

# Infrared Thermal Imaging Combined with Machine Learning for Detecting Disease Activity in Rheumatoid Arthritis: A Proof-of-Concept Study

Neha G. Narula

[Neha.narula@va.gov](mailto:Neha.narula@va.gov)

North Florida/South Georgia Veterans Health System

Michael Bubb

North Florida/South Georgia Veterans Health System

---

## Article

**Keywords:** rheumatoid arthritis, thermography, infrared imaging, machine learning, disease activity, DAS28

**Posted Date:** October 7th, 2026

**DOI:** <https://doi.org/10.21203/rs.3.rs-10985518/v1>

**License:** © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

**Additional Declarations:** No competing interests reported.

---

# Abstract

Standard rheumatoid arthritis (RA) flare assessment relies on components that are either subjective (patient global assessment) or non-specific (ESR, CRP), motivating interest in objective, non-invasive adjuncts. This proof-of-concept study evaluated infrared thermal imaging combined with machine learning for detecting RA flare and classifying disease activity. Eleven seropositive RA patients and nine seronegative controls underwent standardized thermal imaging alongside clinical assessment across 2–3 visits. Whole-hand maximum temperature (Tmax) was compared between RA flare visits and controls. Per-joint thermal zone data were used to train logistic regression and random forest classifiers predicting disease activity, using leave-one-patient-out cross-validation to prevent data leakage. RA patients showed significant clinical improvement with treatment (DAS28-ESR 5.23 to 3.32,  $p = 0.0007$ ). Whole-hand Tmax did not significantly distinguish flare from controls ( $p = 0.53$ ). A leakage-free analysis ( $n = 26$  visits) found no model exceeded baseline for 3-class severity classification. Reclassifying activity using a binary treat-to-target threshold yielded a non-significant trend toward improved classification with random forest (73.1% vs. 57.7% baseline,  $p = 0.08$ ). A single whole-hand temperature value does not reliably distinguish flare from non-flare states. Fine-grained classification was not achievable at this sample size once leakage was corrected. A coarser threshold showed a promising but inconclusive trend, warranting validation in a larger cohort.

## Key Points

- A single whole-hand maximum temperature value did not distinguish RA flare from control states, indicating that global hand temperature is an insufficient biomarker for flare detection.
- Correcting patient-level data leakage in a machine learning severity-classification pipeline eliminated an initially apparent predictive signal, underscoring the necessity of patient-stratified validation in small thermal-imaging studies.
- A coarser, clinically established binary disease-activity threshold (EULAR treat-to-target) showed a stable but statistically inconclusive trend toward improved classification (73.1% vs. 57.7% baseline), suggesting possible utility for coarse-grained activity screening pending validation in a larger cohort.

## 1. Introduction

Rheumatoid arthritis (RA) is a chronic inflammatory arthropathy in which timely detection of disease flares informs treatment decisions. Standard flare assessment relies on composite clinical indices such as the Disease Activity Score in 28 joints (DAS28) [1], incorporating tender/swollen joint counts and acute-phase reactants (ESR or CRP).

A limitation of standard flare-assessment tools is their reliance on components that are either subjective or non-specific to RA: patient global assessment is inherently a self-reported measure subject to individual variation in symptom perception and reporting, while ESR and CRP are acute-phase reactants

that can be elevated by infection, malignancy, other inflammatory or autoimmune conditions, and even obesity, independent of RA-specific joint inflammation. This creates diagnostic ambiguity in patients with overlapping conditions, where an elevated ESR or CRP may not reliably indicate an RA flare. An objective, joint-localized adjunct measure could help disambiguate RA-driven inflammation from these confounding elevations.

These measures also require in-person joint examination and blood draws, motivating interest in non-invasive adjuncts.

Infrared thermal imaging captures cutaneous temperature, which may rise with underlying synovial inflammation, consistent with the classical clinical sign of calor (heat) in inflamed joints [2]; modern infrared cameras allow this cutaneous heat signal to be quantified objectively and non-invasively [17].

A substantial body of work has compared thermographic parameters between RA patients and healthy or osteoarthritic controls, and correlated them with clinical and ultrasonographic measures of synovitis, with mixed results regarding sensitivity and specificity for detecting inflammatory activity [3]. At the hand and wrist, RA patients have shown higher small-joint thermal parameters than healthy controls [4], and thermographic hand/wrist parameters have correlated with power Doppler and grey-scale ultrasound synovitis [6, 9, 10]; however, other work has questioned whether clinically "hot" joints reliably correspond to objectively elevated temperature, describing this association as partly a "myth" in some inflammatory-arthritis contexts [5]. Foot-level studies have reported elevated dorsal-foot temperatures in RA compared with controls, though correlation with ultrasound-detected inflammation at individual joints has been inconsistent [7, 8]. At the knee, thermographic parameters, including maximum, minimum, and average temperature and knee–thigh temperature differentials, have distinguished RA patients from controls with high reported diagnostic accuracy [11] and correlated with colour Doppler ultrasound findings [12]; a large recent cross-sectional study of 95 patients (190 knees) similarly found significant correlations ( $r = 0.21–0.49$ ) between thermal parameters and ultrasound-detected inflammation, with areas under the receiver-operating-characteristic curve (AUC) of 0.63–0.82 for detecting Doppler-positive synovitis [20].

Building on these joint-level associations, several groups have applied automated feature extraction and machine learning to thermal images in an effort to move beyond raw temperature comparisons toward RA activity classification. Early work extracted textural and statistical features from segmented hand thermal images to grade RA severity [15, 16] and applied infrared image analysis to estimate RA activity directly [13]. More recently, a machine-learning-derived composite score, the Thermographic Joint Inflammation Score (ThermoJIS), was developed in a cohort of 449 subjects and validated in 146 RA patients against ultrasound, achieving 94% sensitivity and 51% specificity (AUC 0.78) for active synovitis, improving to 87% sensitivity and 82% specificity when an indeterminate zone was introduced [18]; a related composite index incorporating patient global assessment and CRP (ThermoDAI/ThermoDAI-CRP) correlated strongly with DAS28-CRP, CDAI, and SDAI in the same cohort [14], and both indices were subsequently externally validated in an independent, prospective 77-patient longitudinal cohort [19]. Neural-network-based classifiers applied to combined static and post-vasodilator dynamic thermal

imaging have also been reported, with one recent study achieving 93% accuracy using a support vector machine to classify inflammation severity by Ritchie Articular Index from resting-state thermal features in a 70-patient cohort [21].

Across this literature, however, a recent narrative review noted mixed findings across studies—including inconsistent correspondence between thermographic and clinical/ultrasound measures of joint inflammation—and concluded that further validation work is needed before thermal imaging's role in RA can be established [3]. A 2026 systematic review of 19 studies comparing infrared thermography with joint ultrasound in inflammatory arthritis (13 RA-predominant) reported sensitivity ranging from 79–100% and specificity from 51–94% across studies, and attributed this wide range substantially to methodological heterogeneity in imaging protocols, small sample sizes, and the general absence of pre-specified diagnostic cut-offs, concluding that these limitations likely overestimate true test performance and preclude meta-analytic pooling [22]. Machine learning approaches applied to thermal images have also been explored for automated severity grading; however, such pipelines can be vulnerable to insufficiently patient-independent train/test splitting—a distinct validation concern not addressed in either review above, and one that we identify and correct directly in our own analysis (Section 2.6.2).

This study had two aims, motivated by the need for a more objective adjunct to the subjective and non-specific components of standard flare assessment: (1) to evaluate whether a single whole-hand maximum temperature value distinguishes RA flare visits from seronegative subjects without inflammatory arthritis, and (2) to determine, using a leakage-free, patient-stratified analytic approach, whether per-joint thermal zone data can classify RA disease activity severity.

## **2. Materials and Methods**

### **2.1 Study Design and Participants**

This proof-of-concept study enrolled 11 seropositive RA patients and 9 seronegative control subjects without inflammatory arthritis, under IRB protocol #202102571. RA patients were imaged and clinically assessed across 2–3 visits: an initial visit during active flare, a subsequent treatment-response follow-up visit, and, for patients who achieved clinical remission during the study period ( $n = 4$ : H1, LA1, C1, PA1), a third visit specifically capturing the remission state. For these patients, the remission-state visit represents the same hand and same anatomically-fixed thermal zones as their baseline flare-state visit, imaged under a non-flare condition. RA patients were identified as being in an active flare state at their first (baseline) visit only, with subsequent visits reflecting treatment response. Controls were imaged once for comparison. Reporting follows the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guideline for observational cohort studies, adapted as applicable given the study's exploratory, proof-of-concept design and inclusion of a machine learning classification component.

### **2.2 Thermal Imaging Protocol**

Digital thermal imaging was performed using the WellVu system under standardized conditions across all patients and visits: ambient temperature 68.0°F, imaging distance 1 m, relative humidity 50%, and emissivity 1.0.

Each hand image was annotated with 12 anatomically-fixed thermal zones. Zones 1–11 corresponded to specific joint regions (wrist, MCP, and PIP/DIP joints as anatomically assigned by the WellVu software). **Zone 12 was confirmed to be an ambient/room-temperature reference marker rather than a joint zone** – across every patient and visit, zone 12 readings were substantially cooler (approximately 71–82°F) than all other zones (approximately 86–99°F), and its spatial marker consistently fell outside the hand contour in the thermal image. Zone 12 was therefore excluded from all diagnostic and machine-learning feature sets, leaving 11 joint-based zones.

For each zone, minimum, average, and maximum temperature (Min/Avg/Max, °F) were recorded per hand per visit, yielding 33 features per hand-visit (11 zones × 3 statistics).

## 2.3 Clinical and Laboratory Assessment

DAS28-ESR, ESR, CRP, and swollen/tender joint counts were recorded at each visit. Disease activity was categorized using standard DAS28 thresholds: remission/low activity ( $\leq 3.2$ ), moderate activity (3.2–5.1), and high/severe activity ( $> 5.1$ ). For the primary 3-class machine learning task, these were collapsed to "mild" ( $\leq 3.2$ ), "moderate" (3.2–5.1), and "severe" ( $> 5.1$ ). For a secondary binary task, the EULAR treat-to-target threshold was applied: "low disease activity" ( $\leq 3.2$ ) versus "active disease" ( $> 3.2$ ).

For the four patients (H1, LA1, C1, PA1) who contributed a third visit, this visit was clinically identified as a remission state based on a reduction in DAS28-ESR relative to baseline, patient-reported overall improvement, and resolution of joint swelling and tenderness on examination, rather than a formal ACR/EULAR Boolean remission definition.

## 2.4 Image and Data Exclusion Criteria

At some visits, an RA patient's hand was imaged twice: once using the standard protocol for RA-severity feature extraction, and once as a separate comparison image intended to contrast a joint in documented remission (i.e., neither tender nor swollen on examination) against a separate, incidental site of osteoarthritis (OA) on the same hand. These additional comparison images were identified by explicit filename labeling (e.g., "OA," "NML," "control") – a labeling convention referring to the image's comparison purpose, distinct from the seronegative control subjects described in Section 2.1 – and were excluded from the RA-severity feature set. Only the standard (non-comparison) image from each visit was retained for analysis.

One visit (patient C1, December 11, 2023) contained only OA-comparison-labeled images in the structured zone-data export; the corresponding standard images for that visit (labeled 0001\_IR and 0005\_IR in the original WellVu PDF report) were present only in the PDF export and were manually

transcribed into the analysis dataset, with hand laterality (right/left) confirmed against source documentation.

## 2.5 Primary Analysis: Whole-Hand Tmax

Whole-hand maximum temperature (Tmax) was compared between RA flare visits and seronegative control participants using Welch's t-test.

## 2.6 Secondary Analysis: Machine Learning Classification

**2.6.1 Prior exploratory analysis and identified limitation.** An initial exploratory analysis attempted RA severity classification directly from hand thermal images (33 images total, at 1 image per hand per visit) using a convolutional neural network (EfficientNetB2) and classical machine learning models. This analysis achieved 57% and 28% accuracy across two model configurations, with one model degenerately predicting a single class for all inputs. Critically, **images were split into train/validation/test sets at the image level rather than the patient level**, allowing the same patient's images to appear across multiple splits. This constitutes patient-level data leakage and invalidates the reported performance as a measure of generalization to unseen patients. This analysis is reported here for transparency but is superseded by the corrected analysis below.

**2.6.2 Corrected, leakage-free analysis.** Per-joint zone temperature data (Section 2.2) were extracted for all 11 RA patients across all available visits (2–3 visits each; total  $n = 26$  visit-level records after applying exclusion criteria and incorporating the manually-transcribed C1 visit). Each patient contributed a baseline flare-state visit and a treatment-response follow-up visit; four patients (H1, LA1, C1, PA1) additionally contributed a third visit corresponding to clinical remission. Within this dataset, lower-disease-activity visits — including the four remission-state visits and other visits meeting the low-disease-activity threshold — represent the same patient's hand imaged in a non-flare state, functioning as a within-patient comparison to that patient's own baseline flare-state imaging, distinct from the separate, cross-sectional seronegative control comparison described in Section 2.5. For each visit, features were constructed as:

- **Full feature set:** 66 features (11 zones  $\times$  3 statistics  $\times$  2 hands), with missing-hand data mean-imputed within each cross-validation training fold.
- **Summary feature set:** 4 aggregate features — mean average-zone temperature, maximum zone temperature, standard deviation of average-zone temperatures, and left–right hand asymmetry (difference in mean average-zone temperature between hands).

Two classifiers were evaluated: logistic regression and random forest (300 trees). Both were evaluated using **leave-one-patient-out cross-validation (LOPO-CV)**, in which all visits from a given patient were held out as the test set in each fold while the model trained on all remaining patients' visits. This ensures no patient's data appears in both training and test sets in any fold, addressing the leakage identified in Section 2.6.1.

Two classification targets were evaluated:

- **3-class severity** (mild/moderate/severe, per Section 2.3)
- **2-class disease activity** (low vs. active, per the EULAR treat-to-target threshold)

Model performance was compared against a majority-class baseline (i.e., the accuracy achievable by always predicting the most common class) to contextualize whether observed accuracy reflected genuine predictive signal. For the 2-class target, a one-sided binomial test was used to assess whether observed accuracy significantly exceeded the majority-class baseline rate.

**2.6.3 Sensitivity analysis.** To assess whether the 2-class LOPO-CV findings were sensitive to the specific patient-fold assignment inherent in leave-one-out validation with a small cohort, a group-stratified k-fold cross-validation (5 folds, 20 repeated randomized patient-to-fold assignments, patient independence preserved throughout) was additionally performed for the 2-class target.

## 3. Results

### 3.1 Clinical Response to Treatment

RA patients demonstrated significant clinical improvement with treatment: DAS28-ESR decreased from 5.23 to 3.32 ( $p = 0.0007$ , paired t-test), and ESR decreased from 80.2 to 54.6 ( $p = 0.012$ ). The reported baseline and endpoint DAS28-ESR and ESR values reflect the mean across all 11 patients' first and second visits, respectively; four patients (H1, LA1, C1, PA1) contributed a third visit, which was incorporated into the machine-learning analysis (Section 2.6.2) but not into this baseline-to-endpoint clinical-response summary. Individual patient trajectories, spanning 2–3 visits each, are shown in Fig. 1.

### 3.2 Whole-Hand Tmax Does Not Distinguish Flare from Control

Whole-hand Tmax did not significantly differ between RA flare visits (98.16°F) and seronegative controls (98.01°F; Welch  $t = 0.63$ ,  $p = 0.53$ ), indicating that a single global temperature summary is insufficient to detect RA flare.

### 3.3 Corrected Machine Learning Analysis

**3.3.1 Dataset.** After excluding OA/NML/control comparison images and incorporating the manually-recovered C1 visit, the final dataset comprised 26 visit-level records across 11 patients (patients contributing 2–3 visits each).

**3.3.2 Three-class severity classification.** Class distribution was mild  $n = 11$ , moderate  $n = 7$ , severe  $n = 8$  (majority-class baseline accuracy = 42.3%). No model exceeded this baseline:

| Feature Set                     | Model                          | LOPO-CV Accuracy |
|---------------------------------|--------------------------------|------------------|
| Full (zones 1–11, per hand)     | Logistic Regression            | 30.8%            |
| Full (zones 1–11, per hand)     | Random Forest                  | 38.5%            |
| Summary (cross-zone aggregates) | Logistic Regression            | 38.5%            |
| Summary (cross-zone aggregates) | Random Forest                  | 30.8%            |
| —                               | <i>Majority-class baseline</i> | 42.3%            |

**3.3.3 Two-class disease-activity classification.** Class distribution was active disease n = 15, low disease activity n = 11 (majority-class baseline accuracy = 57.7%). Three of four model/feature-set combinations exceeded this baseline, with random forest on the full feature set showing the largest margin:

| Feature Set                        | Model                          | LOPO-CV Accuracy | vs. Baseline      | p-value*     |
|------------------------------------|--------------------------------|------------------|-------------------|--------------|
| Full (zones 1–11, per hand)        | Logistic Regression            | 50.0%            | –7.7 pts          | 0.84 (n.s.)  |
| <b>Full (zones 1–11, per hand)</b> | <b>Random Forest</b>           | <b>73.1%</b>     | <b>+ 15.4 pts</b> | <b>0.080</b> |
| Summary (cross-zone aggregates)    | Logistic Regression            | 69.2%            | + 11.5 pts        | 0.161        |
| Summary (cross-zone aggregates)    | Random Forest                  | 69.2%            | + 11.5 pts        | 0.161        |
| —                                  | <i>Majority-class baseline</i> | 57.7%            | —                 | —            |

\*One-sided binomial test against the majority-class baseline rate.

None of these results reached conventional statistical significance ( $p < 0.05$ ) against the majority-class baseline, though the random forest/full-feature-set result showed a trend ( $p = 0.080$ ). Figure 2 displays both the 3-class and 2-class results side by side against their respective baselines, illustrating that every 3-class model/feature combination fell below baseline while three of four 2-class combinations exceeded it.

**3.3.4 Sensitivity analysis.** Group-stratified 5-fold cross-validation (20 repeated randomized patient-to-fold assignments) was performed for the 2-class target to assess robustness of the LOPO-CV estimates above:

| Feature Set                        | Model                | LOPO-CV      | SGKF Mean (SD)      | SGKF Range        |
|------------------------------------|----------------------|--------------|---------------------|-------------------|
| Full (zones 1–11, per hand)        | Logistic Regression  | 50.0%        | 56.2% (5.4%)        | 42.3–65.4%        |
| <b>Full (zones 1–11, per hand)</b> | <b>Random Forest</b> | <b>73.1%</b> | <b>69.0% (3.1%)</b> | <b>65.4–73.1%</b> |
| Summary (cross-zone aggregates)    | Logistic Regression  | 69.2%        | 66.5% (4.2%)        | 61.5–73.1%        |
| Summary (cross-zone aggregates)    | Random Forest        | 69.2%        | 63.3% (5.8%)        | 50.0–73.1%        |

The random forest / full-feature-set result was the most stable across validation schemes, with a narrow spread of outcomes (SD 3.1%) closely tracking the LOPO-CV estimate. This consistency suggests the 73.1% LOPO-CV accuracy was not an artifact specific to the particular leave-one-patient-out fold structure. By contrast, logistic regression on the full feature set showed the widest variability (SD 5.4%, range spanning 42.3–65.4%), indicating this particular result is more sensitive to which patients are grouped together during validation and should be interpreted cautiously. Figure 3 shows the confusion matrix for the best-performing and most stable model (random forest, full features, 2-class target).

## 4. Discussion

This study found that a single whole-hand thermal maximum does not reliably distinguish RA flare from non-flare states, consistent with the biological expectation that RA-driven inflammation is localized to specific joints rather than producing a detectable global hand-temperature elevation.

Notably, the mean endpoint ESR of 54.6 mm/hr remained elevated despite the significant reduction in DAS28-ESR; an ESR at this level contributes approximately 2.7 points to the DAS28-ESR score on its own, meaning that achieving low disease activity ( $\text{DAS28} \leq 3.2$ ) in this cohort depended on comparatively low tender/swollen joint counts and patient global assessment offsetting persistently active acute-phase reactants. This may reflect a cohort-specific pattern – for example, a population with more residual acute-phase elevation than has been typical in prior thermal-imaging RA studies – and could partly explain differences between our findings and those of comparator cohorts with more complete biochemical remission.

For machine-learning-based severity classification, correcting a patient-level data leakage issue present in an earlier exploratory analysis (Section 2.6.1) substantially changed the apparent performance: models that appeared to exceed chance in leakage-affected pilot runs performed **at or below a majority-class baseline** once patient-independent validation was enforced for fine-grained (3-class) severity classification. This underscores the importance of patient-stratified validation in small thermal-imaging ML studies, where apparent predictive signal can arise entirely from within-patient similarity across visits rather than genuine disease-state discrimination.

Reframing the classification target using a clinically established binary threshold (EULAR treat-to-target,  $\text{DAS28} \leq 3.2$  vs.  $>3.2$ ) produced a more encouraging, though still statistically inconclusive, result: random forest applied to the full per-zone feature set achieved 73.1% accuracy against a 57.7% baseline ( $p = 0.080$ ). Notably, this result proved the most stable of the four model/feature-set combinations under a sensitivity check using group-stratified repeated k-fold cross-validation (mean 69.0%, SD 3.1%, range 65.4–73.1% across 20 randomized patient-to-fold assignments), suggesting the finding is not an artifact of the particular patient grouping produced by leave-one-out validation. By contrast, logistic regression on the same full feature set showed considerably more instability under this check (LOPO 50.0% vs. SGKF mean 56.2%, SD 5.4%), and should not be interpreted as a meaningful result in either direction.

A likely contributor to the 3-class model's chance-level performance is the narrow spread of baseline disease-activity scores relative to the classification thresholds. Four patients (BA1, C1, AL1, and W1) had baseline DAS28-ESR scores within 0.05 points of the moderate/severe boundary ( $\text{DAS28} = 5.1$ ), and a fifth (H1) fell within 0.16 points of the same boundary; overall, 9 of 11 patients had baseline scores within 0.7 DAS28-ESR units of a class boundary. Given this clustering near threshold values, even a highly accurate underlying predictor of disease activity would be expected to misclassify many visits under a 3-class scheme, independent of any limitation in the thermal features themselves. This provides a plausible explanation for why the coarser 2-class (EULAR treat-to-target) scheme outperformed the 3-class scheme in our data.

Taken together, these findings suggest that if a thermal signal exists, it may be coarse-grained and better captured by a non-linear model (random forest) than a linear one (logistic regression) — potentially distinguishing "at target" from "active disease" more readily than finer gradations of severity — but the present sample (11 patients, 26 visits) remains underpowered to confirm this beyond a trend, since the LOPO-CV estimate did not reach conventional statistical significance against the majority-class baseline ( $p = 0.080$ ).

An additional strength of the disease-activity classification analysis is that, for several patients, disease-activity labels spanned a within-patient trajectory from active flare to clinical remission using the same hand and thermal zones, providing an internal, paired comparison alongside the cross-sectional seronegative control comparison used in the primary Tmax analysis (Section 2.5).

Our findings of chance-level 3-class severity classification and a non-significant 73.1% binary-classification trend contrast with substantially higher performance reported by some recent thermography-based machine learning approaches in RA, including 93% support-vector-machine accuracy for classifying inflammation severity from resting-state thermal features in a 70-patient cohort [21] and 87–94% sensitivity for the ThermoJIS/ThermoDAI indices in a much larger 449-patient development cohort with 146- and 77-patient validation cohorts [14, 18, 19]. This discrepancy likely reflects, at least in part, our substantially smaller and patient-stratified validation cohort (11 patients, 26 visits) and definitional differences between fine-grained disease-activity severity classification, as attempted here, and binary active-synovitis detection against ultrasound, as reported in most

comparator studies. It is also consistent with the conclusion of a recent systematic review of the broader thermography-versus-ultrasound literature that heterogeneous protocols, small samples, and the general absence of pre-specified cut-offs may inflate apparent diagnostic performance across this field [22], reinforcing our view that our own coarse-threshold trend requires confirmation in a larger, prospectively designed cohort before it can be interpreted as a genuine signal.

Even the coarse binary trend observed here (Section 3.3.3) — if confirmed in a larger, prospectively validated cohort — could offer a more objective, joint-localized adjunct to disambiguate RA-driven inflammation from the non-specific and subjective elevations that can confound ESR, CRP, and patient global assessment in patients with overlapping conditions; a single whole-hand temperature value, however, does not appear suited to this role.

## 4.1 Limitations

- **Small sample size.** Eleven patients and 26 total visits substantially limit statistical power, particularly for LOPO-CV, in which each fold's test set is a single patient's 2–3 visits.
- **Imaging protocol asymmetry.** Not all visits included bilateral hand imaging in the structured data export; one visit's zone data required manual transcription from the original PDF report, as it was absent from the structured spreadsheet export.
- **Single-center, retrospective analytic correction.** The corrected ML pipeline was developed iteratively following identification of multiple data-parsing and labeling errors (detailed in supplementary methods notes), which were resolved through direct review of source PDF reports; while corrected, this iterative process underscores the importance of independent data validation in future prospective work.
- **Coarse severity categorization.** DAS28 was collapsed into 2–3 discrete categories for classification purposes; this simplification, while necessary given sample size, discards continuous disease-activity information that a regression-based approach might better exploit in a larger cohort.
- **No held-out validation cohort.** All reported accuracy reflects cross-validation within the same 11-patient cohort; external validation in an independent cohort was not performed.
- **Device-level constraint on pixel extraction.** Zone-level Min, Avg, and Max temperatures were derived from the WellVu system's native per-zone outputs; the device does not expose raw pixel-level thermal data, precluding a custom pixel-extraction approach (e.g., a small-area statistic such as the highest mean temperature within a 3×3-pixel window) that might offer a more noise-robust estimate of true peak tissue temperature than the device's built-in zone statistics. Additionally, the relative contribution of Min, Avg, and Max to model performance (e.g., via feature importance) was not separately examined in this analysis. Future work using devices with pixel-level access should consider evaluating alternative extraction methods and reporting per-statistic feature importance to clarify which zone-level summary carries the most diagnostic signal.

## 4.2 Strengths

- Per-joint (not whole-hand-only) thermal data was analyzed, using real calibrated zone temperatures rather than raw image pixel intensities.
- Imaging conditions (ambient temperature, distance, humidity, emissivity) were standardized across all patients and visits, addressing a limitation identified in an earlier draft of this work.
- Machine learning validation strictly enforced patient-level independence (LOPO-CV), directly addressing and correcting a data leakage issue identified in prior exploratory work.
- Zone 12 was correctly identified as an ambient reference rather than a joint measurement, resolving apparent "sensor error" data points in two patients that had previously been flagged as a data-quality concern.
- For a subset of patients ( $n = 4$ ), thermal and clinical data spanned a within-patient trajectory from active flare to clinical remission, providing a paired comparison state using the same hand and thermal zones alongside the primary cross-sectional control comparison.

## 5. Conclusion

In this proof-of-concept cohort, a single whole-hand temperature value did not distinguish RA flare from control, and fine-grained (3-class) severity classification from per-joint thermal zone data did not exceed chance performance once patient-level data leakage was corrected. A coarser, clinically meaningful binary activity threshold (treat-to-target vs. active disease) showed a promising trend with random forest classification (73.1% vs. 57.7% baseline;  $p = 0.080$ ) that remained stable under a group-stratified cross-validation sensitivity check (mean 69.0%, SD 3.1%), lending some additional confidence that this is not a validation-scheme artifact — though the finding still falls short of conventional statistical significance at this sample size. These findings suggest thermal imaging's potential utility, if any, may lie in coarse activity-status screening rather than fine-grained severity grading, and support validation in a larger, prospectively designed cohort before clinical application.

## Declarations

### Ethics Approval

This study was approved by the Institutional Review Board of the North Florida/South Georgia Veterans Health System (IRB Protocol #202102571) and was performed in accordance with the ethical standards of the responsible institutional research committee and with the 1964 Declaration of Helsinki and its later amendments.

### Consent to Participate

Informed consent was obtained from all individual participants included in the study.

### Author Contributions

N.G.N. conceived and designed the study, acquired and analyzed the data, and drafted the manuscript. M.B. reviewed the manuscript.

### **Data Availability Statement**

The datasets analyzed in this study, including de-identified per-zone thermal imaging data and clinical/laboratory measures, are available from the corresponding author upon reasonable request, subject to institutional and IRB data-sharing restrictions applicable to protocol #202102571.

### **Funding**

This research was supported by BRAVE funding from the U.S. Department of Veterans Affairs (TTP-001-023), Technology Transfer Program

### **Competing Interests**

On behalf of all authors, the corresponding author states that there is no conflict of interest, beyond the equipment loan and training from WellVu disclosed below.

WellVu provided the thermal imaging device used in this study free of charge and provided initial training on image acquisition. WellVu had no role in study design, data collection, analysis, interpretation, or the decision to submit this manuscript for publication. Beyond the device loan and initial training, the authors declare no competing interests.

### **Acknowledgments**

The authors thank WellVu Dr Ron Riegel for providing the thermal imaging device used in this study and for initial training on image acquisition technique. The authors also thank Rama Garikapati for assistance with initial coding work for this project.

### **Declaration of AI and AI-Assisted Technologies in the Research and Writing Process**

Generative AI (Claude, Anthropic) assisted in: (1) debugging and iteratively developing the data extraction and machine learning pipeline (Section 2.6.2), including identification and correction of data-parsing errors; (2) transcribing per-zone thermal values from a source PDF report for one visit where structured data was unexpectedly absent (Section 2.4); (3) statistical framing, including binomial significance testing against majority-class baselines and paired t-test verification of clinical-response statistics; and (4) drafting and revising manuscript text based on author-provided data and decisions.

All AI-assisted outputs were reviewed, verified, and corrected by the human author(s) before inclusion. This process identified and corrected multiple AI-processing errors, including: a parsing routine that failed to detect left-hand thermal data present in source files; a labeling routine that mismatched disease activity scores to visits by list position rather than date, which — combined with a separate source-file error labeling a right-hand image as "left hand" — caused one patient's post-treatment visit to be initially

assigned the wrong disease activity score; and, on one occasion, purported code-execution output that could not be reproduced and did not reconcile with its own reported statistics, which was rejected. All final results were independently verified for internal arithmetic consistency and cross-checked against source documents before acceptance. The corresponding author(s) take full responsibility for the accuracy, integrity, and scientific validity of this manuscript, consistent with current authorship guidance on AI use (e.g., ICMJE, COPE).

## References

1. Prevoo, M. L. et al. Modified disease activity scores that include twenty-eight-joint counts. Development and validation in a prospective longitudinal study of patients with rheumatoid arthritis. *Arthritis Rheum.* **38**, 44–48 (1995).
2. Karimbux, N. Y. Calor, dolor, rubor, tumor. *J. Dent. Educ.* **78**, 1243 (2014).
3. Kow, J. & Tan, Y. K. An update on thermal imaging in rheumatoid arthritis. *Joint Bone Spine.* **90**, 105496 (2023).
4. Uğur, S., İrim, Y., Yücel, A. A., Carlak, H. F. & Kaçar, C. A comparison of thermal characteristics of the small joints of the hands between patients with rheumatoid arthritis and healthy controls. *Arch. Rheumatol.* **39**, 617–623 (2024).
5. Jones, B. et al. Hot joints: myth or reality? A thermographic joint assessment of inflammatory arthritis patients. *Clin. Rheumatol.* **37**, 2567–2571 (2018).
6. Gatt, A. et al. A comparison of thermographic characteristics of the hands and wrists of rheumatoid arthritis patients and healthy controls. *Sci. Rep.* **9**, 17204 (2019).
7. Gatt, A. et al. Thermal characteristics of rheumatoid feet in remission: Baseline data. *PLoS One.* **15**, e0243078 (2020).
8. Gizińska, M., Rutkowski, R., Szymczak-Bartz, L. & Romanowski, W. Straburzyńska-Lupa. Thermal imaging for detecting temperature changes within the rheumatoid foot. *J. Therm. Anal. Calorim.* **145**, 77–85 (2021).
9. Tan, Y. K. et al. Thermography in rheumatoid arthritis: a comparison with ultrasonography and clinical joint assessment. *Clin. Radiol.* **75**, 963e17–963e22 (2020).
10. Tan, Y. K., Hong, C., Li, H., Allen, J. C. Jr & Thumboo J. A novel use of combined thermal and ultrasound imaging in detecting joint inflammation in rheumatoid arthritis. *Eur. J. Radiol.* **134**, 109421 (2021).
11. Vasdev, V. et al. Thermal imaging in rheumatoid arthritis knee joints and its correlation with power Doppler ultrasound. *Med. J. Armed Forces India.* **79**, S189–195 (2023).
12. Umapathy, S. et al. Thermography and colour Doppler ultrasound: a potential complementary diagnostic tool in evaluation of rheumatoid arthritis in the knee region. *Biomed. Tech. (Berl)*. **65**, 289–299 (2020).

13. Frize, M. & Ogungbemile, A. Estimating rheumatoid arthritis activity with infrared image analysis. *Stud. Health Technol. Inf.* **180**, 594–598 (2012).
14. Morales-Ivorra, I. et al. A thermographic disease activity index for remote assessment of rheumatoid arthritis. *RMD Open*. **8**, e002615 (2022).
15. Snehalatha, U., Anburajan, M., Sowmiya, V., Venkatraman, B. & Menaka, M. Automated hand thermal image segmentation and feature extraction in the evaluation of rheumatoid arthritis. *Proc. Inst. Mech. Eng. H* **229**, 319–331 (2015).
16. Umapathy, S., Vasu, S. & Gupta, N. Computer aided diagnosis based hand thermal image analysis: a potential tool for the evaluation of rheumatoid arthritis. *J. Med. Biol. Eng.* **38**, 666–677 (2018).
17. Ring, E. F. & Ammer, K. Infrared thermal imaging in medicine. *Physiol. Meas.* **33**, R33–46 (2012).
18. Morales-Ivorra, I. et al. Assessment of inflammation in patients with rheumatoid arthritis using thermography and machine learning: a fast and automated technique. *RMD Open*. **8**, e002458 (2022).
19. Morales-Ivorra, I. et al. External validation of the machine learning-based thermographic indices for rheumatoid arthritis: a prospective longitudinal study. *Diagnostics (Basel)*. **14**, 1394 (2024).
20. Tan, Y. K. & Lim, G. H. Comparing thermographic heat signatures with joint inflammation detected by ultrasound at the knees in patients with rheumatoid arthritis. *Sci. Rep.* **15**, 31960 (2025).
21. Carlak, H. F., Irim, Y., Uğur, S., Kaçar, C. & Yücel, A. A. A method for the assessment of rheumatoid arthritis using neural network supported static and dynamic thermal analysis. *Phys Eng. Sci. Med* (2026).
22. Bastelberger, J. et al. Infrared thermography in the detection of arthritis: a systematic review of diagnostic performance compared with joint ultrasound. *Rheumatol. Int.* **46**, 60 (2026).

## Figures

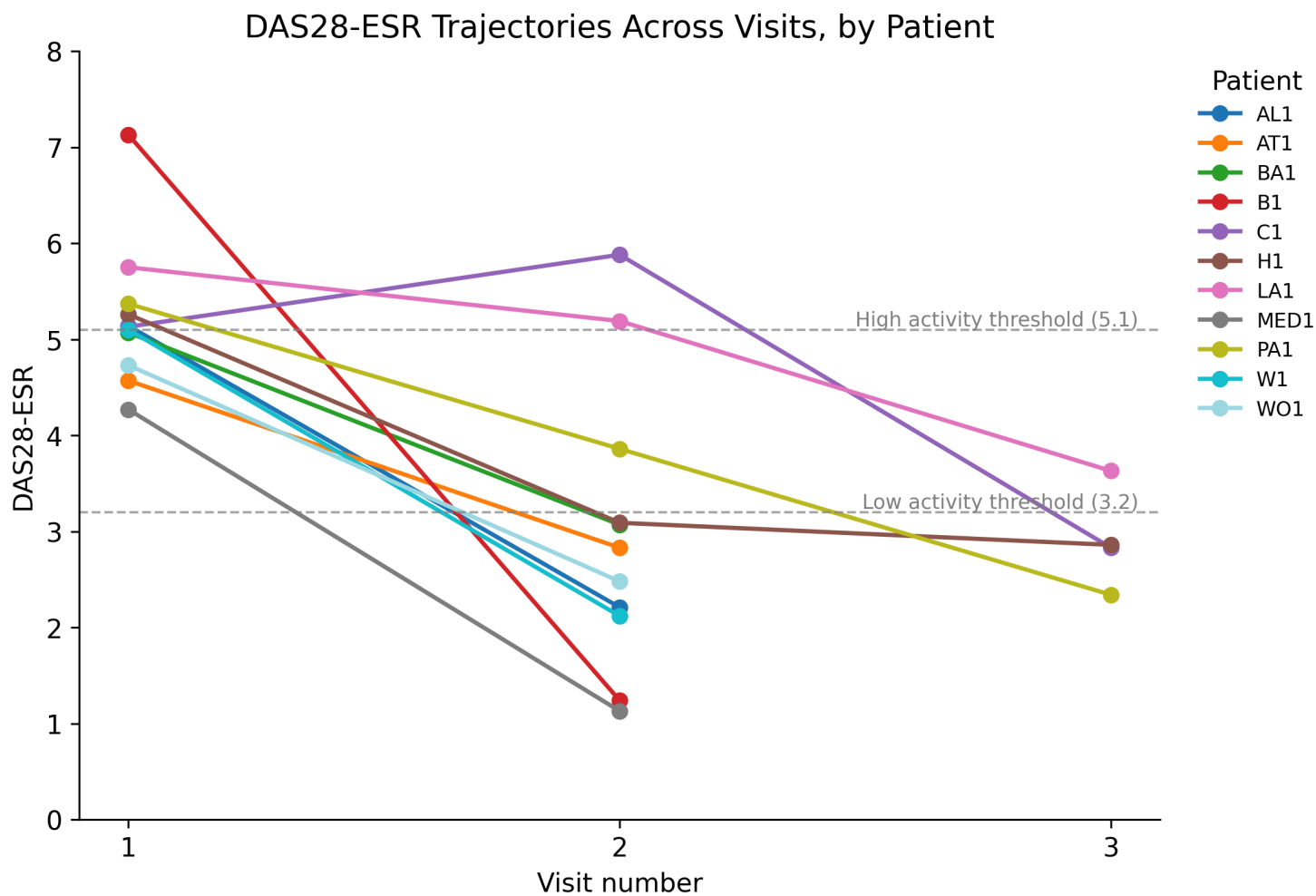


**Figure 1**

DAS28-ESR trajectories across visits, by patient (n=11 patients, 2–3 visits each). Dashed gray lines indicate EULAR disease-activity thresholds used for classification (low activity  $\leq 3.2$ ; high/severe activity  $> 5.1$ ).

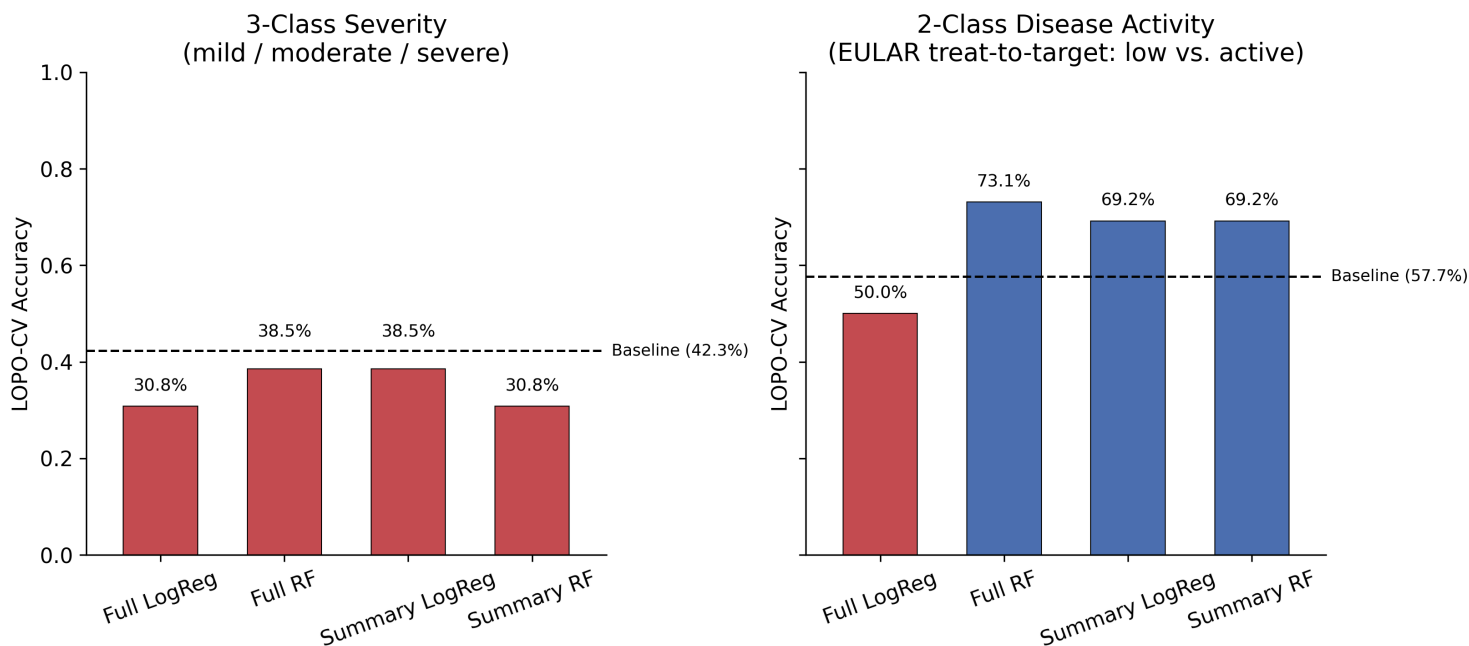


**Figure 2**

Leave-one-patient-out cross-validation (LOPO-CV) accuracy by feature set and model, compared against majority-class baseline (dashed line), for the 3-class severity target (left) and 2-class EULAR treat-to-target activity threshold (right). Red bars fall below baseline; blue bars exceed it.

Random Forest, Full Features  
2-Class (LOPO-CV, n=26)

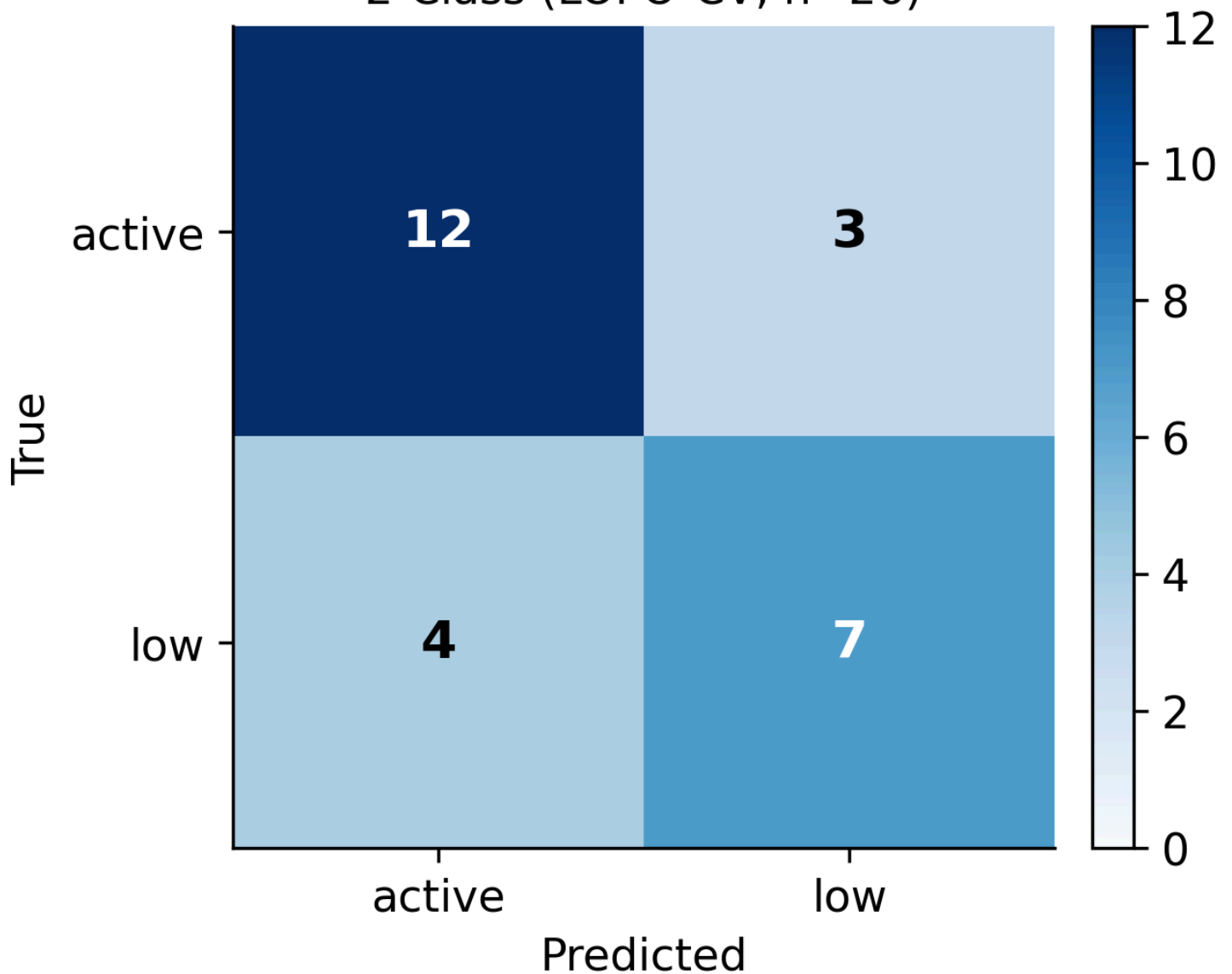


Figure 3

Confusion matrix for the best-performing and most stable model (random forest, full per-zone features, 2-class target), from leave-one-patient-out cross-validation (n=26 visits, 11 patients).