Measurement 290 (2026) 123016

Contents lists available at ScienceDirect

Measurement

journal homepage: www.elsevier.com/locate/measurement

Stable measurement framework for rheumatoid arthritis assessment using late fusion of thermal radiomics and lightweight convolutional representations

Agnieszka Pregowska [a], Konrad Pauk [b], Mikhail Ihnatouski [c], Jolanta Pauk [d,*]

[a] Institute of Fundamental Technological Research, Polish Academy of Sciences, Warsaw, Poland
[b] Medical University of Warsaw, Warsaw, Poland
[c] Scientific and Research Department, Yanka Kupala State University of Grodno, Grodno, Belarus
[d] Biomedical Engineering Institute, Bialystok University of Technology, Bialystok, Poland

ARTICLE INFO

Keywords:
Metrology
Infrared thermography
Sensor fusion
Measurement stability
Calibration
Radiomics
Rheumatoid arthritis

ABSTRACT

Reliable assessment of rheumatoid arthritis (RA) using infrared thermography requires a measurement framework capable of characterizing the dynamic thermal response to controlled excitation while limiting sensitivity to acquisition conditions, inter-subject variability, and computational instability. This study presents a measurement-oriented Active Dynamic Thermography (ADT) framework in which a dual-stream late-fusion stage combines radiomic descriptors with convolutional representations of the acquired thermal response. The measurand was defined as the spatio-temporal thermal response of the hand during post-stimulus rewarming, interpreted as an indirect functional indicator of superficial microvascular and subcutaneous vascular-bed activity associated with inflammatory processes. Controlled cold excitation provided a standardized thermal perturbation for observing vascular recovery dynamics, while radiometric verification and protocol standardization were used to reduce non-biological acquisition variability.

The proposed framework was evaluated on thermographic recordings acquired from 187 patients with clinically confirmed RA and 292 healthy controls. To assess the robustness of the diagnostic output to cohort partitioning, experiments were repeated across ten independent subject-disjoint realizations. FUSION achieved a mean accuracy of 91.0%, balanced accuracy of 91.4%, Matthews correlation coefficient (MCC) of 0.819, ROC-AUC of 0.954, and PR-AUC of 0.901. Compared with the evaluated standalone baselines, FUSION showed higher mean performance and lower descriptive inter-seed variability. Ranking and distributional analyses suggested improved stability. Paired Wilcoxon comparisons showed significant improvements over the standalone CNN baseline for several discriminative metrics after adjustment, whereas differences relative to the stronger INR baseline did not reach the adjusted significance threshold.

The results support the feasibility of a measurement-oriented ADT framework in which complementary thermal descriptors are integrated within a computational decision stage. The fusion component should therefore be interpreted as part of the broader thermal measurement chain rather than as the measurement principle itself. Further test-retest, inter-device, and external validation is required to establish physical repeatability, reproducibility, and clinical generalizability.

1. Introduction

The clinical management of rheumatoid arthritis (RA) requires early and repeated assessment of inflammatory activity before irreversible structural joint damage occurs [1]. Although ultrasound and magnetic resonance imaging (MRI) remain diagnostic benchmarks, their routine use for high-frequency monitoring is limited by cost, examination time, equipment availability, and operator dependence [2]. Infrared thermography (IRT) has therefore attracted interest as a non-invasive point-of-care (PoC) measurement alternative capable of capturing surface-temperature manifestations of metabolic and microcirculatory changes associated with inflammation [3–5].

The biological rationale for using IRT in autoimmune disease is that immune-mediated inflammation can alter local vascular regulation, heat transfer, and skin-temperature distribution [5,6]. In RA,

* Corresponding author.
E-mail address: j.pauk@pb.edu.pl (J. Pauk).

https://doi.org/10.1016/j.measurement.2026.123016
Received 12 February 2026; Received in revised form 24 August 2026; Accepted 26 August 2026
Available online 2 September 2026

persistent synovitis involves immune-cell infiltration, cytokine release, angiogenesis, increased perfusion, and altered vasomotor control [7,8]. Because the hand joints are superficial, these periarticular changes may be reflected in measurable thermal patterns at the skin surface. Active dynamic thermography (ADT) extends this approach by applying a controlled thermal stimulus and analyzing the subsequent rewarming phase [9]. The recovery response may reveal inflammation-related microvascular dysregulation; thus, IRT should be interpreted as a non-invasive functional proxy of vascular activity rather than as a direct image of the immune process [5,10]. In the present measurement framework, the measurand is defined as the spatio-temporal thermal response of the hand to controlled cold excitation, with particular emphasis on the post-stimulus rewarming behavior associated with superficial microvascular and subcutaneous vascular-bed activity. This response is observed indirectly through the calibrated skin-surface temperature field and is interpreted as a functional thermal correlate of inflammation-related vascular alterations rather than as a direct measurement of synovial inflammation.

Beyond RA, IRT has been investigated in biomedical and sports-medicine conditions in which surface-temperature patterns reflect disease- or injury-related functional changes, including diabetic foot monitoring, peripheral vascular disorders, breast pathology screening, musculoskeletal injury assessment, wound monitoring, systemic sclerosis, and Raynaud-related vascular dysfunction [3,5,6,11,12].

Translating these thermal responses into reliable diagnostic outcomes remains a metrological challenge. Diagnostic reconstruction from raw thermal maps is affected by heat-flow physics, acquisition noise, limited reproducibility, and non-stationary physiological fluctuations [13,14]. Existing thermal-analysis pipelines often reduce dynamic rewarming sequences to isolated static frames [10,15], thereby discarding transient and time-varying structures that characterize the physiological response to thermal excitation. A measurement-oriented framework should therefore preserve temporal information while prioritizing representation stability, calibration, and uncertainty awareness, consistent with recent perspectives on metrology for medical artificial intelligence (AI) [16,17].

Feature representation is central to this problem. Deep models can learn high-level latent abstractions from thermal images [18], and recent thermographic studies using prototypical and Siamese networks support representation learning in thermal-image interpretation [11]. Handcrafted radiomics, in contrast, provides predefined descriptors of intensity, texture, gradients, and spatial heterogeneity [19]. In this study, high-level latent abstractions refer to compact internal features automatically learned by neural networks, whereas handcrafted radiomics denotes predefined mathematical descriptors that provide interpretable measurements of thermal patterns.

To address these limitations, a late-fusion framework (hereafter FUSION) was developed in which learning-based and radiomic representations were treated as complementary components of the same measurement system. This design reduced reliance on purely black-box neural representations while preserving physically grounded descriptors. An implicit neural representation (INR) baseline was additionally evaluated as a controlled ablation to assess whether continuous thermal-field reconstruction could improve downstream descriptor stability [20,21].

The primary objective of this study was to develop and internally validate a computationally parsimonious measurement framework for RA assessment using active IRT. This objective was motivated by the need to convert dynamic thermal responses into diagnostic indicators that remain robust despite low thermal contrast, patient-level heterogeneity, environmental sensitivity, and limited clinical dataset size. The framework integrates complementary learned and handcrafted thermal representations through a lightweight linear decision stage and was evaluated using repeated subject-disjoint experiments and non-parametric win-count analysis. The study should be interpreted as an internally validated methodological feasibility study rather than as a fully validated clinical diagnostic system. Accordingly, the proposed approach is formulated as an end-to-end thermal measurement chain comprising controlled thermal excitation, calibrated infrared acquisition, standardized preprocessing, quantitative thermal representation, and computational decision support.

2. Materials and methods

2.1. Dataset characteristics and ethics

The study was approved by the Bioethics Committee of the University Clinical Hospital in Bialystok (No. R-I-002/16/2016), and all participants provided written informed consent. Patients were eligible for inclusion if they were older than 18 years, had clinically confirmed RA for more than 1 year, received biologic therapy either as monotherapy or in combination with methotrexate, and had no relevant concomitant diseases. Exclusion criteria for the RA group included age below 18 years, disease duration shorter than 1 year, glucocorticosteroid treatment, or treatment with non-biologic disease-modifying antirheumatic drugs (DMARDs). Biologic treatment was administered according to the Polish national therapeutic program B.33, which regulates access to reimbursed biologic and targeted synthetic DMARDs for eligible patients with active RA. The program defines uniform qualification and monitoring criteria, including insufficient response to previous conventional DMARD therapy and regular assessment of disease activity and safety. Thus, adherence to B.33 reflected routine clinical practice in Polish rheumatology centers and supported standardized patient qualification and treatment monitoring [22].

Healthy controls were screened to exclude RA, Still's disease, polymyalgia rheumatica, giant-cell arteritis, systemic lupus erythematosus, seronegative inflammatory spondyloarthropathies, other rheumatic diseases, cardiovascular or respiratory diseases, and dermatological disorders. The experimental dataset comprised $N = 479$ participants, including 187 patients with clinically confirmed RA and 292 healthy controls, with one bolometric ADT sequence acquired per participant. Because each sequence corresponded to a unique participant, participant-level partitioning ensured that no individual contributed data to more than one subset. When ROI-level or patch-level descriptors were derived from a sequence, all derived samples from the same participant were retained within the same training, validation, or test subset.

2.2. Data acquisition and active dynamic thermography protocol

Data acquisition was performed using a long-wave infrared (LWIR) camera (FLIR E60bx, 320 × 240 focal plane array, noise-equivalent temperature difference (NETD) < 0.045 °C, spectral range 8–12 μm). Although higher-resolution infrared cameras are currently available, this sensor provided calibrated LWIR measurements with sufficient thermal sensitivity for the controlled ADT protocol used in this study. A bolometric sequence was defined as a time-resolved series of LWIR frames acquired by the microbolometer detector during the complete protocol. Unlike a single static thermogram, each sequence encoded the temporal evolution of hand-surface temperature during baseline stabilization, cold-stress excitation, and post-stimulus rewarming [9]. The camera was factory-calibrated and validated before each measurement session against an external blackbody source with a calibration certificate compliant with the International Organization for Standardization/International Electrotechnical Commission (ISO/IEC) 17025 standard.

A blackbody source is an idealized radiometric reference with emissivity equal to one; in practice, calibrated blackbody devices approximate this behavior and provide traceable temperature references for infrared camera validation [4]. Before each session, the blackbody was stabilized at 30 °C, selected to approximate the temperature range encountered during thermographic measurements of the hand. The camera was positioned using the same working distance and acquisition geometry as during hand imaging, and the blackbody surface

Table 1
Metrological specifications, radiometric verification criteria, and standardized conditions of the ADT measurement protocol.

Category	Parameter	Value/Description
Sensor	Model	FLIR E60bx (LWIR focal plane array)
	Spatial resolution	320 × 240 pixels
	NETD	< 0.045 °C at 30 °C; indicator of detector thermal sensitivity
	Sampling frequency	$f_s = 1$ Hz
Radiometric verification	Reference source	External calibrated blackbody with an ISO/IEC 17025-compliant calibration certificate
	Verification temperature	30 °C
	Acceptance criterion	$\|\Delta T\| \leq 0.1$ °C between the measured blackbody temperature and the certified reference value
	Stability criterion	No systematic drift during a 5-min observation of the reference source
Radiometric settings	Skin emissivity	$\epsilon = 0.98$
	Reflected apparent temperature	Accounted for in the radiometric compensation procedure
Environment	Ambient temperature	23 ± 1 °C
	Relative humidity	55 ± 5%
	Working distance	1.0 m (fixed acquisition geometry)
ADT protocol	Baseline stability	$\sigma < 0.2$ °C over 120 s
	Thermal stimulus	Water–ice slurry at 0 ± 0.5 °C for 5 s
	Minimum induced temperature decrease	≥ 5 °C
	Recovery observation	180 s at 1 Hz (approximately 180 consecutive observations)

Abbreviations: ADT — Active Dynamic Thermography; LWIR — long-wave infrared; NETD — noise-equivalent temperature difference.

was placed in the central field of view. A short radiometric sequence was recorded, and the mean temperature within a central region of interest (ROI) was compared with the certified reference temperature. Acquisition was accepted only when the absolute deviation remained within $|\Delta T| \leq 0.1$ °C and no systematic drift was observed during the 5-minute stability window. Radiometric stability was defined as the absence of systematic drift during observation of the reference source.

Radiometric conversion accounted for reflected apparent temperature (T_{refl}) and environmental thermal background. The reflected apparent temperature denotes the apparent temperature assigned to infrared radiation reflected from the measured surface into the camera, whereas environmental thermal background refers to radiation emitted by surrounding objects and room surfaces. These quantities are important because an infrared camera detects not only skin-emitted radiation, but also reflected and background components reaching the detector [3,4].

The ambient temperature was monitored in real time (23 ± 1 °C), and its influence was compensated within the camera's internal measurement equation to isolate the target's self-emission from parasitic reflections.

Controlled laboratory conditions included a fixed camera-to-hand distance of 1.0 m, standardized hand positioning, participant acclimatization, absence of direct solar radiation, minimized air movement, uniform indoor illumination, and shielding of high-intensity infrared sources. Under these conditions, the background was treated as a diffuse radiator, reducing directional thermal artifacts. Non-biological experimental variance was defined as variability introduced by the measurement and computational process rather than by disease-related physiology. It included thermal sensor noise, camera thermal sensitivity, environmental temperature and humidity fluctuations, hand positioning, segmentation uncertainty, preprocessing operations, stochastic data partitioning, random model initialization, and optimization-related variability.

The metrological characteristics of the measurement system, calibration procedure, environmental conditions, and thermal challenge protocol are summarized in Table 1.

Skin emissivity describes the efficiency with which skin emits infrared radiation relative to an ideal blackbody at the same temperature. The value $\epsilon = 0.98$ has been widely adopted in biomedical thermography since the early radiometric work of Steketee, who reported human skin as a high-emissivity surface in the infrared range [23]. Therefore, $\epsilon = 0.98$ was used in the present study, consistent with established practice for clean human skin in LWIR thermography [3,4,23]. Under opaque-surface assumptions, this corresponds to an approximate reflected component of $1 - \epsilon = 0.02$ in the radiometric temperature calculation.

The ADT protocol consisted of three consecutive phases recorded continuously at a sampling rate of $f_s = 1$ Hz. The first phase was baseline stabilization, during which the participant's hand remained in a standardized position under controlled laboratory conditions until the temperature at the thenar eminence reference region reached a stable pre-stimulus state. Stability was defined as $\sigma < 0.2$ °C over a 120 s observation window, where σ denotes the standard deviation of the temperature signal. This criterion was used to reduce short-term vasomotor fluctuations, residual acclimatization effects, and involuntary movement before thermal stimulation [3,4]. The second phase was cold-stress excitation, in which the hand was immersed for 5 s in a water-ice slurry maintained at 0 ± 0.5 °C. A minimum skin-temperature decrease of at least 5 °C within the monitored hand region was required to confirm sufficient thermal excitation for recovery-phase analysis. This controlled stimulus was intended to increase the thermal gradient between the skin and environment and enhance the observability of vascular recovery dynamics. The third phase was post-stimulus recovery, during which the hand was repositioned under the LWIR camera and recorded for 180 s to monitor rewarming kinetics [9]. After immersion, the skin was gently dried using a standardized non-abrasive procedure to maintain constant emissivity and reduce evaporative cooling artifacts without inducing friction-related thermogenesis. Representative thermographic frames obtained during the baseline and recovery phases are presented in Fig. 2.

The rationale behind ADT was that inflammatory tissues may exhibit altered vascular reactivity and rewarming kinetics compared with healthy tissue. In contrast to conventional thermography, which captures a steady-state temperature distribution, ADT evaluates the transient response after controlled thermal excitation. This active measurement strategy was intended to reveal recovery abnormalities that may remain less visible under equilibrium conditions. From a measurement perspective, the controlled excitation provides a standardized input to the thermophysiological system, so that the resulting rewarming trajectory can be evaluated as a response to a defined perturbation rather than as an uncontrolled steady-state temperature pattern. This is intended to improve observability of dynamic vascular behavior and reduce the dependence of the assessment on a single passive temperature measurement. The pre-stimulus stabilization phase ensured that the dynamic recovery measurement was initiated from a predefined

thermal stability condition, thereby improving protocol standardization without assuming elimination of inter-subject baseline differences.

Quantitatively, the protocol required a minimum post-excitation skin-temperature decrease of at least 5 °C, corresponding to more than 110 times the specified detector NETD of < 0.045 °C. Thus, the imposed thermal perturbation produced a dynamic temperature excursion substantially larger than the nominal thermal sensitivity of the detector. In addition, the 180-s recovery phase acquired at $f_s = 1$ Hz provided approximately 180 consecutive observations of the evolving thermal field, enabling characterization of the rewarming trajectory rather than reliance on a single steady-state temperature measurement. Together, the controlled excitation and time-resolved acquisition increase the information available for observing transient vascular recovery dynamics. However, because the present study did not include a dedicated repeated-acquisition comparison between ADT and passive thermography, this increased observability should not be interpreted as direct experimental evidence of superior metrological repeatability.

2.3. Metrological characteristics and sources of uncertainty

The metrological interpretation of the proposed framework distinguishes between uncertainty associated with physical thermal acquisition and variability introduced by computational processing. At the acquisition level, relevant contributions include detector thermal sensitivity, radiometric verification, emissivity assumptions, reflected apparent temperature, ambient-temperature variation, camera-to-hand geometry, positioning, physiological variability, and the consistency of the cold-stress excitation. These factors may influence the observed skin-surface temperature field and therefore propagate into the derived thermal descriptors.

Calibration-related effects were addressed through session-wise verification against the certified blackbody reference, fixed emissivity, standardized geometry, and controlled environmental conditions. Protocol-related variability was reduced through the predefined baseline-stability criterion, standardized cold-stress temperature and duration, minimum required temperature decrease, and fixed recovery window.

Because only one ADT sequence was acquired per participant, physical measurement repeatability under repeated acquisitions in the same individual could not be estimated directly in the present dataset. Accordingly, the repeated subject-disjoint experiments characterize computational stability and robustness to cohort partitioning rather than metrological repeatability of the thermal acquisition itself. Likewise, inter-device and inter-site reproducibility could not be quantified because all measurements were obtained at a single center using the same LWIR camera.

Inter-subject variability was addressed through subject-disjoint partitioning and repeated evaluation across independently generated cohort splits. This design quantifies the sensitivity of the analytical output to changes in cohort composition while preventing participant-level information leakage. A complete measurement-uncertainty budget and dedicated test–retest and inter-device studies remain necessary for full metrological validation.

To provide a structured metrological characterization of the proposed measurement framework, Table 2 summarizes the assessment of the measurand, radiometric verification, repeatability, reproducibility, robustness, probabilistic calibration, measurement uncertainty, and inter-subject variability.

2.4. Computational workflow and feature extraction

The computational workflow was implemented as a dual-stage analytical pipeline integrating MATLAB and Python.

MATLAB was used for deterministic image preprocessing and anatomical localization, including grayscale normalization, empirical gamma correction, balanced histogram thresholding (BHT), binary mask refinement, Zhang–Suen skeletonization, and repeatable localization of joint-centered ROIs. These operations were selected because MATLAB provides robust matrix-based image-processing routines and convenient visualization tools for verifying segmentation quality, skeleton topology, and anatomical ROI placement.

Python was used for multi-stream feature extraction, model training, late-fusion modeling, and statistical evaluation. Radiomic features were computed using NumPy/SciPy and scikit-image, while wavelet-based descriptors were extracted using PyWavelets. The convolutional stream and INR components were implemented in PyTorch, enabling tensor-based computation and graphics processing unit (GPU)-accelerated feature extraction. The fused feature vectors were subsequently processed using scikit-learn routines for logistic regression, calibration assessment, and performance evaluation. This separation allowed deterministic anatomical preprocessing to be performed before Python-based feature extraction, modeling, and validation.

Preprocessing began with grayscale normalization followed by nonlinear intensity mapping using empirical gamma correction ($\gamma = 1.2$). Gamma correction was applied to enhance low-contrast thermal gradients in the hand region, particularly at the boundaries between cooler background areas and warmer periarticular regions. The value $\gamma = 1.2$ was selected empirically as a mild contrast-enhancement setting that improved the visibility of diagnostically relevant thermal transitions without saturating high-temperature regions or distorting the relative temperature distribution. Gamma-based intensity transformation is commonly used in image processing to modify local contrast through a controlled non-linear mapping [24].

After contrast adjustment, BHT was used to segment the hand from the thermal background. BHT is a histogram-based thresholding method that iteratively balances the gray-level histogram mass on both sides of a candidate threshold until an equilibrium point is reached [25]. Applied to the normalized thermal image, BHT generated a binary hand mask in a deterministic and reproducible manner. This segmentation reduced background influence, stabilized the anatomical search space, and provided a consistent input for subsequent Zhang–Suen skeletonization and joint-centered ROI localization. The use of robust segmentation is consistent with recent work in medical thermal-image analysis, where hotspot localization and spatially informed clustering have been emphasized for extracting clinically meaningful thermal regions [26].

Zhang–Suen skeletonization was then applied to the binary hand mask to obtain a one-pixel-wide, topology-preserving representation of the hand structure [27]. This thinning algorithm iteratively removes boundary pixels according to connectivity-preserving rules while retaining the global shape and branch organization of the segmented hand. The resulting skeleton served as a geometric reference for repeatable localization of joint-centered ROIs, particularly around the metacarpophalangeal and proximal interphalangeal regions. This step reduced variability related to hand size, posture, and minor positioning differences between participants. The resulting anatomical localization procedure is illustrated in Fig. 1.

From the localized ROIs, two complementary feature streams were generated. The radiomics stream extracted handcrafted descriptors of intensity distribution, texture, local gradients, and multiscale thermal variation, whereas the convolutional stream extracted learned spatial embeddings from standardized thermal patches. The radiomic descriptors included first-order intensity statistics, Haralick texture features derived from the gray-level co-occurrence matrix (GLCM), local binary patterns (LBP), gradient-based statistics, and frequency- or wavelet-domain features. First-order features summarized thermal intensity distributions; Haralick descriptors quantified spatial gray-level relationships such as contrast, homogeneity, energy, and correlation [28]; LBP captured local micro-texture variations [29]; gradient-based descriptors characterized the sharpness and directionality of thermal transitions; and frequency- or wavelet-domain descriptors represented multiscale

Table 2
Metrological characteristics of the proposed ADT measurement framework and their assessment in the present study.

Characteristic	Assessment in the present study	Interpretation/limitation
Measurand	Spatio-temporal skin-surface temperature response during post-stimulus rewarming following standardized cold excitation	Physical thermal response interpreted as an indirect functional indicator of inflammation-related vascular activity
Radiometric verification	Session-wise comparison with a calibrated blackbody reference at 30 °C; acceptance criterion $\|\Delta T\| \leq 0.1$ °C	Provides verification of radiometric performance under the acquisition conditions used in the study
Detector thermal sensitivity	NETD < 0.045 °C	Instrument specification; not equivalent to measurement uncertainty
Physical repeatability	One complete ADT acquisition was available per participant	Within-subject test–retest repeatability was not directly quantified
Reproducibility	All measurements were acquired at one center using one LWIR camera	Inter-device and inter-site reproducibility was not established
Acquisition robustness	Standardized ambient conditions, fixed acquisition geometry, predefined baseline stability, controlled thermal stimulus, and fixed recovery window	Protocol standardization was used to reduce non-biological acquisition variability
Computational robustness	Ten independent subject-disjoint train–validation–test realizations	Quantifies sensitivity of the diagnostic output to cohort partitioning and computational stochasticity, not physical measurement repeatability
Inter-subject variability	Subject-disjoint evaluation across the heterogeneous clinical cohort	Included in the internal robustness assessment; subgroup-specific effects were not separately quantified
Probabilistic calibration	ECE, Brier score, and LogLoss evaluated across repeated realizations	Characterizes calibration of the diagnostic output, not radiometric calibration of the infrared camera
Measurement uncertainty	Relevant instrumental, environmental, geometric, physiological, and protocol-related sources were identified	A complete combined measurement-uncertainty budget was not estimated from the available dataset

Abbreviations: ADT — Active Dynamic Thermography; LWIR — long-wave infrared; NETD — noise-equivalent temperature difference; ECE — expected calibration error.

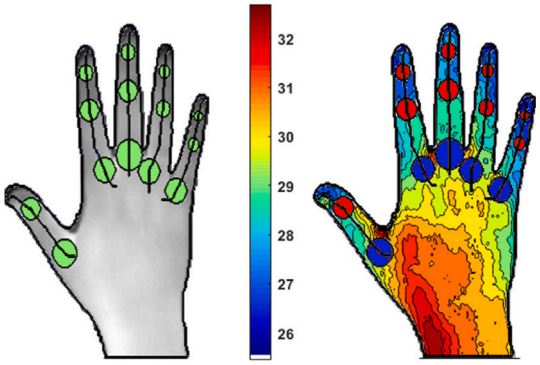


Fig. 1. Anatomical ROI localization using Zhang-Suen skeletonization. The skeletonized hand structure (left) provides a topology-preserving geometric reference for automated identification of metacarpophalangeal (MCP) and proximal interphalangeal (PIP) joint regions (right). These ROIs were used for thermographic feature extraction and recovery-phase analysis, enabling spatially consistent assessment of inflammation-related thermal dynamics across subjects.

spatial information [30]. These handcrafted features provided interpretable and reproducible measures of thermal heterogeneity associated with periarticular inflammatory processes, complementing the learned convolutional representations [19].

A late-fusion strategy was used to integrate the two complementary feature streams after independent extraction, following established multimodal machine-learning principles for combining heterogeneous representations [31]. Radiomics provided interpretable and physically grounded descriptors, whereas convolutional embeddings captured nonlinear spatial patterns that were difficult to define manually. The fused feature vector was constructed by concatenating the radiomic descriptors with the EfficientNet-B0 convolutional embedding and was subsequently processed by a lightweight linear classifier implemented as L_2-regularized logistic regression. This design allowed the classifier to operate on a compact joint representation, reduced the risk of dominance by the high-dimensional convolutional feature space, and supported representation stability under stochastic subject-level partitioning. The classifier produced class membership probabilities used for threshold optimization and performance evaluation.

EfficientNet-B0 was selected as the convolutional backbone because it provides an effective balance between predictive performance, parameter efficiency, and computational cost. Through compound scaling of network depth, width, and input resolution, EfficientNet architectures achieve competitive feature extraction performance while maintaining a relatively small computational footprint [32]. These characteristics are particularly advantageous for thermographic measurement systems intended for resource-constrained edge-computing and PoC deployment.

The reliability of the resulting feature vectors was supported by standardization of the measurement environment. Control of skin emissivity ($\epsilon = 0.98$), fixed camera-to-hand geometry, and constant acquisition conditions was intended to reduce non-biological experimental variance and improve the consistency of the extracted thermal features. These features should be interpreted as indirect functional indicators of inflammation-related thermal heterogeneity rather than as direct measurements of synovial immunopathology.

2.5. Computational components of the measurement chain

This section describes the evaluated model architectures from a system-identification perspective, emphasizing their role as operators acting on thermal signal representations rather than as purely data-driven classifiers. By treating the rewarming sequence as a dynamic response of a biological system to a thermal input, the framework aims to characterize stable thermal signatures associated with RA-related inflammation. Central to this approach is the fusion processing stage of FUSION, which integrates two complementary representational streams: a convolutional pathway for hierarchical spatial patterns and a radiomic pathway for structured biophysical descriptors. The architecture was designed to remain computationally lightweight while preserving representation stability in small-sample clinical regimes.

Let $\mathbf{z}_{\text{cnn}} \in \mathbb{R}^{d_{\text{cnn}}}$ denote the feature embedding extracted by the EfficientNet-B0 convolutional branch, where d_{cnn} represents the dimensionality of the learned convolutional feature space [32]. Let $\mathbf{z}_{\text{rad}} \in$

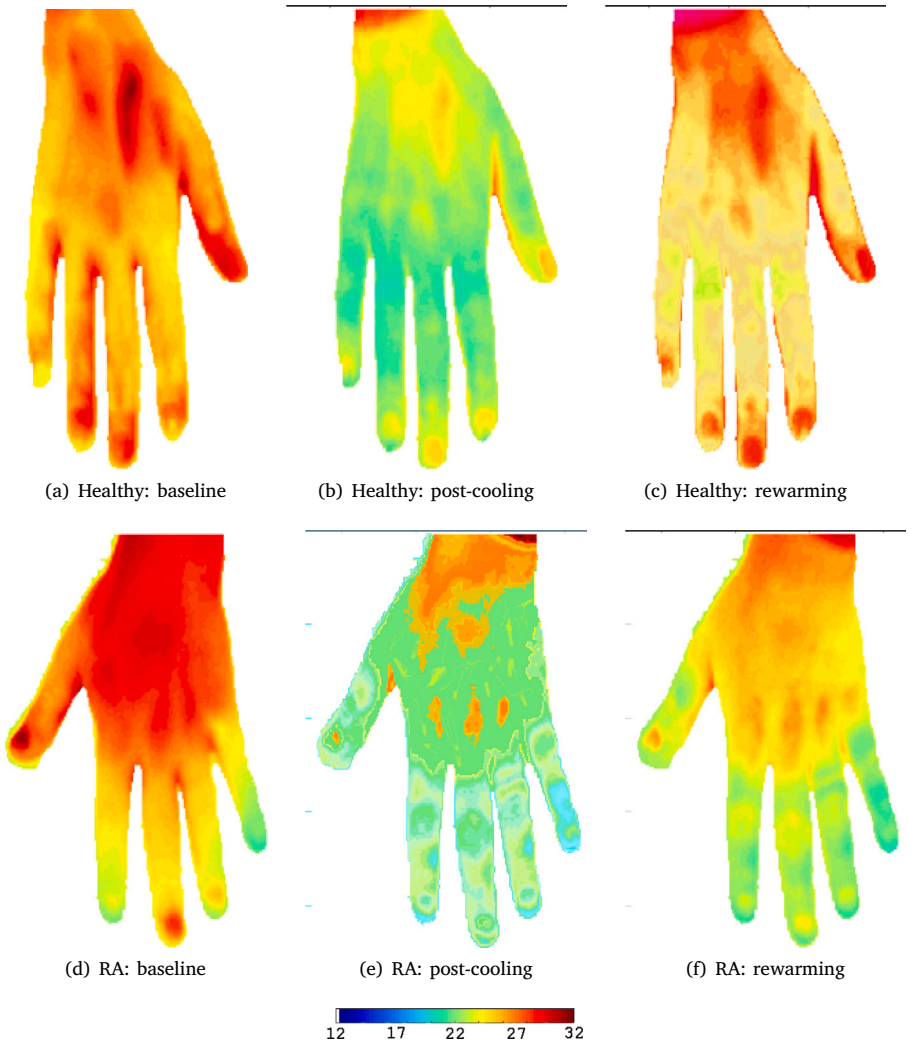


Fig. 2. Representative false-color thermographic images acquired during the ADT protocol. The upper row shows a healthy control participant, whereas the lower row shows a patient with RA. Images correspond to baseline, immediately post-cooling, and rewarming phases of the measurement protocol. All thermograms are displayed using the same false-color temperature scale to facilitate visual comparison of thermal recovery dynamics. The accompanying color bar indicates the corresponding temperature range in degrees Celsius.

$\mathbb{R}^{d_{\text{rad}}}$ denote the handcrafted radiomic feature vector, where d_{rad} corresponds to the dimensionality of the radiomic descriptor space [19]. The fused representation was defined by feature-vector concatenation:

$$\mathbf{z} = [\mathbf{z}_{\text{cnn}}; \mathbf{z}_{\text{rad}}] \tag{1}$$

The shallow probing head was intentionally designed as a lightweight decision stage operating on the fused CNN–radiomics representation. Rather than learning additional deep nonlinear transformations, its role was to evaluate the discriminative information already encoded within the convolutional embeddings and radiomic descriptors. In this context, deeper nonlinear heads refer to classification modules composed of multiple fully connected layers and nonlinear activation functions stacked on top of the feature representation. Although such architectures may increase model capacity, they also introduce additional trainable parameters, computational complexity, and energy consumption. The proposed shallow linear probing head was therefore adopted to reduce model complexity, limit overfitting, and support computationally efficient deployment in resource-constrained measurement systems.

To isolate the specific contribution of continuous signal modeling, an INR configuration was evaluated separately as a controlled ablation baseline. This baseline interpreted the discrete spatio-temporal sequence as a continuous coordinate-dependent thermal field and allowed the effect of continuous regularization to be assessed independently from the CNN–radiomics fusion strategy.

In this context, continuous regularization refers to representing the thermal field as a smooth coordinate-dependent function rather than as a collection of discrete pixel values. By enforcing continuity, the INR suppresses high-frequency sampling artifacts, local sensor noise, and small spatial perturbations that may arise during acquisition. Consequently, neighboring spatial locations are constrained to exhibit consistent thermal behavior, leading to smoother reconstructed temperature fields and more stable downstream descriptors. From a measurement perspective, this regularization reduces sensitivity to stochastic fluctuations and may improve the stability of extracted features across independent computational realizations [20,21].

Following the principles of linear classifier probes [33], the parsimonious decision stage was used to evaluate whether the fused CNN–radiomics representation contained sufficient discriminative information without requiring deeper nonlinear classification heads.

2.5.1. Convolutional neural network

A convolutional neural network (CNN) can be interpreted as a learned multiscale spatial filter that extracts local and hierarchical structures from thermal images [34,35]. This interpretation links CNN-based feature extraction with classical signal-decomposition approaches, while allowing the filters to be adapted to the data.

The convolutional stream served as a learned spatial representation operator. Given a preprocessed thermal image $\mathbf{X} \in \mathbb{R}^{H\times W}$, the CNN applied successive convolutional filters and nonlinear activation functions to generate feature maps:

$$\mathbf{F}_i = \sigma(\mathbf{X} * \mathbf{K}_i + b_i), \tag{2}$$

where $\mathbf{K}_i$ denotes a learnable kernel, b_i is a bias term, and $\sigma(\cdot)$ is a nonlinear activation function.

Successive layers progressively increased the effective receptive field, enabling the network to capture local temperature gradients, spatial correlations, and higher-level thermal patterns. This hierarchical representation was used as a compact convolutional embedding for subsequent late fusion with radiomic descriptors.

Although CNN-based features provide expressive spatial information, they are data-driven and may be sensitive to stochastic training and limited-sample variability. Therefore, the CNN stream was used primarily as a feature extractor, and the final decision stage was constrained to a lightweight linear head. This design retained the spatial representational capacity of the convolutional branch while limiting model complexity and reducing the risk of overfitting in the subject-disjoint evaluation setting.

2.5.2. Implicit neural representation

A coordinate-based INR was evaluated as a controlled ablation baseline, with each thermal frame modeled as a continuous function of spatial coordinates. Specifically, a lightweight multilayer perceptron (MLP) was fitted independently for each sample to learn

$$f_\theta : \mathbf{p} \in \mathbb{R}^2 \mapsto \hat{T}(\mathbf{p}), \tag{3}$$

where $\mathbf{p}$ denotes spatial coordinates and $\hat{T}(\mathbf{p})$ is the reconstructed thermal intensity. This formulation follows recent advances in coordinate-based neural fields, which represent signals as continuous functions rather than discrete samples [20,21,34].

Radiomic descriptors were subsequently computed from the reconstructed implicit field using the same deterministic extraction routine as in the primary radiomics stream, including GLCM, LBP, gradient-based, and frequency-domain features. To apply these discrete operators, the continuous field $\hat{T}(\mathbf{p})$ was resampled onto a standardized grid. This procedure allowed the INR baseline to be evaluated under the same downstream feature-extraction and classification protocol as the main radiomics pipeline.

In this configuration, the INR served as a continuous signal reconstruction stage rather than as a separate discriminative classifier. Its purpose was to assess whether the continuity-based regularization described above improved the stability of downstream descriptors. However, because the INR was fitted independently per sample, it did not provide the global inductive bias or hierarchical spatial abstraction characteristic of the convolutional stream.

2.5.3. Linear probing baseline (CNN_LR)

The CNN_LR model was evaluated as a controlled baseline combining learned convolutional features with a linear classifier. In this configuration, the CNN branch generated a latent feature vector $\mathbf{z}_{\text{cnn}}$, which was mapped directly to class probabilities without additional nonlinear classification layers. This design reduced the effective model complexity and limited the risk of overfitting in the limited-sample setting.

Because the underlying representation remained entirely data-driven, the CNN_LR model could still exhibit sensitivity to subject-level partitioning and lacked the explicit biophysical interpretability provided by handcrafted radiomic descriptors. Linear probing is a well-established strategy for evaluating the discriminative quality of learned representations [33]. In this study, CNN_LR was included as a secondary descriptive linear-probing reference to assess whether convolutional spatial embeddings alone were sufficient without radiomic late fusion. It was not included in the primary inferential comparison family, which was restricted to CNN, INR, and FUSION.

2.5.4. Multimodal late fusion and decision logic

The proposed FUSION framework integrates complementary signal representations by combining learned convolutional features with handcrafted radiomic descriptors at the feature level. The multimodal late-fusion stage operates on the fused representation defined in Eq. (1), where convolutional and radiomic feature embeddings are concatenated prior to classification.

The diagnostic output was generated by applying a lightweight linear classifier to the concatenated feature vector $\mathbf{z}$ to estimate class membership probabilities. The two streams contributed complementary information — learned spatial embeddings and structured radiomic descriptors — which were combined while limiting dependence on either representation alone.

From a measurement perspective, this design was intended to reduce sensitivity to stochastic data partitioning and to improve representation stability in a limited-sample clinical setting. By constraining the fusion head to a compact linear decision stage, the framework limited model complexity, reduced the risk of overfitting, and preserved computational parsimony. This configuration provided the basis for evaluating whether CNN–radiomics late fusion improved stability and performance relative to the reported baseline architectures. The computational configuration and principal hyperparameter settings used in the evaluated framework are summarized in Table 3.

2.6. Statistical performance characterization

For rigorous characterization of the performance reliability of the proposed framework, model performance was treated as a stochastic variable dependent on the data distribution. Let D denote the experimental dataset, and let $\{S_k\}_{k=1}^{K}$ be a set of $K = 10$ independent random subject-level partitions of D into disjoint training and test subsets. For each model architecture $\mathcal{M}$ and each partition S_k, a multi-metric performance vector was computed:

$$\mathbf{m}_{\mathcal{M}}^{(k)} = \left[\text{ACC}, \text{BACC}, \text{MCC}, \text{PR-AUC}, \text{ROC-AUC}, \text{Brier}, \text{ECE}, \text{LogLoss}\right]^{\top}. \tag{4}$$

This formulation enabled empirical estimation of performance variability induced by stochastic subject-level partitioning. Unlike conventional single-split evaluations, the multi-metric approach characterized both discriminative performance and probabilistic calibration. The inclusion of Expected Calibration Error (ECE), Brier score, and LogLoss provided information on the reliability of predicted probabilities, complementing threshold-based classification metrics. Such calibration metrics are widely used to assess whether predicted probabilities of modern neural networks match observed outcome frequencies [36]. From a measurement perspective, treating model outputs as repeated stochastic observations allowed calibration and uncertainty-related behavior to be assessed beyond isolated classification accuracy.

2.7. Stability and ranking consistency criteria

To assess whether the observed performance gains were systematic rather than incidental, the Wilcoxon signed-rank test was performed on paired metric differences across repeated realizations. For the primary inferential analysis, the comparison family comprised CNN, INR, and FUSION across the five primary discriminative metrics. CNN_LR was

Table 3
Computational configuration and hyperparameter specifications for the evaluated processing framework.

Category	Parameter	Value
Preprocessing	Image size	64×64
	Intensity normalization	$[0, 1]$
	Gamma correction	$\gamma = 1.2$
	Segmentation	Balanced Histogram Thresholding
	Skeletonization	Zhang-Suen thinning
TRUE INR baseline	Representation	coordinate-based MLP fitted per sample
	Input	(x, y) coordinates (+ optional Fourier features)
	Output	reconstructed thermal field $\hat{T}(x, y)$
	Downstream	radiomics computed from $\hat{T}$ + logistic regression
Radiomics	GLCM gray levels	16
	GLCM distances	$\{1, 2, 3\}$ pixels
	GLCM angles	$0, \pi/4, \pi/2, 3\pi/4$
	LBP points (P)	8
	LBP radius (R)	1
	FFT radial bands	6
	Wavelet type	Daubechies-2 (db2)
	Wavelet decomposition level	2
CNN architecture	Backbone	EfficientNet-B0
	Input channels	1 (grayscale, replicated if required)
	Output embedding	d_{cnn} (as implemented)
	Regularization	Dropout (as implemented)
	Training	End-to-end fine-tuning
Training	Optimizer	AdamW
	Learning rate	10^{-3}
	Weight decay	10^{-4}
	Batch size	64
	Max epochs	60
	Early stopping patience	10 epochs
Classification	Classifier	Linear probing head
	Architecture	LayerNorm + Dropout + FC
	Output layer	Binary classification
	Threshold optimization criteria	ACC/MCC/BACC/F1/Youden
	Threshold grid size	501
	Class weighting	Balanced
Evaluation	Train/val/test split	70%/15%/15%
	Random seeds	10
	Metrics	ACC, BACC, MCC, ROC-AUC, PR-AUC
	Calibration metrics	Brier, ECE, LogLoss
	Statistical test	Wilcoxon signed-rank

Abbreviations: ADT — Active Dynamic Thermography; GLCM — Gray-Level Co-occurrence Matrix; LBP — Local Binary Pattern; FFT — Fast Fourier Transform; CNN — Convolutional Neural Network; INR — Implicit Neural Representation; ACC — Accuracy; BACC — Balanced Accuracy; ROC-AUC — Area Under the Receiver Operating Characteristic Curve; PR-AUC — Area Under the Precision-Recall Curve; ECE — Expected Calibration Error.

retained as a secondary descriptive linear-probing reference and was not included in this inferential comparison family.

In this context, the ten independent random realizations referred to ten repeated training and evaluation runs generated using different random seeds, not to ten patients. In each realization, the cohort was repartitioned at the subject level, ensuring that all bolometric sequences from a given participant were assigned exclusively to one data subset.

For each metric, paired differences were computed as

$$\Delta m^{(k)} = m^{(k)}_{\text{FUSION}} - m^{(k)}_{\text{baseline}}, \tag{5}$$

where $m^{(k)}_{\text{FUSION}}$ and $m^{(k)}_{\text{baseline}}$ denote the metric values obtained by the FUSION and the baseline model in realization k, respectively. The metric vector included accuracy (ACC), balanced accuracy (BACC), Matthews correlation coefficient (MCC), precision–recall area under the curve (PR-AUC), receiver operating characteristic area under the curve (ROC-AUC), Brier score, ECE, and LogLoss. The analysis was performed separately for ACC, BACC, MCC, ROC-AUC, and PR-AUC across the ten subject-disjoint realizations, with adjusted p-values reported for multiple pairwise comparisons.

Because only ten independent realizations were available, the Wilcoxon signed-rank test had limited statistical power. Consequently, some pairwise comparisons did not reach the conventional significance threshold ($p < 0.05$). Therefore, statistical testing was complemented by non-parametric win-count analysis, which quantified ranking stability across repeated subject-disjoint partitions. Ranking- and count-based comparisons are established tools for evaluating classifier stability across repeated experimental settings [37,38]. A win was assigned to the model achieving the best value of a given metric in each realization. Higher values were considered better for discriminative metrics such as ACC, MCC, ROC-AUC, and PR-AUC, whereas lower values were preferred for error-based metrics such as ECE and Brier score.

Formally, for discriminative metrics, the winning model in realization k was defined as

$$M^*_k = \arg\max_M m^{(k)}_M. \tag{6}$$

For error-based metrics, the corresponding criterion was

$$M^*_k = \arg\min_M m^{(k)}_M. \tag{7}$$

The final win count was computed as the number of realizations in which each model obtained the best metric value. This statistic quantified how consistently a model ranked first under stochastic subject-level partitioning and provided information complementary to mean performance and Wilcoxon hypothesis testing.

Together, the Wilcoxon signed-rank test and win-count analysis provided a dual characterization of model performance: a formal paired comparison of metric differences and descriptive evidence for ranking

consistency. Because several Wilcoxon comparisons did not reach statistical significance, the stability advantage of FUSION was interpreted descriptively rather than as statistically confirmed superiority.

2.8. Distributional analysis and robustness assessment

Rather than relying solely on point estimates of expected performance, the proposed framework was evaluated through the empirical distributions generated across the set of stochastic realizations $\{S_k\}$.

Distributional robustness was assessed using boxplot-based summaries of repeated subject-disjoint realizations, including the median, interquartile range (IQR), distributional outliers, and lower-tail behavior, following standard exploratory data-analysis principles [39,40]. Reduced dispersion and compact IQR were interpreted as indicators of robustness against sampling variability and patient-level heterogeneity. Boxplot visualizations of the primary metrics, specifically ACC and MCC, were used to characterize these effects.

Tail-distribution analysis was further used to assess susceptibility to worst-case data partitions. In the present study, catastrophic performance failures refer to unusually poor realizations resulting from unfavorable subject-disjoint partitions, where diagnostic performance deteriorates substantially relative to the median behavior observed across repeated experiments. Such failures are reflected by lower-tail observations, extreme outliers, or pronounced reductions in ACC, MCC, ROC-AUC, and PR-AUC. Therefore, robustness was evaluated not only through average performance but also through the frequency and magnitude of lower-tail events. Reduced lower-tail dispersion and the absence of severe outliers were interpreted as indicators of improved resistance to performance degradation.

2.9. Software implementation and validation rigor

The computational pipeline was implemented in Python using PyTorch for neural-network development and optimization, scikit-learn for statistical learning and performance evaluation, NumPy/SciPy for numerical computation, scikit-image for image processing and radiomic feature extraction, and PyWavelets for discrete wavelet transform (DWT) descriptors. Python was selected because of its mature scientific computing ecosystem and widespread use in reproducible biomedical image analysis.

DWT descriptors were included to characterize thermal patterns at multiple spatial scales. By decomposing thermographic images into frequency subbands, DWT enabled simultaneous representation of global temperature distributions and localized thermal irregularities associated with tissue heterogeneity. The Daubechies 2 (db2) mother wavelet was selected because of its compact support, favorable localization properties, and computational efficiency, making it suitable for representing both smooth thermal gradients and localized temperature variations. To reduce evaluation bias, all experiments were repeated across 10 independent random seeds. Stratified subject-level partitioning was applied to the training, validation, and test subsets using a 70:15:15 ratio, ensuring that class distributions remained balanced while preventing sequences from the same participant from appearing in more than one subset. In the present study, Open Science principles refer to transparency, reproducibility, traceability, and independent verification of scientific results. Accordingly, preprocessing procedures, feature extraction routines, model architectures, hyperparameters, random seed configurations, evaluation protocols, and statistical analysis procedures were documented to facilitate reproducibility and future external validation.

The reported results were internally validated using repeated subject-disjoint train–validation–test partitions across 10 independent realizations. Evaluation included discrimination metrics (ACC, BACC, MCC, ROC-AUC, PR-AUC), calibration metrics (ECE, Brier score, LogLoss), inter-seed variability analysis, win-count statistics, and non-parametric significance testing. This multi-level strategy assessed predictive performance, computational stability, robustness to subject-level partitioning, calibration quality, and uncertainty-related behavior.

Table 4
Baseline demographic profile of the study cohort. Continuous variables are reported as mean ± SD.

Variable	Patients with RA	Healthy controls
Age (years)	51.7 ± 10.6	52.1 ± 10.2
Body mass (kg)	73.2 ± 10.9	76.3 ± 10.2
Body height (m)	1.75 ± 0.13	1.73 ± 0.11
BMI (kg/m^2)	25.5 ± 2.13	25.4 ± 2.7
Disease duration (years)	8.3 ± 4.6	N/A

Abbreviations: RA — rheumatoid arthritis; BMI — body mass index; SD — standard deviation; N/A — not applicable.

Table 5
Clinical and biochemical characteristics of the study cohort. Values are presented as mean ± SD unless otherwise indicated.

Parameter	Patients with RA	Healthy controls
ESR (mm/h)	42.16 ± 15.36	16.11 ± 4.31
CRP (mg/L)	22.16 ± 13.25	2.85 ± 0.90
Erythrocytes (10^6/μL)	4.29 ± 0.33	4.23 ± 0.43
Leukocytes (10^3/μL)	7.14 ± 1.69	6.05 ± 1.65
Platelets (10^3/μL)	276.32 ± 99.34	256.12 ± 78.13
DAS28	5.17 ± 0.45	N/A
Tender joint count	7.16 ± 2.76	0
Swollen joint count	6.02 ± 2.07	0
GM-CSF (pg/mL)	6.88 ± 2.39	< 6.5
IL-6 (pg/mL)	3.57 ± 1.22	3.10–10.3
TNF-α (pg/mL)	23.17 ± 10.12	< 13.7

Values for GM-CSF, IL-6, and TNF-α in healthy controls are reported as reference ranges or assay thresholds when population statistics were unavailable.
Abbreviations: RA — rheumatoid arthritis; ESR — erythrocyte sedimentation rate; CRP — C-reactive protein; DAS28 — Disease Activity Score in 28 joints; GM-CSF — granulocyte–macrophage colony-stimulating factor; IL-6 — interleukin-6; TNF-α — tumor necrosis factor alpha; N/A — not applicable.

3. Numerical results and performance stability analysis

3.1. Clinical and demographic characteristics of the study cohort

Demographic characteristics of the study participants are summarized in Table 4. Patients with RA had a mean age of 51.7 ± 10.6 years and a mean disease duration of 8.3 ± 4.6 years, while healthy controls had a comparable mean age of 52.1 ± 10.2 years.

3.2. Clinical and biochemical characteristics of the study cohort

The clinical and biochemical characteristics of the study cohort are summarized in Table 5.

Compared with healthy controls, patients with RA showed higher ESR, CRP, DAS28, tender and swollen joint counts, and TNF-α, supporting the interpretation that the RA cohort represented a clinically active inflammatory group.

As shown in Table 6, significant positive correlations were observed between ESR and CRP, DAS28 and tender joint count, DAS28 and swollen joint count, tender and swollen joint counts, and IL-6 and TNF-α. A negative correlation was observed between GM-CSF and platelet count.

3.3. Overall diagnostic performance and measurement stability

This section reports the experimental characterization of the proposed framework as an internally validated diagnostic measurement approach, with emphasis on computational stability, robustness to subject-level partitioning, and probabilistic calibration. Because repeated physical acquisitions were not available for individual participants, these analyses should not be interpreted as a direct estimate of metrological test-retest repeatability. Beyond average classification performance, the analysis considered inter-seed dispersion, ranking consistency, and probabilistic calibration to provide a broader assessment of model behavior.

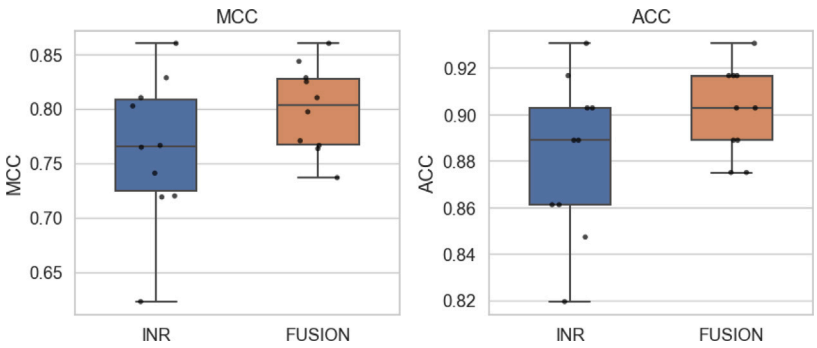


Fig. 3. Comparative distribution of test-set MCC and Accuracy across 10 independent subject-disjoint train–validation–test partitions for the INR baseline and FUSION frameworks. Each partition was generated using a different random seed and involved the complete study cohort rather than a subset of ten patients. Individual markers represent independent experimental realizations. FUSION showed higher median performance and reduced dispersion, supporting improved robustness to stochastic data partitioning.

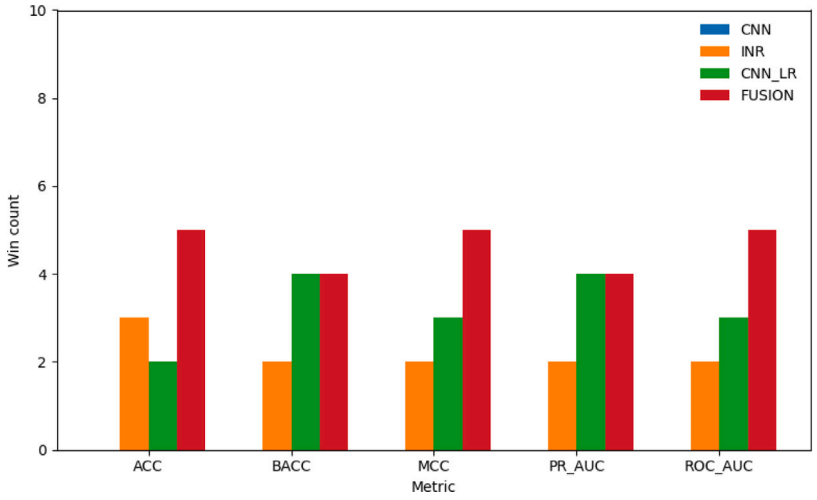


Fig. 4. Win-count analysis across 10 independent subject-disjoint experimental realizations for the evaluated performance metrics: Accuracy (ACC), Balanced Accuracy (BACC), Matthews correlation coefficient (MCC), receiver operating characteristic area under the curve (ROC-AUC), and precision–recall area under the curve (PR-AUC). Bars represent the number of realizations in which a given model achieved the highest value of a particular metric. For each metric and realization, only the single best-performing model was counted; consequently, win counts sum to 10 within each metric. Higher win counts indicate greater ranking consistency and improved robustness to stochastic variations in subject partitioning across independent experimental realizations.

Table 6
Statistically significant correlations between laboratory parameters and disease-activity indices in patients with RA.

Variable pair	Correlation coefficient	Direction
ESR — CRP	0.65	Positive
DAS28 — tender joint count	0.58	Positive
DAS28 — swollen joint count	0.49	Positive
Tender joint count — swollen joint count	0.71	Positive
IL-6 — TNF-α	0.92	Positive
GM-CSF — platelets	−0.46	Negative

Abbreviations: ESR — erythrocyte sedimentation rate; CRP — C-reactive protein; DAS28 — disease activity score based on 28 joints; IL-6 — interleukin 6; TNF-α — tumor necrosis factor alpha; GM-CSF — granulocyte-macrophage colony-stimulating factor. Only statistically significant correlations are shown.

As shown in Table 8, FUSION achieved the highest mean performance across all primary discriminative metrics, with ACC = 0.910, BACC = 0.914, MCC = 0.819, ROC-AUC = 0.954, and PR-AUC = 0.901. CNN_LR provided a competitive linear-probing baseline, particularly for ROC-AUC and PR-AUC, but remained below FUSION in all reported primary metrics. These results indicate higher mean discrimination and lower descriptive inter-seed variability for FUSION compared with the standalone CNN, INR, and CNN_LR baselines. The MCC is a balanced agreement measure that is informative even under class imbalance, with values near 1 indicating strong agreement between predicted and reference labels [41]. TRUE INR was treated as a per-sample optimization ablation and was not included in the main feed-forward model ranking. CNN_LR was included as a linear-probing baseline to assess whether the convolutional embedding alone was sufficient for stable discrimination without radiomic late fusion.

3.4. Calibration analysis and diagnostic reliability

As shown in Table 7, FUSION achieved the lowest Brier score (0.084 ± 0.019) and ECE (0.094 ± 0.021), indicating favorable probability calibration in terms of squared probabilistic error and expected calibration deviation. However, CNN_LR achieved the lowest LogLoss (0.358 ± 0.137), whereas FUSION obtained a LogLoss of 0.395 ± 0.195. These results indicate that calibration performance was metric-dependent rather than uniformly superior for FUSION.

3.5. Comparative stability: FUSION versus INR baseline

A focused comparison between the INR baseline and the proposed FUSION framework is presented in Fig. 3, which shows the distribution of test-set MCC and Accuracy across identical random subject-level splits. Each marker corresponds to an individual experimental

Table 7
Calibration performance metrics (lower values indicate superior alignment) evaluated on the test set. Results are reported as mean ± SD across 10 independent experimental seeds. Bold values denote the optimal performance per metric, highlighting the model's ability to maintain probabilistic measurement reliability.

Metric	CNN	TRUE INR	INR	CNN_LR	FUSION
BRIER	0.167 ± 0.046	0.216 ± 0.031	0.097 ± 0.026	0.090 ± 0.022	**0.084 ± 0.019**
ECE	0.191 ± 0.069	0.203 ± 0.052	0.120 ± 0.029	0.095 ± 0.022	**0.094 ± 0.021**
LOGLOSS	0.581 ± 0.154	0.738 ± 0.122	0.449 ± 0.184	**0.358 ± 0.137**	0.395 ± 0.195

Abbreviations: CNN — convolutional neural network; TRUE INR — coordinate-based INR baseline; INR — implicit neural representation model; CNN_LR — convolutional neural network with logistic regression classifier; FUSION — proposed multimodal late-fusion framework; BRIER — Brier score; ECE — expected calibration error; LOGLOSS — logarithmic loss.

Table 8
Performance stability across ten independent subject-disjoint realizations. Results are reported as mean ± SD.

Model	ACC	BACC	MCC	ROC-AUC	PR-AUC
CNN	0.819 ± 0.033	0.822 ± 0.028	0.649 ± 0.053	0.910 ± 0.012	0.824 ± 0.037
INR	0.882 ± 0.034	0.887 ± 0.035	0.763 ± 0.068	0.940 ± 0.025	0.874 ± 0.062
FUSION	**0.910 ± 0.018**	**0.914 ± 0.024**	**0.819 ± 0.038**	**0.954 ± 0.012**	**0.901 ± 0.034**
CNN_LR	0.881 ± 0.037	0.886 ± 0.038	0.760 ± 0.076	0.942 ± 0.024	0.897 ± 0.049

Abbreviations: CNN — convolutional neural network; INR — implicit neural representation; FUSION — proposed multimodal late-fusion framework; CNN_LR — convolutional neural network with logistic regression classifier; ACC — accuracy; BACC — balanced accuracy; MCC — Matthews correlation coefficient; ROC-AUC — area under the receiver operating characteristic curve; PR-AUC — area under the precision–recall curve.

Table 9
Number of experimental realizations (out of 10) in which each model achieved the highest value of a given evaluation metric.

Model	ACC	BACC	MCC	PR_AUC	ROC_AUC
CNN	0	0	0	0	0
INR	3	2	2	2	2
CNN_LR	2	4	3	4	3
FUSION	**5**	**4**	**5**	**4**	**5**

For each metric and experimental realization, only the single best-performing model was counted. Consequently, win counts sum to 10 within each metric column.
Abbreviations: INR — implicit neural representation model; CNN_LR — convolutional neural network with logistic-regression classifier; FUSION — proposed multimodal late-fusion framework; ACC — accuracy; BACC — balanced accuracy; MCC — Matthews correlation coefficient; PR-AUC — area under the precision–recall curve; ROC-AUC — area under the receiver operating characteristic curve.

seed, providing a transparent view of sensitivity to data partitioning. Compared with the INR baseline, FUSION shifted the performance distribution toward higher values and reduced the IQR.

These results indicate that continuous thermal-field reconstruction alone did not match the stability of the combined CNN–radiomics representation.

3.6. Win-count analysis and ranking stability

To assess reliability beyond aggregated point estimates, win-count analysis was conducted across 10 independent experimental realizations, where each realization corresponded to a complete subject-disjoint train–validation–test partition generated using a different random seed. This non-parametric approach evaluated how consistently each model achieved the highest ranking under repeated data partitioning.

As shown in Fig. 4 and Table 9, FUSION achieved the highest winning frequency among the evaluated models for the primary discriminative metrics. Specifically, it ranked first in 50% of the realizations for both Accuracy and Matthews Correlation Coefficient. This result suggests improved ranking stability under repeated stochastic realizations and indicates that the observed performance advantage was not limited to a single favorable partition. Although FUSION did not win every realization, it achieved the highest win frequency across the primary metrics; no competing model exceeded four wins for any reported metric. Thus, the win-count analysis supports a descriptive ranking advantage, but not a dominant or statistically conclusive superiority.

3.7. Computational complexity and PoC feasibility

The practical feasibility of the evaluated architectures for PoC scenarios was assessed using the complexity analysis summarized in Table 10. The standalone CNN based on EfficientNet-B0 contained approximately 5.3×10^6 parameters and showed an inference latency of 8.133 ms per sample. In the proposed FUSION framework, the final linear decision stage was computationally lightweight, requiring only 774 floating-point operations (FLOPs) with a latency of 0.0020 ms per sample after feature extraction. These values refer specifically to the linear classifier operating on the extracted CNN and radiomic descriptors, rather than to the complete preprocessing and feature-extraction pipeline. FLOPs were used as a hardware-independent measure of computational complexity at this decision stage. Deterministic preprocessing steps, including BHT segmentation and Zhang–Suen skeletonization, together with targeted radiomic feature extraction, further supported reproducible signal conditioning before classification. From an instrumentation perspective, this design may facilitate future implementation in resource-constrained thermographic systems, although validation on dedicated embedded hardware remains necessary.

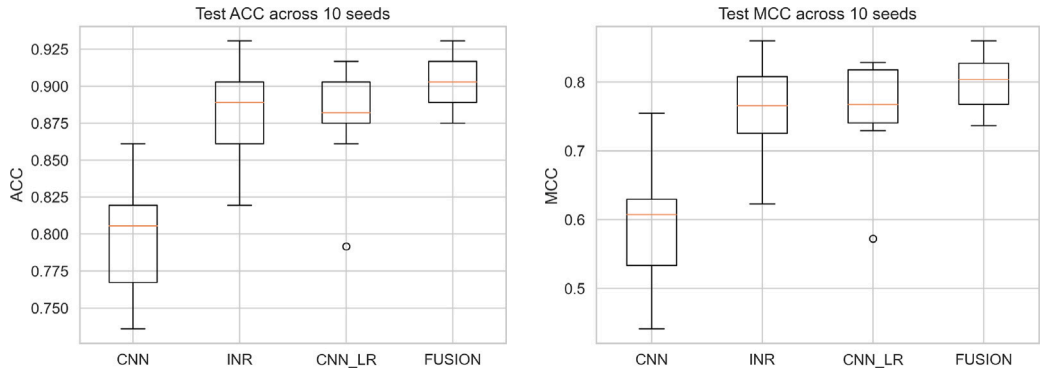
High-performance GPU acceleration may be useful during CNN training or feature extraction, whereas the final linear classifier can be executed efficiently on conventional central processing units (CPUs). This supports a resource-efficient instrumentation design by limiting the computational burden of the classification stage while preserving compatibility with future PoC deployment.

As summarized in Table 11, the proposed FUSION configuration reduced training energy consumption from 11.22 Wh for the standalone CNN to 9.73 Wh, corresponding to an approximately 13.3% reduction, while also requiring less training time.

From a measurement-instrumentation perspective, computational efficiency is a system-level design constraint that may affect portability, power consumption, thermal management, and deployment feasibility. In the present internal evaluation, the proposed FUSION architecture achieved a favorable accuracy–complexity trade-off relative to the standalone CNN baseline. This result is relevant for future PoC thermal instrumentation, where computational resources and power budgets may be limited.

3.8. Global performance distribution and robustness

Fig. 5 summarizes the distribution of ACC and MCC across repeated subject-disjoint realizations. FUSION showed higher median performance and lower dispersion than the standalone baselines, supporting improved robustness to stochastic data partitioning.



(a) Test accuracy across 10 random seeds. (b) Test MCC across 10 random seeds.

Fig. 5. Distribution of ACC and MCC across 10 independent subject-disjoint experimental realizations generated using different random seeds. Boxplots summarize the variability of model performance under repeated stochastic data partitioning. The FUSION framework exhibits the highest median performance together with reduced inter-seed dispersion compared with the standalone CNN, CNN_LR and INR baselines, indicating improved computational robustness and ranking stability.

Table 10
Computational complexity and resource use of the evaluated architectures. FLOPs and inference latency were reported per single measurement sample.

Model	Trainable params	FLOPs	Latency (ms)
TRUE INR	132	–	– (per-sample fit)
INR	132	262	0.0010[a]
CNN (EfficientNet-B0)	$\approx 5.3 \times 10^6$	3.99×10^7	8.133
CNN_LR	257	512	0.0020[a]
FUSION	linear decision stage	774	0.0020[a]

[a] Latency estimated analytically for the lightweight decision stage. Note that TRUE INR differs fundamentally from the remaining architectures in that it requires per-sample optimization of a coordinate-based MLP rather than feed-forward inference. As a result, its computational cost depends on the number of optimization iterations and cannot be expressed as a single, architecture-level FLOPs or latency value. For this reason, TRUE INR is reported as a performance ablation baseline rather than a deployable inference model, and its computational complexity is not directly comparable to feed-forward CNN or fusion architectures.
Abbreviations: TRUE INR — coordinate-based INR baseline fitted independently to each sample; INR — implicit neural representation model; CNN — convolutional neural network; CNN_LR — convolutional neural network with logistic-regression classifier; FUSION — proposed multimodal late-fusion framework; FLOPs — floating-point operations.

Table 11
Training time and energy consumption measured across ten independent realizations.

Model	Training time (s)	Energy (Wh)
CNN	1447.4	11.22
INR	819.4	3.44
FUSION	1034.4	9.73

Abbreviations: CNN — convolutional neural network; INR — implicit neural representation; FUSION — proposed multimodal late-fusion framework; Wh — watt-hour.

3.9. Statistical power and ranking consistency

Wilcoxon testing (Table 12) showed favorable paired differences for FUSION, but the limited number of realizations reduced statistical power. Therefore, these results were interpreted together with win-count and distributional analyses.

The pairwise Wilcoxon analysis showed that FUSION achieved higher mean values than INR for ACC, BACC, MCC, ROC-AUC, and PR-AUC. However, after adjustment for multiple comparisons, these differences did not reach the conventional significance threshold, with MCC showing the strongest tendency toward improvement. In contrast, FUSION significantly outperformed the standalone CNN baseline after adjustment for ACC, BACC, MCC, and ROC-AUC, whereas the PR-AUC difference showed a non-significant trend. These findings support the interpretation that FUSION provides a consistent descriptive stability advantage, while formal superiority over the strongest INR baseline remains limited by the small number of repeated realizations.

3.10. Ablation study and design implications

Two INR-related configurations were distinguished in this study. The INR baseline denotes the feed-forward implicit-representation configuration included in the main comparative evaluation. In addition, the TRUE INR configuration denotes a coordinate-based MLP fitted independently for each sample and was used as a controlled per-sample ablation. Because TRUE INR requires per-sample optimization rather than standard feed-forward inference, it was interpreted separately from the main baseline models. The comparative results reported above suggest that continuous coordinate-based approximation alone was insufficient to match the performance of representations incorporating hierarchical convolutional features. The performance gains observed in the proposed framework were therefore attributed to the complementary integration of structured handcrafted radiomics and hierarchical convolutional embeddings.

4. Discussion

The ADT results should be interpreted as dynamic recovery responses rather than as static skin-temperature differences. The post-cold-stress rewarming phase reflects vascular recovery and local heat redistribution, processes that may be altered in inflammatory tissue. For this reason, the extracted radiomic and convolutional descriptors were considered functional indicators of inflammation-related thermal heterogeneity and recovery dynamics, rather than direct structural markers of synovial pathology. Because the temperature field changed continuously after thermal excitation, the analyzed thermographic sequences represented non-stationary biomedical signals with both temporal and spatial variability. Accordingly, the thermal recovery response constitutes the physical measurand of the proposed framework, whereas inflammatory activity represents its biological and clinical interpretation rather than a directly measured quantity.

Previous IRT-based RA studies have reported promising discrimination based on static or dynamic thermal descriptors, but direct numerical comparison remains difficult because of differences in acquisition protocols, cohort definitions, anatomical regions, and validation strategies. In contrast to earlier single-protocol or single-split analyses, the present study emphasized repeated subject-disjoint evaluation, calibration metrics, and ranking stability. Therefore, the main methodological contribution lies not only in the achieved mean ROC-AUC of 0.954, but

Table 12
Pairwise Wilcoxon signed-rank comparisons across ten independent subject-disjoint realizations. Mean values are reported for FUSION and the corresponding baseline model. The median paired difference is expressed as FUSION minus baseline.

Metric	Comparison	Mean baseline	Mean FUSION	Median diff.	p	p_{adj}
ACC	FUSION vs. INR	0.882	0.910	+0.0208	0.0355	0.1419
BACC	FUSION vs. INR	0.887	0.914	+0.0170	0.0488	0.1465
MCC	FUSION vs. INR	0.763	0.819	+0.0387	0.0137	0.0684
ROC-AUC	FUSION vs. INR	0.940	0.954	+0.0130	0.1055	0.2109
PR-AUC	FUSION vs. INR	0.874	0.901	+0.0236	0.2324	0.2324
ACC	FUSION vs. CNN	0.819	0.910	+0.0903	0.0020	0.0293
BACC	FUSION vs. CNN	0.822	0.914	+0.0958	0.0020	0.0293
MCC	FUSION vs. CNN	0.649	0.819	+0.1629	0.0020	0.0293
ROC-AUC	FUSION vs. CNN	0.910	0.954	+0.0467	0.0020	0.0293
PR-AUC	FUSION vs. CNN	0.824	0.901	+0.0901	0.0098	0.0586

Holm-adjusted p-values were computed over the full family of 15 pairwise tests, including all pairwise comparisons among CNN, INR, and FUSION across the five primary discriminative metrics. CNN_LR was evaluated descriptively as a secondary linear-probing reference and was not included in this inferential family.
Abbreviations: ACC — accuracy; BACC — balanced accuracy; MCC — Matthews correlation coefficient; ROC-AUC — area under the receiver operating characteristic curve; PR-AUC — area under the precision–recall curve; CNN — convolutional neural network; INR — implicit neural representation; FUSION — proposed multimodal late-fusion framework; p — unadjusted Wilcoxon signed-rank test p-value; p_{adj} – Holm-adjusted p-value.

also in the stability-oriented evaluation of a late CNN-radiomics fusion framework.

The results suggest that CNN–radiomics late fusion improved diagnostic performance and ranking stability, while showing favorable, although metric-dependent, probabilistic calibration in the internal subject-disjoint evaluation, in line with prior evidence that late feature-level fusion and radiomics–AI integration can enhance multimodal medical-image analysis [42,43]. Reliable probability calibration is particularly relevant for clinical decision-support settings, where overconfident outputs may limit the practical utility and safe deployment of diagnostic models [44]. These observations are consistent with previous reports indicating that IRT can capture inflammation-related thermal alterations in musculoskeletal and rheumatic conditions, while also highlighting the need for robust computational processing of thermal patterns [5,6,10]. In contrast to studies based mainly on static thermographic descriptors, the present framework emphasizes dynamic recovery information, stability across repeated subject-disjoint evaluations, and calibrated probabilistic assessment. However, the findings should be interpreted as methodological feasibility evidence rather than definitive clinical validation, because the evaluation remained internal and was based on a limited number of repeated realizations.

From a measurement-science perspective, the main contribution of this work is the treatment of ADT-based RA assessment as a calibrated dynamic measurement problem rather than only as an image-classification task. The measurement chain links a defined thermal excitation, observation of the post-stimulus thermal response, deterministic preprocessing, complementary quantitative thermal descriptors, probabilistic decision support, and repeated computational evaluation within a single framework. Within this interpretation, the CNN-radiomics fusion stage is not the measurement itself, but a computational component of the measurement chain that transforms complementary descriptors of the observed thermal response into an assessment output. The primary physical information remains the calibrated spatio-temporal temperature response acquired during ADT.

4.1. Limitations and future perspectives

Several limitations should be acknowledged. The present study remains an internally validated methodological feasibility study. Although demographic, clinical, and biochemical cohort characteristics were included, model performance was not evaluated within clinically defined subgroups, such as sex-specific strata, disease stage, disease severity, or medication-related subgroups. Future studies should therefore assess subgroup-related variability in thermographic features and model outputs.

Thermographic outcomes were not prospectively validated against independent laboratory or imaging markers of inflammation. Future studies should compare dynamic thermographic descriptors with biomarkers such as CRP, ESR, RF, anti-CCP, IL-6, TNF-α, broader cytokine panels, ultrasound, and MRI markers of synovitis.

A further metrological limitation is that the dataset contained one ADT acquisition per participant. Consequently, within-subject test-retest repeatability of the complete physical measurement procedure could not be quantified. The repeated computational realizations used in this study assess stability with respect to subject partitioning and model stochasticity, but they do not replace repeated physical measurements under repeatability conditions. Dedicated repeated-acquisition experiments are therefore required to estimate within-subject repeatability and to construct a comprehensive measurement-uncertainty budget.

All thermographic data were acquired at a single center using one LWIR camera and a unified acquisition protocol. Therefore, the results should not be interpreted as hardware-independent or externally validated. LWIR thermography captures superficial skin-temperature patterns rather than detailed anatomical structures [3,4]. In addition, thermal infrared radiation is sensitive to optical barriers and reflected radiation; glass or water may block or distort LWIR transmission, and glass can act as an infrared mirror, potentially introducing reflected images of the operator, participant, or surrounding objects into the thermal scene [45]. Thus, thermographic findings should be interpreted as surface-temperature descriptors rather than direct structural images. Larger multicenter studies using diverse sensor configurations are required to establish inter-site reproducibility, sensor-to-sensor robustness, and clinical generalizability.

The preprocessing pipeline did not include an explicit cold-image-based absolute thermal-contrast channel. Future work should investigate baseline-to-cold temperature-drop maps, cold-to-recovery rewarming maps, and normalized thermal-contrast representations to determine whether contrast-based preprocessing improves sensitivity to vascular recovery abnormalities and inflammation-related thermal heterogeneity.

5. Conclusions

This study evaluated a measurement-oriented framework for RA assessment based on the dynamic thermal response elicited by controlled ADT excitation. Within this framework, CNN-radiomics late fusion served as a computational information-integration stage rather than as the measurement principle itself. By integrating learned convolutional features with biophysically grounded radiomic descriptors, the proposed FUSION processing stage achieved the highest mean

performance across ten subject-disjoint realizations, with ACC = 0.910, BACC = 0.914, MCC = 0.819, ROC-AUC = 0.954, and PR-AUC = 0.901. Distributional and ranking analyses suggested lower descriptive inter-seed variability and improved ranking consistency. Paired Wilcoxon comparisons showed significant improvements over the standalone CNN baseline for ACC, BACC, MCC, and ROC-AUC after adjustment, whereas differences relative to the stronger INR baseline did not reach the adjusted significance threshold. Therefore, these findings should be interpreted as internally validated methodological feasibility evidence rather than as externally confirmed clinical superiority.

The main contribution of this work is a computationally parsimonious measurement framework that links calibrated thermal acquisition, deterministic preprocessing, multimodal feature extraction, calibrated probabilistic assessment, and computational stability evaluation. The lightweight linear decision stage, with an estimated inference latency of 0.0020 ms after feature extraction, supports the feasibility of real-time implementation; however, this computational efficiency should be interpreted as an implementation advantage rather than as evidence of clinical validation. The deterministic preprocessing pipeline, including BHT segmentation and Zhang–Suen skeletonization, further supported repeatable anatomical localization and reduced sensitivity to postural and sampling variability. Overall, the results suggest that combining radiomic descriptors, convolutional representations, and late feature-level fusion may improve the internal stability of the measurement-derived diagnostic output in RA. Nevertheless, the proposed framework should be interpreted as a methodological feasibility approach requiring further validation. Future research should focus on larger multicenter cohorts, robustness across diverse infrared sensor configurations, and validation of dynamic thermographic descriptors against independent laboratory and imaging markers of RA activity.

The present results demonstrate internal computational stability and robustness to subject-level partitioning, but do not establish full metrological repeatability or reproducibility of the physical ADT measurement. Future validation should therefore include repeated within-subject acquisitions, multicenter and inter-device studies, and quantitative propagation of measurement uncertainty across the complete acquisition and processing chain.

CRediT authorship contribution statement

Agnieszka Pregowska: Writing – original draft, Software, Methodology, Conceptualization, Formal analysis, Visualization . **Konrad Pauk:** Writing – review & editing, Validation, Formal analysis, Data curation . **Mikhail Ihnatouski:** Resources, Methodology, Software, Formal analysis, Visualization, Validation. **Jolanta Pauk:** Writing – original draft, Writing – review & editing, Supervision, Project administration, Investigation, Funding acquisition.

Funding

The study was partially supported by the Polish Ministry of Science and Higher Education as part of project WZ/WM–IIB/2/2024.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors thank Dr. Agnieszka Wasilewska for her contribution to thermographic data acquisition.

Data availability

The source code, anonymized data, and supplementary materials supporting the findings of this study are available from the corresponding author upon reasonable request. The thermographic datasets contain clinical measurements and are therefore not publicly available due to privacy and ethical restrictions. Anonymized data may be shared subject to institutional and ethical approval.

References

[1] J.S. Smolen, R.B.M. Landewé, S.A. Bergstra, et al., EULAR recommendations for the management of rheumatoid arthritis with synthetic and biological disease-modifying antirheumatic drugs: 2022 update, Ann. Rheum. Dis. 82 (2023) 3–18, http://dx.doi.org/10.1136/ard-2022-223356.

[2] F. Costantino, L. Carmona, M. Boers, et al., EULAR recommendations for the reporting of ultrasound studies in rheumatic and musculoskeletal diseases (RMDs), Ann. Rheum. Dis. 80 (2021) 840–847, http://dx.doi.org/10.1136/annrheumdis-2020-219816.

[3] B.B. Lahiri, S. Bagavathiappan, T. Jayakumar, J Philip, Medical applications of infrared thermography: A review, Infrared Phys. Technol. 55 (2012) 221–235, http://dx.doi.org/10.1016/j.infrared.2012.03.007.

[4] E.F.J. Ring, K Ammer, Infrared thermal imaging in medicine, Physiol. Meas. 33 (2012) R33–R46, http://dx.doi.org/10.1088/0967-3334/33/3/R33.

[5] G. Schiavon, E.M. Zanetti, L. Atzori, et al., Infrared thermography for the evaluation of inflammatory and degenerative joint diseases: A systematic review, Ther. Adv. Musculoskelet. Dis. 13 (2021) http://dx.doi.org/10.1177/1759720X211045756, 1759720X211045756.

[6] N. Nallathambi, M. Vathulya, V Nadarajah, Clinical application of infrared thermography in rheumatic and musculoskeletal disorders, Mediterr. J. Rheumatol. 36 (2025) 168–177, http://dx.doi.org/10.31138/mjr.20240412.sh.

[7] J.-W. Kim, J.-S. Kong, S. Lee, S.-A. Yoo, J.-H. Koh, J. Jin, W.-U. Kim, Angiogenic cytokines can reflect the synovitis severity and treatment response to biologics in rheumatoid arthritis, Exp. Mol. Med. 52 (2020) 843–853, http://dx.doi.org/10.1038/s12276-020-0443-8.

[8] F. Khodadust, A. Ezdoglian, M.M. Steinz, J.R. van Beijnum, G.J.C. Zwezerijnen, G. Jansen, S.W. Tas, C.J. van der Laken, Systematic review: Targeted molecular imaging of angiogenesis and its mediators in rheumatoid arthritis, Int. J. Mol. Sci. 23 (2022) 7071, http://dx.doi.org/10.3390/ijms23137071.

[9] J. Pauk, A. Wasilewska, M Ihnatouski, Infrared thermography sensor for disease activity detection in rheumatoid arthritis patients, Sensors 19 (2019) 3444, http://dx.doi.org/10.3390/s19163444.

[10] Y.K. Tan, G.H. Lim, Comparing thermographic heat signatures with joint inflammation detected by ultrasound at the knees in patients with rheumatoid arthritis, Sci. Rep. 15 (2025) 31960, http://dx.doi.org/10.1038/s41598-025-17901-6.

[11] M.C. Ergene, A. Bayrak, M Ceylan, The use of prototypical and siamese networks in the determination of lower extremity injuries in professional football players with thermographic data, Quant. InfraRed Thermogr. J. (2025) http://dx.doi.org/10.1080/17686733.2025.2594955.

[12] R. Usamentiaga, A. Fidanza, B. Yousefi, G. Iacutone, G. Logroscino, S Sfarra, Advancing knee injury prevention and anomaly detection in rugby players through automated processing of infrared thermography: A novel biothermodynamics approach, Therm. Sci. Eng. Prog. 65 (2025) 103782, http://dx.doi.org/10.1016/j.tsep.2025.103782.

[13] A.Z. Nowakowski, M Kaczmarek, Artificial intelligence in IR thermal imaging and sensing for medical applications, Sensors 25 (3) (2025) 891, http://dx.doi.org/10.3390/s25030891.

[14] J. Vicnesh, M. Salvi, Y. Hagiwara, et al., Application of infrared thermography and artificial intelligence in healthcare: A systematic review of over a decade (2013–2024), IEEE Access 12 (2024) 1–25, http://dx.doi.org/10.1109/ACCESS.2024.3522251.

[15] J. Gauci, O. Falzon, K.P. Camilleri, Principal component analysis for dynamic thermal video analysis, Infrared Phys. Technol. 104 (2020) 103359, http://dx.doi.org/10.1016/j.infrared.2019.103359.

[16] H. Rabus, N. Agarwal, A. Boubekki, et al., Metrology for artificial intelligence in medicine, Meas.: Sensors 38 (2025) 101789, http://dx.doi.org/10.1016/j.measen.2024.101789.

[17] M. Wieczorowski, D. Kucharski, P. Sniatala, et al., A novel approach to using artificial intelligence in coordinate metrology including nano scale, Measurement 217 (2023) 113051, http://dx.doi.org/10.1016/j.measurement.2023.113051.

[18] K. Üreten, H. Erbay, H.H. Maraş, Detection of rheumatoid arthritis from hand radiographs using a convolutional neural network, Clin. Rheumatol. 39 (2020) 969–974, http://dx.doi.org/10.1007/s10067-019-04487-4.

[19] P. Lambin, R.T.H. Leijenaar, T.M. Deist, J. Peerlings, E.E.C. de Jong, J. van Timmeren, S. Sanduleanu, R.T.H.M. Larue, A.J.G. Even, A. Jochems, Y. van Wijk, H. Woodruff, J. van Soest, T. Lustberg, E. Roelofs, W. van Elmpt, A. Dekker, F.M. Mottaghy, J.E. Wildberger, S. Radiomics Walsh, The bridge between medical imaging and personalized medicine, Nat. Rev. Clin. Oncol. 14 (2017) 749–762, http://dx.doi.org/10.1038/nrclinonc.2017.141.

[20] V. Sitzmann, J. Martel, A. Bergman, et al., Implicit neural representations with periodic activation functions, in: Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 33, 2020, pp. 7462–7473.
[21] M. Tancik, P.P. Srinivasan, B. Mildenhall, et al., Fourier features let networks learn high frequency functions in low dimensional domains, in: Proc. 34th Int. Conf. Neural Inf. Process. Syst., NIPS, 2020, pp. 7537–7547.
[22] Agency for health technology assessment and tariff system. recommendation no. 23/2025 of 28 february 2025 of the president of the agency for health technology assessment and tariff system on the appraisal of remsima (infliximab) within drug program b.33: treatment of patients with active rheumatoid arthritis and juvenile idiopathic arthritis (icd-10: m05, m06, m08). warsaw: aotmit;, 2025, Available from: https://bip.aotm.gov.pl/assets/files/zlecenia_mz/2024/205/REK/2025%2002%2028%20BP%20Rekomendacja%20nr%2023%202025%20Remsima%20RZS%20BIP_REOPTR.pdf.
[23] J Steketee, Spectral emissivity of skin and pericardium, Phys. Med. Biol. 18 (1973) 686–694, http://dx.doi.org/10.1088/0031-9155/18/5/307.
[24] R.C. Gonzalez, R.E. Woods, Digital Image Processing, fourth ed., Pearson, New York, 2018.
[25] A. Anjos, H.R. Shahbazkia, Bi-level image thresholding: A fast method, Biosignals 2 (2008) 70–76.
[26] T. Gupta, I. Sreedevi, R. Jindal, Segmenting hotspots from medical thermal images using density-based modified FC-$P_c$fs with spatial information, Quant. InfraRed Thermogr J. 22 (2025) 368–404, http://dx.doi.org/10.1080/17686733.2024.2408714.
[27] T.Y. Zhang, C.Y. Suen, A fast parallel algorithm for thinning digital patterns, Commun. ACM 27 (1984) 236–239, http://dx.doi.org/10.1145/357994.358023.
[28] R.M. Haralick, K. Shanmugam, I. Dinstein, Textural features for image classification, IEEE Trans. Syst. Man Cybern. 3 (1973) 610–621, http://dx.doi.org/10.1109/TSMC.1973.4309314.
[29] T. Ojala, M. Pietikäinen, T. Mäenpää, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Trans. Pattern Anal. Mach. Intell. 24 (2002) 971–987, http://dx.doi.org/10.1109/TPAMI.2002.1017623.
[30] S.G. Mallat, A theory for multiresolution signal decomposition: the wavelet representation, IEEE Trans. Pattern Anal. Mach. Intell. 11 (1989) 674–693, http://dx.doi.org/10.1109/34.192463.
[31] T. Baltrušaitis, C. Ahuja, L.-P. Morency, Multimodal machine learning: A survey and taxonomy, IEEE Trans. Pattern Anal. Mach. Intell. 41 (2019) 423–443, http://dx.doi.org/10.1109/TPAMI.2018.2798607.
[32] M. Tan, Q.V. Le, Efficientnet: rethinking model scaling for convolutional neural networks, in: Proceedings of the 36th International Conference on Machine Learning, 2019, pp. 6105–6114.
[33] G. Alain, Y. Bengio, Understanding intermediate layers using linear classifier probes, in: International Conference on Learning Representations, ICLR, Toulon, France, 2017.
[34] S. Mallat, Understanding deep convolutional networks, Philos. Trans. R. Soc. A 374 (2016) 20150203, http://dx.doi.org/10.1098/rsta.2015.0203.
[35] J. Bruna, S. Mallat, Invariant scattering convolution networks, IEEE Trans. Pattern Anal. Mach. Intell. 35 (2013) 1872–1886, http://dx.doi.org/10.1109/TPAMI.2012.230.
[36] C. Guo, G. Pleiss, Y. Sun, et al., On calibration of modern neural networks, in: Proc. 34th Int. Conf. Mach. Learn., ICML, 2017, pp. 1321–1330.
[37] A. Benavoli, G. Corani, J. Demsar, et al., Time for a change: a tutorial for comparing hypothesis testing in machine learning, J. Mach. Learn. Res. 18 (2017) 1–52.
[38] L.I. Kuncheva, Combining Pattern Classifiers: Methods and Algorithms, second ed., Wiley, Hoboken, 2014.
[39] J.W. Tukey, Exploratory Data Analysis, Addison-Wesley, Reading, MA, 1977.
[40] J. Demšar, Statistical comparisons of classifiers over multiple data sets, J. Mach. Learn. Res. 7 (2006) 1–30.
[41] D. Chicco, G. Jurman, The advantages of the matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation, BMC Genomics 21 (2020) 6, http://dx.doi.org/10.1186/s12864-019-6413-7.
[42] K. Gadzicki, R. Khamsehashari, C. Zetzsche, Early vs late fusion in multimodal convolutional neural networks, in: Proc. 2020 IEEE 23rd Int. Conf. Inf. Fusion, FUSION, IEEE, 2020, pp. 1–6, http://dx.doi.org/10.23919/FUSION45037.2020.9190246.
[43] A. Maniaci, S. Lavalle, C. Gagliano, et al., The integration of radiomics and artificial intelligence in modern medicine, Life 14 (2024) 1248, http://dx.doi.org/10.3390/life14101248.
[44] C.J. Kelly, A. Karthikesalingam, M. Suleyman, et al., Key challenges for delivering clinical impact with artificial intelligence, BMC Med. 17 (2019) 195, http://dx.doi.org/10.1186/s12916-019-1426-2.
[45] R. Usamentiaga, P. Venegas, J. Guerediaga, L. Vega, J. Molleda, F.G. Bulnes, Infrared thermography for temperature measurement and non-destructive testing, Sensors 14 (2014) 12305–12348, http://dx.doi.org/10.3390/s140712305.