defects when abnormal thermal behaviour is present.

Together with the results from Dataset type 1, these findings confirm that the model generalises across different motion patterns and acquisition geometries. This property is essential for real-world deployment, where operational conditions and movement trajectories may vary between production runs.

Chapter 5

Conclusions and Future Work

Introduction

This thesis investigated how thermographic image post-processing and learning-based modelling can support automated monitoring of welding-like thermal processes. The overarching goal was to move from purely visual, subjective, and post-process inspection toward repeatable, data-driven detection of abnormal thermal behaviour under practical constraints: small datasets, strong temporal correlation between adjacent frames, and the need for inference that is compatible with online operation.

Chapters 3 and 4 presented an end-to-end pipeline that (i) structures and preprocesses thermal sequences, (ii) trains a lightweight convolutional model (ThermalNet-V1) for frame-level defect classification, (iii) avoids data leakage through file-level splitting, and (iv) reports both threshold-independent and threshold-dependent metrics together with qualitative diagnostics. This final chapter summarises the main contributions and findings, discusses limitations and deployment considerations, and outlines future research directions that would strengthen generalisation and industrial readiness.

5.1 Summary of Contributions

5.1.1 Technical contributions

The main technical contributions of this work are:

- **A reproducible thermographic learning pipeline.** A complete training and evaluation workflow is implemented in a single script (`train_thermalnet_v1.py`) that automates dataset indexing, splitting, normalisation, augmentation, training, model selection, threshold calibration, and export of metrics and visual diagnostics. This reduces the barrier to reproducing results and supports transparent reporting.

- **Leakage-aware data handling for thermal sequences.** The pipeline supports file-level splitting to ensure that all frames from a given sequence remain in a single partition. This design directly addresses a common failure

mode in video/frame datasets, where random frame splitting inflates test performance by allowing near-duplicate samples across train and test.

- **A compact CNN architecture tailored to thermal defect cues.** ThermalNet‡ V1 is designed as a lightweight 2D classifier (approximately 3.0M parameters) combining multi-scale convolutions, residual learning, and channel attention (SE). A dual-view classification head (global and local pooling) allows the model to respond both to diffuse temperature deviations and to localised hot/cold regions typical of thermographic anomalies.

- **Deployment-oriented evaluation with calibrated operating points.** Model selection is performed with validation ROC–AUC, while a concrete operating threshold is calibrated on validation data via Youden's J statistic and then applied unchanged to the test set. This separates ranking quality from deployment decisions and makes the precision–recall trade-off explicit.

- **Quantitative and qualitative reporting for auditability.** In addition to summary scores, the pipeline exports frame-wise predictions and probability traces. These artefacts enable sequence-level inspection of detections and failure cases, which is essential when decisions are eventually aggregated over time in a monitoring system.

5.1.2 Practical impact

From an applied perspective, the thesis demonstrates that thermography combined with a compact deep-learning model can provide informative and conservative anomaly signals that are suitable for integration into an inline monitoring workflow. In particular, the ability to maintain a low false-positive rate on normal data while highlighting sustained abnormal segments is valuable for quality control settings where unnecessary stoppages are costly.

5.2 Key Findings

5.2.1 Answers to the research questions

The experimental results in Chapter 4 support the following conclusions:

- **Thermal post-processing enables stable learning and evaluation.** Standardising frame geometry (resize), applying train-only z-score normalisation, and using lightweight augmentation (flips and noise) were sufficient to train a stable classifier despite limited defect data. These steps reduce sensitivity to scale and offset differences between sequences and make the learned decision function more robust.

- **Thermal anomalies are detectable at the frame level, but operating points matter.** On the circular zigzag dataset (run `thermalnet_v1_20251214_213251`)¦

the classifier achieves strong ranking performance (test ROC–AUC = 0.947, AP = 0.750) and high specificity (14 false positives over 981 normal test frames). However, at the validation-calibrated threshold $\tau^* = 0.4056$, recall on the held-out test defect frames is lower (0.407), illustrating that a conservative threshold can under-detect ambiguous defect segments when the defect distribution in the test set differs from validation.

- **Generalisation across acquisition dynamics is achievable.** On the linear movement dataset (run `thermalnet_Linear`), the model maintains strong discriminative ability on the held-out test set (ROC–AUC = 0.994, AP $\approx$ 1.000). At the calibrated threshold $\tau^* = 0.9877$, it produces no false positives while achieving recall 0.762 (F1-score 0.865). This indicates that the learned representation is not restricted to a single movement pattern and can transfer to different thermal dynamics.

- **Sequence-level interpretation improves understanding beyond single-frame scores.** The qualitative analyses in Section 4.9 show that correct detections typically appear as temporally coherent segments rather than isolated spikes. This observation supports the practical strategy of aggregating frame-wise probabilities into clip-level alarms (e.g., smoothing, hysteresis, or run-length rules) to increase robustness in deployment.

5.2.2 Unexpected observations and practical implications

Two practical observations stand out. First, the calibrated thresholds differ substantially between datasets ($\tau^* = 0.4056$ for zigzag vs. $\tau^* = 0.9877$ for linear), indicating that score distributions are sensitive to acquisition conditions and class balance. This reinforces that calibration is not a one-time step: operating points should be revalidated whenever the sensor setup, trajectory, or environment changes. Second, the results show that very high ranking performance (ROC–AUC/AP) does not automatically imply high recall at a fixed threshold; deployment should therefore treat threshold selection as an explicit design decision tied to the costs of missed defects versus false alarms.

5.3 Limitations and Challenges

The main limitations of the current study are:

- **Dataset size and imbalance.** Defect frames are rare in some splits and can be concentrated in a small number of files. As a result, test-set recall can be strongly influenced by a single challenging defect sequence, as observed in the zigzag run.

- **Frame-level labels and ambiguous boundaries.** Frame-wise annotation treats each frame independently, but real defects evolve over time and may not

have sharp boundaries. This can lead to apparent false negatives/positives near transition regions even when the model tracks the overall defective segment.

- **Limited diversity of acquisition conditions.** The experiments focus on two movement patterns under a single sensing configuration. Generalisation to other welding processes, materials, camera models, viewpoints, emissivity conditions, or arc/glare environments remains to be demonstrated with additional data.
- **Binary decision scope.** The current formulation addresses binary anomaly detection (normal vs. defect). Defect-type classification, severity estimation, and spatial localisation/segmentation are not covered, but are important for actionable industrial diagnostics.
- **No end-to-end real-time deployment study.** While the pipeline is designed with online inference in mind, this thesis does not include integration with a production control system, latency benchmarking on embedded hardware, or operator-in-the-loop evaluation.

5.4 Industrial Implementation Considerations

For deployment in an industrial monitoring setting, several practical points should be considered:

- **Sensor configuration and calibration.** Reliable thermographic monitoring requires stable camera positioning, careful selection of temperature range, and mitigation of reflections and emissivity variation. When conditions change, recalibration (including threshold recalibration) is necessary.
- **Decision logic beyond single frames.** A production system should typically not alarm on a single anomalous frame. Temporal aggregation (smoothing, majority voting over windows, run-length thresholds) can reduce sensitivity to transient artefacts while preserving detection of sustained abnormal behaviour.
- **Operating-point selection based on costs.** The calibrated τ^* provides a principled default, but the final threshold should be chosen based on application requirements (e.g., prioritising very low false-positive rates to avoid stoppages, or prioritising high recall for safety-critical parts).
- **Traceability and continuous monitoring.** Logging frame-wise probabilities, alarms, and associated process metadata supports root-cause analysis and enables periodic model audits. In practice, monitoring for dataset drift (new materials, new joint types, sensor ageing) is as important as initial model accuracy.
- **Integration with quality workflows.** To maximise value, the system should connect to existing QA pipelines (traceability IDs, inspection reports, operator interfaces) and provide interpretable evidence (example frames, probability traces) rather than only a binary decision.

5.5 Future Research Directions

5.5.1 Technical improvements

Several modelling directions would likely improve robustness and reduce dependence on frame-wise thresholding:

- **Sequence-level models.** Incorporating temporal context directly (e.g., 3D CNNs, ConvLSTM/GRU, temporal attention, or Transformer-based video encoders) could improve detection of subtle anomalies and reduce boundary ambiguity.
- **Uncertainty and calibration.** Explicit uncertainty estimation (e.g., temperature scaling, Bayesian approximations, ensembles) could allow risk-aware decisions and help operators interpret low-confidence cases.
- **Self-supervised and semi-supervised learning.** Given the scarcity of labelled defect data, representation learning from large volumes of unlabelled normal sequences could improve generalisation and reduce annotation effort.
- **Explainability and localisation.** Methods such as saliency maps or class-activation mapping adapted to thermal imagery could highlight contributing regions, supporting trust and enabling more actionable diagnostics.

5.5.2 Application extensions

To broaden applicability and industrial relevance, future work should expand evaluation to additional processes and outputs:

- **Broader process coverage.** Validation on real welding scenarios across multiple processes (e.g., MIG/TIG/laser), materials, and joint types would establish external validity and identify domain-specific failure modes.
- **Multi-class and severity modelling.** Moving beyond binary decisions to defect categorisation and severity scoring would better align with inspection standards and downstream decision-making.
- **Spatial defect localisation.** Extending the approach to segmentation or weakly supervised localisation would support root-cause analysis and targeted rework rather than only pass/fail screening.

5.5.3 System enhancements

Finally, deployment-oriented engineering can make the approach more practical:

- **Edge optimisation.** Quantisation, pruning, and runtime optimisation would reduce latency and allow deployment on industrial PCs or embedded GPUs at camera frame rate.

- **Multi-modal sensing and fusion.** Combining thermography with complementary signals (visible light, acoustic emission, current/voltage, or position encoders) could improve robustness in environments where thermal measurements alone are unstable.
- **Industry 4.0 integration.** Linking model outputs to manufacturing execution systems (MES) and process databases would enable closed-loop quality analytics and long-term process improvement.

5.6 Broader Impact and Significance

This thesis contributes to the growing body of work that applies modern machine learning to non-destructive testing and process monitoring. Scientifically, it demonstrates a transparent, leakage-aware approach to modelling correlated thermal data and provides evidence that compact CNNs can learn defect-sensitive thermal representations that generalise across different acquisition dynamics. Practically, it motivates thermography-based monitoring as a pathway to earlier defect detection, reduced inspection effort, and improved process understanding. In the longer term, such systems can support safer products and more sustainable manufacturing by reducing scrap, rework, and energy waste through earlier intervention and continuous quality feedback.

Bibliography

[1] X. P. V. Maldague. *Theory and Practice of Infrared Technology for Nondestructive Testing*. Wiley Series in Microwave and Optical Engineering. New York: Wiley, 2001 (cit. on pp. 1, 8, 9, 11, 14, 16, 17, 21, 23, 30, 34, 39).

[2] R. Usamentiaga et al. "Infrared Thermography for Temperature Measurement and Non-Destructive Testing". In: *Sensors* 14.7 (2014), pp. 12305–12348 (cit. on pp. 1, 7–17, 21, 23, 34, 39).

[3] S. Lagüela et al. "Application of Infrared Thermography to the Analysis of Welding Processes". In: *QIRT Journal* (2012) (cit. on pp. 1, 8, 10, 12, 13, 17).

[4] B. Yousefi et al. "Application of Deep Learning in Infrared Nondestructive Testing". In: *Proc. Quantitative InfraRed Thermography (QIRT)*. 2018 (cit. on pp. 2, 12, 13, 18, 21).

[5] D. Buongiorno et al. "Inline Defective Laser Weld Identification Using Thermal Image Sequences and Deep Learning". In: *Applied Sciences* 12 (2022) (cit. on pp. 2, 10, 12, 21, 25, 36).

[6] G. Piecuch et al. "Diagnostics of Welding Process Based on Thermovision Images Using CNN". In: *IOP Conf. Ser.* 2019 (cit. on pp. 2, 12, 21).

[7] S. Kou. *Welding Metallurgy*. Second. Wiley, 2003 (cit. on p. 7).

[8] American Welding Society. *Welding Handbook, Vol. 1: Welding Science and Technology*. Ninth. AWS, 2011 (cit. on pp. 7, 8).

[9] N. Williams and J. Parker. "Review of Resistance Spot Welding of Steel Sheets—Part I". In: *International Materials Reviews* 49.2 (2004) (cit. on pp. 7, 11, 17).

[10] *Structural Welding Code – Steel*. Standard specification. American Welding Society, 2020 (cit. on p. 7).

[11] *Non-destructive testing of welds – Visual testing of fusion-welded joints*. International standard. International Organization for Standardization, 2016 (cit. on p. 8).

[12] G. C. Holst. *Electro-Optical Imaging System Performance*. 5th ed. Bellingham, WA: SPIE, 2008 (cit. on p. 9).

[13] L. Kästner et al. "Classification of Spot-Welded Joints in Laser Thermography Using CNNs". In: *IEEE Access* 9 (2021) (cit. on pp. 11, 12, 17, 20).

[14] S. Verspeek et al. “Spot Weld Inspections Using Active Thermography”. In: *Applied Sciences* (2022) (cit. on pp. 11, 12, 17).

[15] C. Hellier. *Handbook of Nondestructive Evaluation*. McGraw-Hill, 2001 (cit. on p. 11).

[16] D. Tran et al. “Learning Spatiotemporal Features with 3D Convolutional Networks”. In: *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*. 2015 (cit. on pp. 13, 14, 20).

[17] Y. Liu et al. “Stable 3D Deep Convolutional Autoencoder for Defect Detection”. In: *Materials* 17 (2024) (cit. on pp. 13, 14, 20).

[18] Q. Fang and X. Maldague. “Defect Depth Estimation in Infrared Thermography Using Deep Learning”. In: *Applied Sciences* (2020) (cit. on pp. 13, 15, 18).

[19] H. Bang et al. “Defect Identification Using Thermography and Deep Learning”. In: *Composite Structures* 246 (2020) (cit. on pp. 13, 15).

[20] S. Kapoor and A. Narayanan. “Leakage and the Reproducibility Crisis in Machine-Learning-Based Science”. In: *Patterns* 4.9 (2023), p. 100804. DOI: 10.1016/j.patter.2023.100804 (cit. on pp. 13, 22, 30, 35).

[21] S. Ioffe and C. Szegedy. “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift”. In: *Proc. Int. Conf. Mach. Learn. (ICML)*. 2015 (cit. on pp. 20, 31).

[22] K. He, X. Zhang, S. Ren, and J. Sun. “Deep Residual Learning for Image Recognition”. In: *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*. 2016 (cit. on pp. 20–23, 32).

[23] S. Elfwing, E. Uchibe, and K. Doya. “Sigmoid-Weighted Linear Units for Neural Network Function Approximation in Reinforcement Learning”. In: *Neural Networks* 107 (2018) (cit. on pp. 20, 31).

[24] P. Ramachandran, B. Zoph, and Q. V. Le. “Searching for Activation Functions”. In: *Proc. Int. Conf. Learn. Represent. (ICLR)*. 2018 (cit. on pp. 20, 31).

[25] D. P. Kingma and J. Ba. “Adam: A Method for Stochastic Optimization”. In: *Proc. Int. Conf. Learn. Represent. (ICLR)*. 2015 (cit. on pp. 20, 22, 23, 32, 35, 37).

[26] I. Loshchilov and F. Hutter. “Decoupled Weight Decay Regularization”. In: *Proc. Int. Conf. Learn. Represent. (ICLR)*. 2019 (cit. on p. 20).

[27] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. “Rethinking the Inception Architecture for Computer Vision”. In: *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*. 2016, pp. 2818–2826 (cit. on pp. 21, 23, 32).

[28] F. Yu and V. Koltun. “Multi-Scale Context Aggregation by Dilated Convolutions”. In: *International Conference on Learning Representations (ICLR)* (2016) (cit. on pp. 21, 23, 32).

[29] M. Lin, Q. Chen, and S. Yan. *Network In Network*. arXiv:1312.4400. 2013. DOI: 10.48550/arXiv.1312.4400. arXiv: 1312.4400 [cs.NE] (cit. on pp. 21, 22, 32).

[30] J. Hu, L. Shen, and G. Sun. "Squeeze-and-Excitation Networks". In: *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*. 2018 (cit. on pp. 21, 23, 32).

[31] I. Loshchilov and F. Hutter. "SGDR: Stochastic Gradient Descent with Warm Restarts". In: *Proc. Int. Conf. Learn. Represent. (ICLR)*. 2017 (cit. on pp. 22, 23, 32, 35, 37).

[32] T. Fawcett. "An Introduction to ROC Analysis". In: *Pattern Recognition Letters* 27.8 (2006), pp. 861–874 (cit. on pp. 22, 33, 38–40).

[33] J. Davis and M. Goadrich. "The Relationship Between Precision-Recall and ROC Curves". In: *Proc. Int. Conf. Mach. Learn. (ICML)*. 2006, pp. 233–240 (cit. on pp. 22, 33, 38–40).

[34] E. F. Schisterman, N. J. Perkins, A. Liu, and H. Bondell. "Optimal Cut-point and Its Corresponding Youden Index to Discriminate Individuals Using Biomarker Results". In: *Epidemiology* 19.5 (2008), pp. 651–655 (cit. on pp. 22, 33, 34, 41).

[35] G. C. Holst. *Electro-Optical Imaging System Performance*. 4th ed. Bellingham, WA: SPIE Press, 2017 (cit. on pp. 25, 30).

Dedications

*I dedicate this work to my **mother**, **sister**, and **brother**, whose love, patience, and constant support sustained me throughout this journey.*

*I am also deeply grateful to those who supported and guided me during the process of **immigration**, whose help made this path possible.*

*Finally, I thank my **friends** for their presence, companionship, and unwavering support throughout this experience.*