INTL JOURNAL OF ELECTRONICS AND TELECOMMUNICATIONS, 2026, VOL. 72, NO. 1, PP. 1–9
Manuscript received September 15, 2025; revised February, 2026. doi: 10.24425/ijet.2026.157875

Spatiotemporal Deep Learning on dynamic infrared thermography for classification of post-COVID-19 and post-myocardial infarction patients

Jakub Skwierczyński, Krzysztof Krupka, Andrzej Rusiecki, and Łukasz Jeleń

***Abstract*—Dynamic infrared thermography is emerging as a noninvasive technique for monitoring microvascular health, yet its interpretation remains largely qualitative and labor-intensive. This work systematically benchmarks four deep learning architectures: 2D CNN, 3D CNN, CNN–LSTM, and CNN–Transformer, evaluated for automated DIRT sequence classification in a clinically relevant cohort of post-COVID-19 and post-myocardial infarction patients. The study introduces a rigorous pipeline encompassing thermal image acquisition, standardized preprocessing, tailored data augmentation, and stratified cross-validation to ensure reliable evaluation. Purely spatial models such as the 2D CNN underperform, achieving a macro F1 score of 73.5% and accuracy of 80.1%, while temporally aware models yield substantial gains: CNN–LSTM reaches a macro F1 score of 91.4% and accuracy of 92.7%, and the CNN–Transformer achieves 88.8% and 90.6% prior to hyperparameter optimization. After automated hyperparameter optimization, both models converge to a macro F1 score of 93.8% and accuracy of 94.8%, with the Transformer requiring less than half the parameters. Functional ANOVA analysis highlights that learning rate is the most influential factor for LSTM tuning, while dropout dominates for the Transformer. These findings establish a foundation for robust, sequence-aware DIRT analysis, demonstrating that modern deep learning models, when rigorously validated, can transform DIRT into a quantitative biomarker for longitudinal vascular assessment.**

***Keywords*—dynamic infrared thermography; deep learning; thermoregulation analysis; hyperparameter optimization; CNN**

I. Introduction

DYNAMIC infrared thermography (DIRT) tracks the rebound of skin temperature after a brief thermal stimulus and delivers a time-resolved map of microvascular function. The roots of the technique reach back to Sir William Herschel, who in 1800 observed an unseen band beyond red light that raised a thermometer more than any visible color, now known as infrared radiation. A century later, Kálmán Tihanyi showed that this invisible radiation could be captured much like ordinary light by patenting an infrared-sensitive camera in 1929 [1], [2].

Jakub Skwierczyński, Andrzej Rusiecki and Łukasz Jeleń are with the Department of Computer Engineering, Wrocław University of Science and Technology, Wrocław, Poland (e-mail: lukasz.jelen@pwr.edu.pl);

Krzysztof Krupka is with The Karol Godula Upper Silesian Academy of Entrepreneurship in Chorzów, Poland (e-mail: kkrupka@wp.pl)

The modern term DIRT was introduced by de Weerd et al. [3] as an extension of conventional infrared thermography (IRT). In IRT, a single static image merely labels regions as colder or hotter and does not capture microcirculatory dynamics. DIRT improves on this approach by monitoring how surface temperature evolves after a brief thermal challenge. Skin blood flow is the body's primary thermoregulatory reserve, so vessels respond vigorously to metabolic, thermal, and pharmacological stimuli [4]. By recording temperature at every frame, DIRT can quantify small oscillations that reflect vascular tone and can classify longer-term trends as rising, falling, or stabilized. These additional measures reveal subtle changes produced by inflammation, tumor angiogenesis, or dominant perforators, and the technique is now applied to nasal airflow research, perforator planning, joint monitoring, and bedside vascular screening [5]–[8]. As clinical adoption widens, DIRT generates large volumes of time-resolved data for each patient encounter, making manual inspection infeasible and creating a clear need for automated, sequence-aware interpretation.

Despite these advances, interpretation remains largely qualitative in practice. Subtle differences in temperature patterns require models that combine spatial and temporal information across the entire DIRT sequence. Deep architectures such as 3D CNNs, CNN-LSTM hybrids, and CNNs augmented with Transformer encoder blocks can learn these spatiotemporal features directly from raw data [9], [10]. On the same clinical cohort, an initial study [11] evaluated 2D CNN baselines trained on preprocessed DIRT scans, demonstrating the feasibility of deep learning for thermal sequence analysis. However, a systematic evaluation of sequence-aware architectures has yet to be conducted.

Expanding upon that earlier work, this study presents an end-to-end pipeline designed to process complete DIRT sequences, maintain spatial fidelity and temporal consistency through dedicated preprocessing, and benchmark four deep learning architectures on a curated cohort of post-COVID-19 and post-myocardial infarction patients. The findings demonstrate that integrating sequence-based DIRT with modern deep learning frameworks can support the development of a robust, noncontact biomarker for longitudinal assessment of microvascular health, contributing to earlier diagnosis and more efficient clinical workflows.

© The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0, https://creativecommons.org/licenses/by/4.0/), which permits use, distribution, and reproduction in any medium, provided that the Article is properly cited.

II. Materials and Methods

This section describes the detailed methodology developed for analyzing DIRT sequences using advanced deep learning architectures. The primary aim was to establish a robust, automated method to distinguish subtle microvascular differences between patients recovering from COVID-19 and those who had experienced a myocardial infarction. To achieve this goal, an integrated pipeline involving thermal data acquisition, specialized preprocessing techniques, model selection, training procedures, and comprehensive evaluation metrics was established, as outlined in Fig. 1.

A. Research Methodology

The research methodology developed in this study focuses on analyzing temperature variability by comparing two distinct thermal regions: the body's core temperature and cortical temperature. The core temperature was approximated through thermal measurements taken at the forehead's center, known as the glabella. Prior research established that the glabella reliably reflects the body's internal temperature due to its proximity to deeper vascular structures [12]–[14].

In contrast, cortical temperature measurements capture surface level thermal dynamics influenced by multiple factors, including ambient environmental conditions, thermoregulatory processes, and specific measurement methodologies [15]–[17]. After capturing the DIRT sequences, three primary parameters were extracted from the infrared thermal data: the temperature trend indicating directional thermal changes, the absolute temperature values, and the differential adjustments between consecutive temperature measurements.

These collected parameters underwent a factor analysis to explore potential interdependencies among the body's various physiological subsystems. This approach enabled the identification of subtle thermal signatures associated with underlying pathological conditions. Leveraging the sensitivity of dynamic infrared thermography in detecting early shifts in thermoregulatory autoregulation, this methodology provides a noninvasive, radiation-free approach suitable for routine clinical screening and monitoring of microvascular health.

B. Dataset Overview

The dataset used in this study consists of dynamic infrared thermography sequences collected from 96 patients referred for cardiovascular assessment following COVID-19 infection or myocardial infarction. Specifically, the data includes scans from 66 individuals recovering from a COVID-19 infection and 30 individuals who had experienced a myocardial infarction. For each participant, a sequence of eight thermal images was captured, comprising four anterior scans and four posterior scans, as illustrated in Fig. 2. Each individual scan has a spatial resolution of 142×19 pixels.

Due to the retrospective and non-systematic nature of data collection, the available clinical metadata are limited. In particular, clinical information such as age, body mass index, comorbidities, medication status, and exact time since the incident were not consistently available between subjects. No healthy control group was included. Consequently, the analysis focuses on discriminating between patient subgroups present in the dataset rather than on absolute deviations from normal microvascular function. Furthermore, even with the small cohort size and the aforementioned limitations, the dataset reflects real-world data availability and should be viewed as an exploratory benchmark designed to evaluate the feasibility and robustness of sequence-aware deep learning for DIRT analysis, rather than to establish population-level generalizations. Each scanning session consisted of four measurements performed using a thermal imaging camera. The protocol began with an initial baseline measurement immediately after the patient removed clothing, followed by subsequent measurements at standardized intervals: two scans at consecutive 30-second intervals and a final scan after an additional 4-minute interval. This standardized scanning approach ensured consistency across all captured thermal sequences, facilitating accurate and reproducible tracking of temperature recovery dynamics for each participant.

C. Data Preprocessing

Raw DIRT sequences acquired from the thermal microcamera contained inherent variability and noise, such as environmental background, sensor artifacts, and participant-specific temperature fluctuations. To ensure data uniformity and optimize input for effective modeling, a structured multistep preprocessing pipeline was implemented, as depicted in Fig. 1 and exemplified in Fig. 2.

Initially, adaptive foreground segmentation isolated the patient's silhouette from the background. Otsu's thresholding method determined an optimal temperature cutoff automatically, effectively separating warmer body pixels from cooler environmental regions. To prevent loss of peripheral body areas, an ambient-aware threshold adjustment was employed, considering both Otsu's threshold and an ambient-based median temperature offset. Subsequently, connected component analysis retained only the largest connected component, effectively isolating the patient's silhouette from extraneous pixels and artifacts. Morphological hole filling was applied next, ensuring a continuous, solid body mask. Pixels outside the final mask were suppressed to eliminate background interference.

Due to minor variations in initial scan dimensions, each thermal scan was spatially resampled via bilinear interpolation to a standardized resolution of 142×18 pixels, ensuring uniform spatial dimensions across the dataset.

The effectiveness of this preprocessing pipeline is demonstrated in Fig. 3. Prior to processing, pixel intensity distributions exhibited clear bimodal patterns representing ambient temperatures and body surfaces. After preprocessing, background variability was substantially reduced, enhancing the anatomical relevance and signal-to-noise ratio of the resulting data, thereby facilitating reliable downstream modeling.

Following spatial harmonization, each thermal scan underwent temperature intensity normalization by linearly scaling pixel values from the physiologically relevant range to a $[0, 1]$ interval. This normalization aligned pixel intensities with the expected input range of neural network layers and enhanced

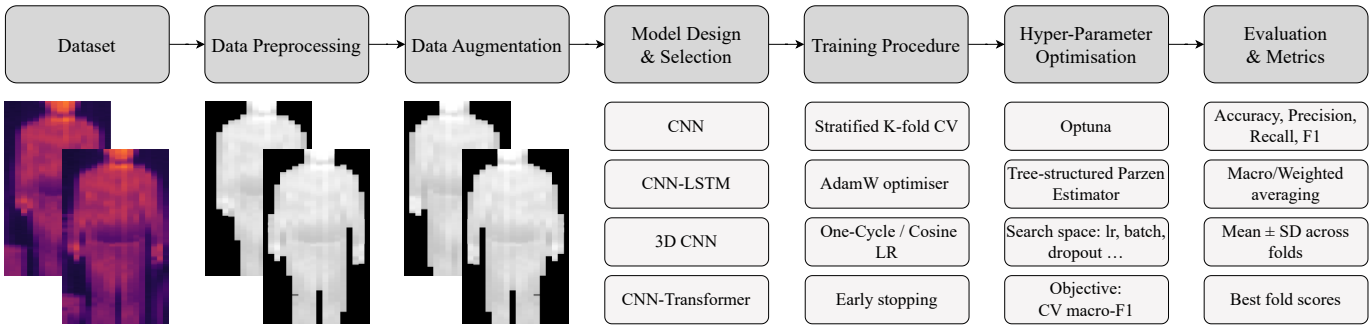


Fig. 1. Overview of the proposed deep learning classification pipeline

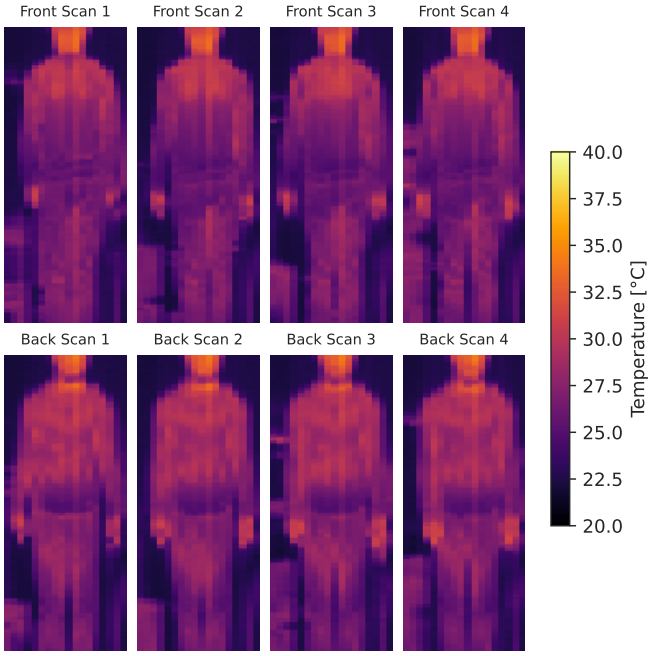


Fig. 2. Exemplary DIRT sequence for a single patient, comprising four anterior and four posterior thermal scans

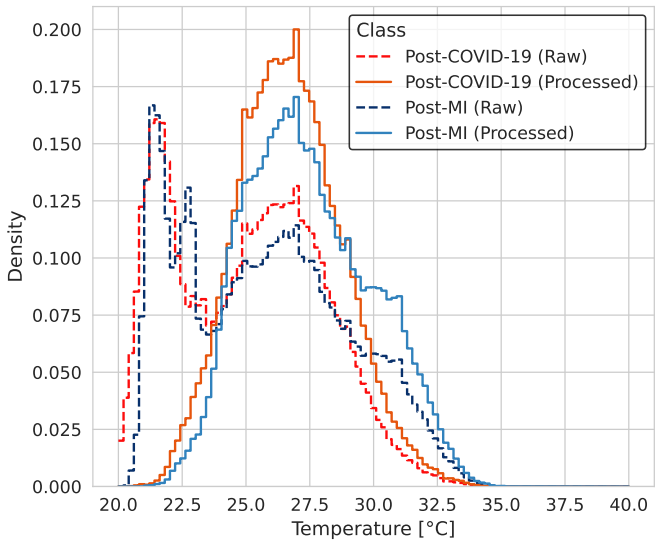


Fig. 3. Pixel temperature distributions before and after preprocessing, comparing Post-COVID-19 and Post-MI thermal scans.

numerical stability. An example of a preprocessed scan is shown in the second step of the pipeline depicted in Fig. 1, where it can be observed that the background was effectively removed, leaving a clearly delineated and normalized patient thermogram ready for subsequent analysis.

Finally, normalized thermal images were structured into tensors suitable for model training. Specifically, four sequentially aligned thermal scans, each consisting of anterior and posterior views concatenated horizontally to form images of dimensions 142×36 pixels, were stacked temporally. This resulted in tensors of shape $4 \times 1 \times 142 \times 36$, capturing comprehensive spatiotemporal information. These tensors served directly as inputs to the deep learning models described in subsequent sections.

D. Data Augmentation

Due to the relatively limited size of the dataset, data augmentation techniques were employed to enhance the generalization capability of the deep learning models and reduce the risk of overfitting [18]. Data augmentation was applied exclusively to the training set within each cross-validation fold, leaving the validation sets untouched, thereby preserving the integrity of model evaluation.

Augmentation procedures included horizontal flipping (illustrated in Fig. 1), random horizontal shifts, additive Gaussian noise, and simulated temperature drift through random scaling. These transformations were applied consistently across each temporal sequence to maintain the temporal coherence of the DIRT data. Specifically, horizontal flipping was randomly performed with a probability of 0.5, while horizontal shifts introduced small random translations of up to ± 3 pixels, simulating realistic sensor positioning variations. Gaussian noise was added to simulate sensor imperfections, with noise drawn from a zero-mean Gaussian distribution. Additionally, global temperature scaling was applied, randomly adjusting each image's temperature values within a $\pm 3\%$ range to mimic minor sensor drift.

E. Deep Learning Architectures

To capture the spatiotemporal patterns in the preprocessed DIRT tensors, we evaluated four complementary neural net-

work architectures. First, a 2D CNN baseline treats the four time steps as input channels and learns purely spatial filters [19]. Second, the CNN–LSTM hybrid augments this by feeding per frame CNN features into an LSTM layer to model temporal dynamics [20]. Third, a 3D CNN applies volumetric convolutions to jointly learn across space and time [9]. Finally, the CNN–Transformer replaces the recurrent module with a self-attention encoder that integrates information across all scans in parallel [21]. The detailed layer configurations and parameter counts for each model are summarized in Tables I–IV. All models were implemented in PyTorch 2.7.1 with Python 3.12.3 [22].

1) 2D CNN: A lightweight two-layer 2D convolutional network was implemented as a spatial baseline. The model treats the four time steps as input channels, applies successive Conv2d–AvgPool2d blocks to extract spatial features, and flattens the result into two fully connected layers before the final softmax. With only 40,946 parameters, this architecture provides a fast reference point for purely spatial classification (Table I) [19].

TABLE I
2D CNN ARCHITECTURE SUMMARY

Layer	Output Shape	Params
Conv2d (4, 32)	[1, 32, 142, 36]	1,184
AvgPool2d	[1, 32, 35, 9]	–
Conv2d (32, 16)	[1, 16, 35, 9]	4,624
AvgPool2d	[1, 16, 8, 2]	–
Flatten	[1, 208]	–
Linear (208, 128)	[1, 128]	26,752
Linear (128, 64)	[1, 64]	8,256
Linear (64, 2)	[1, 2]	130
Total	–	**40,946**

2) CNN–LSTM Hybrid: To capture temporal dependencies, the 2D CNN was augmented with a recurrent module. Frame-wise feature maps are pooled and passed through a linear layer before being fed into a single layer LSTM with 512 hidden units. The LSTM's final state is then classified via a dense layer. This hybrid retains spatial convolutional strengths while explicitly modeling sequence dynamics (Table II) [20].

TABLE II
CNN-LSTM ARCHITECTURE SUMMARY

Layer	Output Shape	Params
Conv2d (1, 32)	[4, 32, 142, 36]	288
BatchNorm2d	[4, 32, 142, 36]	64
Conv2d (32, 64)	[4, 64, 142, 36]	18,432
BatchNorm2d	[4, 64, 142, 36]	128
MaxPool2d	[4, 64, 71, 18]	–
Conv2d (64, 128)	[4, 128, 71, 18]	73,728
BatchNorm2d	[4, 128, 71, 18]	256
MaxPool2d	[4, 128, 35, 6]	–
AdaptiveAvgPool2d	[4, 128, 1, 1]	–
Flatten	[4, 128]	–
Linear (128, 128)	[4, 128]	16,512
LSTM (128, 512)	[1, 4, 512]	790,528
LayerNorm	[1, 512]	1,024
Linear (512, 2)	[1, 2]	1,026
Total	–	**901,986**

3) 3D CNN: A volumetric convolutional architecture was designed to learn spatiotemporal filters directly. The network employs three stages of Conv3d–BatchNorm3d–MaxPool3d, reducing the temporal and spatial dimensions jointly, followed by global average pooling and two fully connected layers. This 3D CNN, with 1.14 M parameters, fuses motion and texture in a single pass (Table III) [9].

TABLE III
3D CNN ARCHITECTURE SUMMARY

Layer	Output Shape	Params
Conv3d (1, 64)	[1, 64, 4, 142, 36]	1,728
BatchNorm3d	[1, 64, 4, 142, 36]	128
MaxPool3d	[1, 64, 2, 71, 18]	–
Conv3d (64, 128)	[1, 128, 2, 71, 18]	221,184
BatchNorm3d	[1, 128, 2, 71, 18]	256
MaxPool3d	[1, 128, 1, 35, 9]	–
Conv3d (128, 256)	[1, 256, 1, 35, 9]	884,736
BatchNorm3d	[1, 256, 1, 35, 9]	512
MaxPool3d	[1, 256, 1, 17, 4]	–
AdaptiveAvgPool3d	[1, 256, 1, 1, 1]	–
Flatten	[1, 256]	–
Linear (256, 128)	[1, 128]	32,896
Dropout	[1, 128]	–
Linear (128, 2)	[1, 2]	258
Total	–	**1,141,698**

4) CNN–Transformer Hybrid: Finally, a hybrid model replaces the recurrent layer with a Transformer encoder. Convolutional blocks produce per scan embeddings that are concatenated and enriched by multihead self-attention across the four time steps. A position-aware Transformer block aggregates these embeddings before classification, enabling flexible modeling of long-range temporal interactions (Table IV) [21].

TABLE IV
CNN–TRANSFORMER ARCHITECTURE SUMMARY

Layer	Output Shape	Params
Conv2d (1, 32)	[4, 32, 142, 36]	288
BatchNorm2d	[4, 32, 142, 36]	64
ReLU	[4, 32, 142, 36]	–
MaxPool2d	[4, 32, 71, 18]	–
Conv2d (32, 64)	[4, 64, 71, 18]	18,432
BatchNorm2d	[4, 64, 71, 18]	128
ReLU	[4, 64, 71, 18]	–
MaxPool2d	[4, 64, 35, 9]	–
Conv2d (64, 128)	[4, 128, 35, 9]	73,728
BatchNorm2d	[4, 128, 35, 9]	256
ReLU	[4, 128, 35, 9]	–
AdaptiveAvgPool2d	[4, 128, 1, 1]	–
Flatten	[4, 128]	–
Linear (128, 256)	[4, 256]	33,024
MultiheadAttention	[1, 4, 256]	263,168
LayerNorm	[1, 4, 256]	512
Linear (256, 256)	[1, 4, 256]	65,792
Dropout	[1, 4, 256]	–
Linear (256, 256)	[1, 4, 256]	65,792
LayerNorm	[1, 4, 256]	512
LayerNorm	[1, 256]	512
Linear (256, 2)	[1, 2]	514
Total	–	**523,746**

F. Training Procedure

Given the limited size of our dataset, stratified 5-fold cross-validation was employed to obtain robust performance estimates and reduce the risk of overfitting [23]. In each fold, the training subset was augmented as described in Section II-D,

while the corresponding validation subset remained unaltered to ensure an unbiased assessment of generalization.

All models were trained to minimize categorical cross-entropy loss using the AdamW optimization algorithm [24]. A one-cycle learning rate policy with cosine annealing was utilized: the learning rate increased linearly to a peak value during the initial phase of training and then gradually decreased to a small final value [25]. This schedule promotes efficient convergence and reduces the likelihood of the model becoming trapped in sharp local minima. Weight decay was uniformly applied to all trainable parameters to serve as a regularization mechanism and to further mitigate overfitting.

Hyperparameter tuning was conducted using the Optuna framework, which employs a Tree-structured Parzen Estimator to explore the search space efficiently [26]. The optimization process considered a range of learning rates, weight decay coefficients, dropout probabilities, and architecture-specific parameters. To accelerate convergence, trials were pruned early based on intermediate validation macro F1 scores, enabling the elimination of poorly performing configurations. For the selected models, the hyperparameter combination that achieved the highest average macro F1 score across the five folds was selected.

Training was carried out for up to 300 epochs per fold. Early stopping was applied if the validation macro F1 score did not improve for 20 consecutive epochs. The model weights corresponding to the epoch with the best validation macro F1 were restored prior to final evaluation. All reported performance metrics, including accuracy, precision, recall, and F1 score, are presented as the mean and standard deviation across the five cross-validation folds, offering a comprehensive evaluation of each model's robustness and generalization performance.

All experiments were conducted on a workstation equipped with an NVIDIA GeForce GTX 1660 GPU using CUDA 12.4.

G. Evaluation Metrics

To comprehensively assess the performance of the trained models on the binary classification task, four commonly used evaluation metrics were employed: accuracy, precision, recall, and F1 score. These measures jointly capture different aspects of predictive behavior, allowing for a nuanced evaluation of model reliability. In medical imaging and diagnostic tasks, relying on a single metric may obscure clinically important tendencies, such as an overemphasis on the majority class or an asymmetry between sensitivity and specificity. Therefore, a balanced set of complementary metrics was adopted to provide a thorough performance characterization.

Accuracy: Accuracy quantifies the overall proportion of correct predictions. It is defined as the number of correctly classified samples divided by the total number of samples:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{1}$$

While intuitive, accuracy alone can be misleading in imbalanced settings, as it may be biased toward the majority class.

Precision: Precision, also known as positive predictive value, measures the proportion of correctly predicted positive cases among all predicted positives:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

High precision implies that the model makes few false positive errors, which is essential when false alarms must be minimized.

Recall: Recall (sensitivity) captures the proportion of true positive samples correctly identified by the model:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

This metric is critical in medical diagnostics, where missing positive cases (e.g., undetected COVID-19) may carry clinical risk.

F1 score: The F1 score is the harmonic mean of precision and recall, providing a balanced measure that considers both false positives and false negatives:

$$\text{F1 score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

The F1 score is especially valuable in imbalanced datasets, as it penalizes extreme disparities between precision and recall.

Averaging Strategy: Given the moderate class imbalance in our dataset (66 Post-COVID-19 vs. 30 Post-MI participants), performance was evaluated using the macro-averaging strategy for precision, recall, and F1 score. This approach computes each metric independently for both classes and then takes their unweighted mean:

$$M_{\text{macro}} = \frac{1}{2}(M_{\text{COVID}} + M_{\text{MI}}) \tag{5}$$

Macro-averaging assigns equal importance to both diagnostic groups, ensuring that the minority class contributes equally to the overall performance estimate. This choice provides a fairer assessment of the model's ability to generalize across classes, avoiding the bias that could arise from class imbalance and aligning with best practices in medical classification tasks where sensitivity to minority cases is clinically relevant.

III. Results

This section reports the empirical performance of the four candidate architectures and analyzes the effect of automated hyperparameter search on the two strongest contenders. Section III-A compares the macro-averaged cross-validation scores of all models, whereas Section III-B focuses on the hyperparameter optimization phase for the CNN-LSTM and CNN-Transformer networks, including functional ANOVA importance and parallel coordinates visualizations.

A. Cross-Validation Performance

2D CNN: The purely spatial baseline achieved an average macro accuracy of 0.80 and an F1 score of 0.74 (Table V). The relatively wide inter-quartile ranges in Fig. 4 highlight its sensitivity to the sampling of training data, confirming that temporal cues are essential for reliable discrimination. The purely spatial baseline, which most closely approximates the classical frame-based DIRT analysis, performed substantially worse than all temporally aware models.

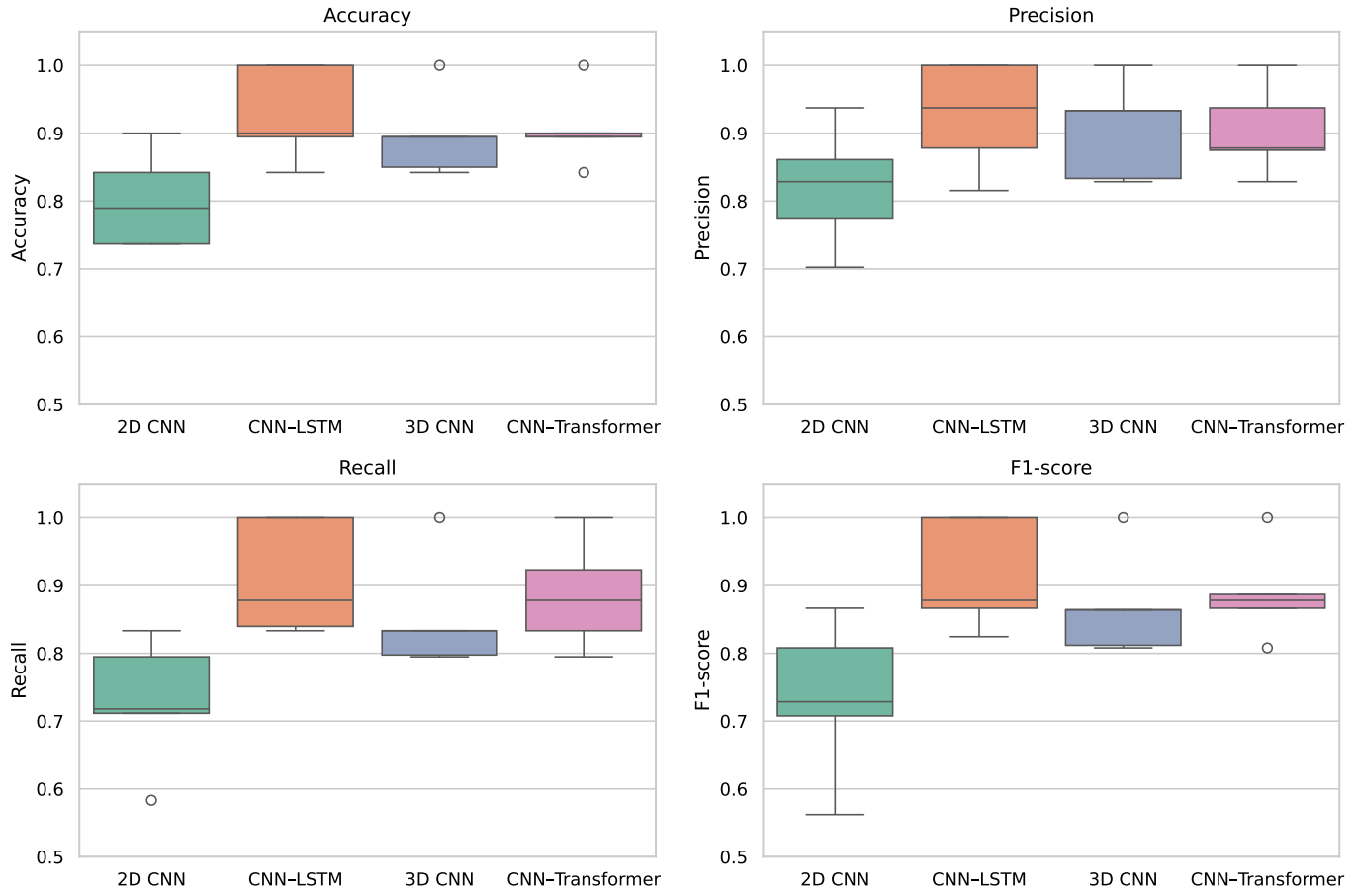


Fig. 4. Distribution of macro-averaged accuracy, precision, recall, and F1 score over the five cross-validation folds for each model

TABLE V
CROSS-VALIDATION RESULTS: MACRO-AVERAGED PERFORMANCE METRICS (MEAN ± STANDARD DEVIATION) ACROSS 5 FOLDS

Model	Accuracy	Precision	Recall	F1 score
2D CNN	0.801 ± 0.063	0.821 ± 0.079	0.728 ± 0.086	0.735 ± 0.103
CNN-LSTM	**0.927** ± 0.063	**0.926** ± 0.072	**0.910** ± 0.075	**0.914** ± 0.073
3D CNN	0.896 ± 0.056	0.906 ± 0.066	0.852 ± 0.076	0.870 ± 0.070
CNN-Transf.	0.906 ± 0.051	0.904 ± 0.059	0.886 ± 0.071	0.888 ± 0.062

CNN–LSTM: Augmenting the spatial encoder with an LSTM yielded the strongest overall performance, achieving a macro F1 score of 0.91, as well as the highest precision (0.93) and recall (0.91) among all tested models (Table V). The narrow boxes in Fig. 4 further indicate excellent fold-to-fold stability, suggesting that the recurrent layer effectively captures the subtle rewarming dynamics characteristic of endothelial injury.

3D CNN: The volumetric model ranked third, reaching a macro F1 score of 0.87 (Table V). Although its accuracy approached that of the CNN–Transformer, the 3D CNN exhibited slightly lower recall (Fig. 4), implying occasional misses of positive cases despite strong spatiotemporal coupling.

CNN–Transformer: Replacing recurrence with self-attention yielded a macro F1 score of 0.89 (Table V). While precision remained high, recall was about 2 percentage points lower than for CNN-LSTM (Fig. 4), likely due to the limited context from only four time steps.

B. Hyperparameter Tuning Outcomes

For the hyperparameter optimization, we selected the two best-performing baseline models from Table V: CNN-LSTM and CNN-Transformer and conducted 50 Optuna trials for each architecture. The search space included the learning rate (10^{-6}–10^{-3}), dropout rate (0–0.5), batch size $\{2, 4, 8\}$, and a set of architecture-specific parameters such as hidden size, Transformer depth, and number of attention heads. The configurations that achieved the highest mean macro F1 scores are listed below, and the corresponding post-optimization results are summarized in Table VI:

- **CNN–LSTM**: $lr = 2.48\times10^{-5}$; *batch size* $= 2$; *dropout* $= 0.44$; *hidden size* $= 512$; $d_{\text{model}} = 128$; *layers* $= 1$.
- **CNN–Transformer**: $lr = 7.29\times10^{-6}$; *batch size* $= 2$; *dropout* $= 0.25$; $d_{\text{model}} = 256$; *heads* $= 2$; *depth* $= 1$.

Both networks converged to an identical macro F1 score of 0.938, indicating that, given sufficient tuning, self-attention mechanisms can match the predictive performance of their recurrent counterparts while using fewer parameters. Notably, the optimal learning rate for the Transformer model was an

TABLE VI
MACRO-AVERAGED 5-FOLD CV AFTER HYPERPARAMETER TUNING

Model	Accuracy	Precision	Recall	F1 score
CNN–LSTM	0.948 ± 0.047	0.955 ± 0.047	0.935 ± 0.062	0.938 ± 0.056
CNN–Transf.	0.948 ± 0.046	0.955 ± 0.047	0.935 ± 0.062	0.938 ± 0.055

order of magnitude lower than that of the LSTM, reflecting its higher sensitivity to optimization dynamics.

Functional ANOVA (fANOVA) [27] was employed to systematically quantify the marginal contribution of each hyperparameter to the observed variation in validation macro F1 scores during the optimization process. This approach decomposes the overall variance in model performance into the relative importance of individual hyperparameters, allowing for a clear attribution of which factors are most influential in tuning deep neural networks.

For the CNN–LSTM model, the results summarized in Fig. 6 show that the learning rate is by far the most influential factor, explaining about 70% of the total variance. This finding reflects the strong sensitivity of recurrent models to optimization parameters, where even small adjustments in the learning rate can substantially affect convergence and generalization. Dropout emerges as the next most significant parameter, albeit with much lower importance (12%), while other architectural choices such as hidden size, LSTM layers, and batch size play only marginal roles in determining the final macro F1 score.

By contrast, for the CNN–Transformer, fANOVA (Fig. 7) reveals a distinct pattern: dropout has the highest importance (58%), suggesting that regularization is paramount for effective training of self-attention architectures in this setting. Transformer-specific parameters such as network depth (13%) and the number of attention heads (7%) also contribute, but to a much lesser extent. The learning rate and batch size, while still relevant, are not as dominant as in the recurrent baseline.

To further clarify the relationship between hyperparameters and model performance, Fig. 5 shows a parallel coordinates plot of the 30 best CNN–Transformer configurations ranked by macro F1 score. Each line represents a single Optuna trial, spanning the most influential hyperparameters, with color intensity indicating the corresponding F1 value. Distinct trends can be observed: the highest-performing runs typically use lower dropout rates around 0.25, a shallow network depth of one layer, learning rates between 10^{-5} and 10^{-4}, and small batch sizes of two. These observations align with the fANOVA analysis, confirming that, under these conditions, stronger regularization or deeper architectures do not provide additional benefits, and that optimal results are achieved with a conservative and well-calibrated learning schedule.

IV. RESULT DISCUSSION

The experimental results presented in Section III provide several important insights into the performance of modern deep learning architectures for automated analysis of dynamic infrared thermography sequences in post-COVID-19 and post-myocardial infarction patients. The findings emphasize how explicitly modeling temporal dynamics contributes to improved diagnostic accuracy and reliability.

A key observation is the clear superiority of sequence-aware models over the purely spatial 2D CNN baseline. The 2D CNN, which uses only spatial information from each frame, consistently underperformed across all main metrics, particularly in recall and F1 score, and exhibited higher variability between folds. These outcomes align with clinical understanding that temporal rewarming dynamics contain crucial microvascular cues, which cannot be captured by static spatial representations alone. Similar trends were previously noted in [11], where 2D CNNs trained on preprocessed thermograms achieved an overall accuracy of approximately 85%, with weaker sensitivity for COVID-19 detection. Preserving the full thermal information improved classification in that study, yet the absence of temporal modeling limited its discriminative power. The present findings confirm and extend these observations by showing that explicit sequence modeling substantially enhances robustness and recall.

Introducing a recurrent LSTM component led to the best overall results, with the CNN–LSTM model achieving the highest and most stable macro-averaged metrics. The model effectively captured both global and subtle rewarming trends, enabling robust classification even under moderate class imbalance. The low variance across folds demonstrates strong generalization and resilience to variations in patient data. Compared with earlier spatial CNN approaches [11], incorporating temporal memory mechanisms yields a notable gain in both F1 and recall, underscoring that rewarming dynamics carry significant diagnostic information beyond spatial structure. Such stability and consistency are crucial for real-world clinical deployment, where reproducibility across patient populations determines reliability.

The 3D CNN architecture, which fuses spatial and temporal dimensions through volumetric convolutions, ranked between the purely spatial and recurrent models. While its overall accuracy surpassed that of the 2D CNN, its recall remained slightly lower than that of the CNN–LSTM. This suggests that, although 3D convolutions can integrate space and time implicitly, they may struggle to capture finer temporal dependencies that recurrent models learn more effectively. These results parallel earlier conclusions that richer thermal inputs improve accuracy, but explicit sequence modeling remains essential for optimal performance.

Following systematic hyperparameter optimization, the CNN–Transformer achieved a macro F1 score comparable to the recurrent model while requiring substantially fewer parameters: $523{,}746$ versus $901{,}986$ for the LSTM. This efficiency, coupled with competitive accuracy, highlights the promise of attention mechanisms for medical imaging scenarios where model size and computational cost are critical factors. The functional ANOVA analysis further indicated that Transformer performance depends primarily on dropout and network depth, while LSTM performance is driven by learning rate sensitivity. In contrast to [11], where model tuning was limited to basic parameter adjustment and oversampling strategies, the present analysis systematically explores the optimization landscape, offering a clearer understanding of how architecture-specific hyperparameters shape model behavior.

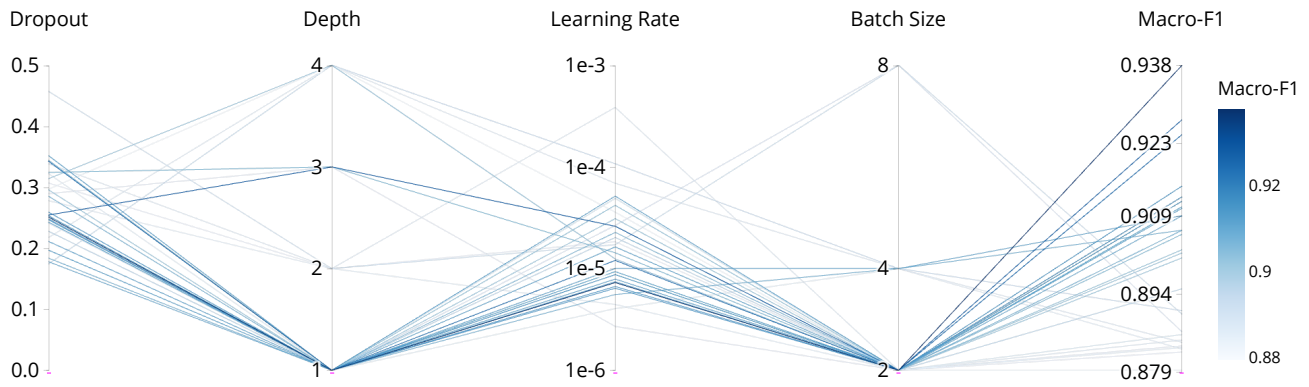


Fig. 5. Parallel coordinates plot of the 30 best CNN–Transformer configurations

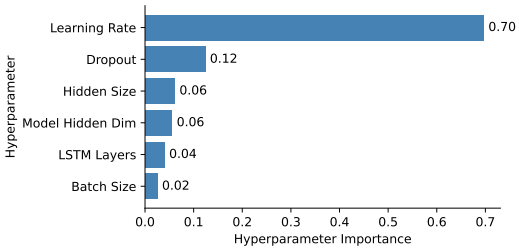


Fig. 6. Functional ANOVA importance of the tuned hyperparameters for the CNN-LSTM model

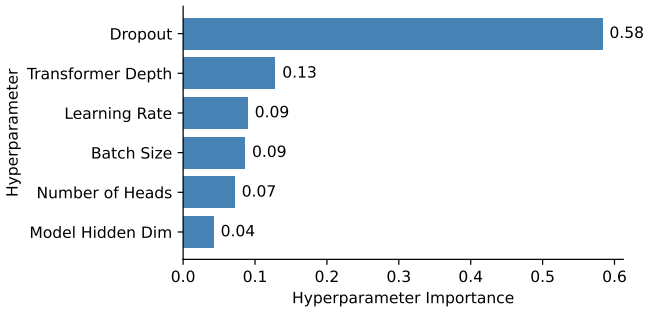


Fig. 7. Functional ANOVA importance of the tuned hyperparameters for the CNN-Transformer model

From a broader methodological perspective, the combination of automated hyperparameter search, variance-based importance analysis, and stratified cross-validation provides a robust and transparent framework for benchmarking. The observed best-performing configurations, characterized by moderate dropout, shallow depth, small batch sizes, and conservative learning rates, promote strong generalization, particularly in studies with limited data availability. These systematic insights extend beyond previous DIRT analyses and offer practical guidance for deploying deep learning models in thermal imaging pipelines.

Overall, the results demonstrate that incorporating temporal information and principled optimization significantly improves the automation and consistency of DIRT-based diagnostics. When viewed alongside earlier findings [11], the evidence indicates a clear progression: preserving rich thermal information is essential, yet only through explicit temporal modeling and careful regularization can consistently high macro F1 performance be achieved. The comprehensive methodology adopted here, including cross-validation and transparent evaluation, enhances reproducibility and strengthens the case for DIRT as a quantitative, sequence-driven biomarker of microvascular health.

Despite the encouraging performance of the proposed method, several limitations should be considered when interpreting these results. First, the small dataset that does not include a healthy control group, and provides only limited clinical characteristic. Some of the important factors were not explicitly modeled, although they are known to influence microvascular reactivity and thermal recovery dynamics. As a result, part of the observed class separation may reflect latent confounders or cohort-specific characteristics rather than disease-specific effects alone. In light of the above, even if cross-validation reduces overfitting, the reported results should be interpreted as indicative of separability within the studied cohorts and may not directly generalize to more heterogeneous clinical populations. Finally, although deep sequence models clearly outperform purely spatial baselines, comparisons with simpler, interpretable classical approaches remain an important direction for future work.

V. Conclusions

This study presents a comprehensive benchmarking of modern deep neural architectures for automated classification of dynamic infrared thermography sequences in the context of post-COVID-19 and post-myocardial infarction patient classification. The experiments demonstrate that effective modeling of temporal information is indispensable for this task. The purely spatial 2D CNN baseline, which achieved a mean macro-averaged F1 score of 0.735 ± 0.103 and recall of 0.728 ± 0.086, consistently lagged behind models capable of learning spatiotemporal dependencies. Incorporating temporal structure through volumetric convolutions, recurrence, or self-attention mechanisms resulted in significant improvements in model performance. The 3D CNN increased the macro F1 score to 0.870 ± 0.070, while the CNN-Transformer reached

0.888 ± 0.062, with both models also showing higher precision and accuracy.

The highest performance was achieved by the CNN-LSTM hybrid, which, prior to hyperparameter optimization, attained a macro F1 score of 0.914 ± 0.073 and recall of 0.910 ± 0.075, demonstrating excellent stability and generalization across folds. After automated hyperparameter search, both the recurrent (CNN-LSTM) and self-attention (CNN-Transformer) architectures converged to virtually identical mean macro F1 scores of 0.938 and accuracy of 0.948, while maintaining high precision and recall (see Table VI). This outcome underscores the ability of careful optimization and regularization to bridge the performance gap between recurrent and attention-based architectures, with the Transformer reaching equivalent predictive accuracy while using roughly 58% of the parameters required by the LSTM model.

These findings are consistent with earlier work on 2D CNN baselines trained on preprocessed thermograms [11], which suggested that preserving rich thermal information improves classification. The present benchmarking extends that evidence by demonstrating that explicit sequence modeling is a key driver of the additional improvements in macro F1 and recall.

Looking forward, future work should extend these approaches to larger and more diverse patient populations, explore the feasibility of multiclass and regression tasks, and focus on integrating model interpretability to facilitate adoption in clinical workflows. The present results provide compelling evidence that sequence-aware deep learning, when thoroughly validated, can transform DIRT from a qualitative imaging tool into a robust and quantitative biomarker for longitudinal microvascular health monitoring.

REFERENCES

[1] W. Herschel, "Experiments on the solar, and on the terrestrial rays that occasion heat; with a comparative view of the laws to which light and heat, or rather the rays which occasion them, are subject, in order to determine whether they are the same, or different. part ii. by william herschel, ll. d. f. r. s," *Philosophical Transactions of The Royal Society of London*, vol. 90, pp. 437–538, 01 1800. [Online]. Available: https://doi.org/10.1098/rstl.1800.0020

[2] E. F. J. Ring, "The historical development of thermal imaging in medicine," *Rheumatology*, vol. 43, no. 6, pp. 800–802, 06 2004. [Online]. Available: https://doi.org/10.1093/rheumatology/keg009

[3] L. de Weerd, J. B. Mercer, and S. Weum, "Dynamic infrared thermography," *Clinics in Plastic Surgery*, vol. 38, no. 2, pp. 277–292, Apr. 2011. [Online]. Available: https://doi.org/10.1016/j.cps.2011.03.013

[4] L. A. Holowatz, C. S. Thompson-Torgerson, and W. L. Kenney, "The human cutaneous circulation as a model of generalized microvascular function," *Journal of Applied Physiology (1985)*, vol. 105, no. 1, pp. 370–372, Jul. 2008, epub 2007 Oct 11. [Online]. Available: https://doi.org/10.1152/japplphysiol.00858.2007

[5] L. de Weerd, S. Weum, and J. B. Mercer, "The value of dynamic infrared thermography (dirt) in perforator selection and planning of free diep flaps," *Annals of Plastic Surgery*, vol. 63, no. 3, pp. 274–279, Sep. 2009. [Online]. Available: https://doi.org/10.1097/SAP.0b013e3181b597d8

[6] O. Hennessy and S. M. Potter, "Use of infrared thermography for the assessment of free flap perforators in autologous breast reconstruction: A systematic review," *JPRAS Open*, vol. 23, pp. 60–70, 2020. [Online]. Available: https://doi.org/10.1016/j.jpra.2019.11.006

[7] J. Lindemann, K. Wiesmiller, T. Keck, and K. Kastl, "Dynamic nasal infrared thermography in patients with nasal septal perforations," *American Journal of Rhinology & Allergy*, vol. 23, no. 5, pp. 471–474, Sep. 2009. [Online]. Available: https://doi.org/10.2500/ajra.2009.23.3351

[8] A. Casas-Alvarado, A. Ogi, D. Villanueva-García, J. Martínez-Burnes, I. Hernández-Avalos, A. Olmos-Hernández, P. Mora-Medina, A. Domínguez-Oliva, and D. Mota-Rojas, "Application of infrared thermography in the rehabilitation of patients in veterinary medicine," *Animals*, vol. 14, no. 5, 2024.

[9] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," 2015. [Online]. Available: https://doi.org/10.48550/arXiv.1412.0767

[10] P. Ramachandran, N. Parmar, A. Vaswani, I. Bello, A. Levskaya, and J. Shlens, "Stand-alone self-attention in vision models," 2019. [Online]. Available: https://doi.org/10.48550/arXiv.1906.05909

[11] Ł. Jeleń, K. Krupka, and A. Rusiecki, "Dynamic infrared thermography and machine learning for advanced diagnostic applications in biomedical imaging," in *Advances in Dependable Systems and Networks*, W. Zamojski, J. Mazurkiewicz, J. Sugier, T. Walkowiak, and J. Kacprzyk, Eds. Cham: Springer Nature Switzerland, 2025, pp. 75–84. [Online]. Available: https://doi.org/10.1007/978-3-031-92734-8_8

[12] L. P. J. Teunissen, J. Klewer, A. de Haan, J. D. de Koning, and H. A. M. Daanen, "Non-invasive continuous core temperature measurement by zero heat flux," *Physiological Measurement*, vol. 32, no. 5, pp. 559–570, 2011. [Online]. Available: https://doi.org/10.1088/0967-3334/32/5/005

[13] A. Chen, J. Zhu, Q. Lin, and W. Liu, "A comparative study of forehead temperature and core body temperature under varying ambient temperature conditions," *International Journal of Environmental Research and Public Health*, vol. 19, p. 15883, 2022. [Online]. Available: https://doi.org/10.3390/ijerph192315883

[14] S. Mendt, M. A. Maggioni, M. Nordine, M. Steinach, O. Opatz, D. L. Belavy, D. Felsenberg, J. Koch, P. Shang, H. C. Gunga, and A. C. Stahn, "Circadian rhythms in bed rest: Monitoring core body temperature via heat-flux approach is superior to skin surface temperature," *Chronobiology International*, vol. 34, no. 5, pp. 666–676, 2017. [Online]. Available: https://doi.org/10.1080/07420528.2016.1224241

[15] W. Bierman, "The temperature of the skin surface," *JAMA*, vol. 106, no. 12, pp. 1158–1162, 1936. [Online]. Available: https://doi.org/10.1001/jama.1936.02770140020007

[16] R. Lenhardt and D. Sessler, "Estimation of mean body temperature from mean skin and core temperature," *Anesthesiology*, vol. 105, no. 6, pp. 1117–1121, 2006. [Online]. Available: https://doi.org/10.1097/00000542-200612000-00011

[17] P. Shilco, Y. Roitblat, N. Buchris, J. Hanai, S. Cohensedgh, E. Frig-Levinson, J. Burger, and M. Shterenshis, "Normative surface skin temperature changes due to blood redistribution: A prospective study," *Journal of Thermal Biology*, vol. 80, pp. 82–88, 2019. [Online]. Available: https://doi.org/10.1016/j.jtherbio.2019.01.009

[18] Z. Wang, P. Wang, K. Liu, P. Wang, Y. Fu, C.-T. Lu, C. C. Aggarwal, J. Pei, and Y. Zhou, "A comprehensive survey on data augmentation," 2025. [Online]. Available: https://doi.org/10.48550/arXiv.2405.09591

[19] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[20] B. Abdelhalim and F. Titouna, "Automatic sports video classification using cnn-lstm approach," 12 2023.

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 6000–6010.

[22] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," 2019. [Online]. Available: https://doi.org/10.48550/arXiv.1912.01703

[23] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*, ser. IJCAI'95. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1995, p. 1137–1143.

[24] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2019. [Online]. Available: https://doi.org/10.48550/arXiv.1711.05101

[25] L. N. Smith, "Cyclical learning rates for training neural networks," 2017. [Online]. Available: https://doi.org/10.48550/arXiv.1506.01186

[26] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," 2019. [Online]. Available: https://doi.org/10.48550/arXiv.1907.10902

[27] F. Hutter, H. Hoos, and K. Leyton-Brown, "An efficient approach for assessing hyperparameter importance," in *Proceedings of the 31st International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, E. P. Xing and T. Jebara, Eds., vol. 32, no. 1. Bejing, China: PMLR, 22–24 Jun 2014, pp. 754–762.