Journal of Soft Computing Exploration

Homepage: htttps://shmpublisher.com/index.php/joscex

p-ISSN: 2746-7686
e-ISSN: 2746-0991

.

Lightweight thermogram classification for diabetic foot screening: A handcrafted-feature and RFE-SVM pipeline for low-resource settings

Bedy Purnama[1,2], Bayu Erfianto[1], Linlin Lindayani[3]
[1]School of Computing, Telkom University, Indonesia
[2]Center of Excellence Artificial Intelligence for Learning and Optimization, Telkom University, Indonesia
[3]Sekolah Tinggi Ilmu Keperawatan PPNI Jawa Barat, Indonesia

Article Info

Article history:

Received May 1, 2026
Revised June 23, 2026
Accepted July 7, 2026

Keywords:

Diabetic foot thermograph
Handcrafted feature engineering
Recursive feature elimination
Support vector machine
Nested cross-validation

ABSTRACT

Diabetic foot problems continue to be a significant cause of morbidity among people with diabetes, requiring accurate and scalable methods for early identification. Infrared thermography is a non-invasive technique for detecting asymmetry in plantar temperature, which is associated with the risk of inflammation. However, the use of infrared thermography is limited by the small size of labeled datasets in the biomedical domain, posing issues for the generalizability of models and the reliability of validation. In this work we aim at studying the effectiveness of a carefully verified handcrafted-feature based model for diabetic foot thermogram categorization with limited data. The suggested pipeline combines multi-descriptor feature representation, Recursive Feature Elimination (RFE) and margin-based classification by means of Support Vector Machines in a nested cross-validation setting, together with bootstrap stability analysis. Experiments on a publicly accessible plantar thermogram dataset (N = 167) show strong performance with holdout AUC of 0.942 and nested cross-validation AUC of 0.959 ± 0.045. Calibration findings show a low Brier score of 0.065 with balanced sensitivity (0.88) and specificity (0.889). Further study shows that discriminative performance stems from distributed gradient–texture interactions, not sparse feature subsets. The results show that a careful validation and a systematic design of the representation can overcome the limitations of data and allow the development of reliable thermographic screening systems.

Corresponding Author:

Bedy Purnama,
School of Computing,
Telkom University,
Bandung 40257, Indonesia.
Email: bedypurnama@telkomuniversity.ac.id
https://doi.org/10.52465/joscex.v7i3.81

1. INTRODUCTION

Diabetic foot problems continue to be a major global health problem and are among the main causes of non-traumatic lower limb amputations in the world [1], [2]. Recent studies have found that many people with diabetes develop diabetes-related peripheral neuropathy, and the International Diabetes Federation says that in many groups the prevalence of diabetic foot ulcers is often between 10% and 30%. [2] These

consequences are related to an increased risk of hospitalization, amputation, worse quality of life and premature death [2], [3], [4]. Therefore, early diagnosis of plantar temperature asymmetry has become an important component of preventative diabetic foot care efforts [5].

In recent times, infrared thermography has been established as a non-invasive, contactless imaging technique used to identify modest thermal changes related to inflammation and microvascular dysfunction [6]. Temperature changes usually occur before tissue damage becomes obvious. Therefore, thermography can be used for early screening and longitudinal evaluation of diabetic foot problems in addition to standard visual inspection [3], [7]. The availability of portable thermal cameras and smartphone based imaging equipment has increased further the research on automated thermogram analysis [8].

Alongside these technological improvements, deep convolutional neural networks (CNNs) have become the main approach for thermographic image classification [6], [9]. Transfer learning with pretrained architectures such as DenseNet, ResNet and EfficientNet has showed good performance on small- and medium-sized datasets [8], [10]. The classification performance in thermographic diabetic foot studies has further been enhanced using ensemble classifiers using CNN-extracted features in hybrid frameworks [11].

However, the performance of deep learning models is still strongly dependent on the dataset size, annotation quality, and evaluation technique [12], [13]. In biomedical applications with limited samples, high-capacity networks are more vulnerable to overfitting and unstable decision boundaries, especially without sufficient separation of model optimization and evaluation [14], [15], [16]. Recent reviews in medical AI have further illustrated that methodologic concerns such as non-nested cross-validation and data leaking can lead to large inflation of performance reported [17], [18], [19].

The Plantar Thermogram Database for the Study of Diabetic Foot Complications [20] is a publicly available database and is frequently used as a benchmark in thermographic research on diabetic foot complications. However, the dataset is limited to 167 thermograms, which is a typical high dimensional low sample size (HDLSS) learning challenge. In this case, the main problems are featuring redundancy, overfitting and validation leakage and model performance is sensitive to feature representation, classifier selection and validation process [21], [22], [23].

These issues have reawakened interest in hand-crafted feature engineering in the context of limited-data medical imaging [15], [21], [24]. Descriptors such as Histogram of Orientated Gradients (HOG) [25], Local Binary Patterns (LBP) [26], and Gray-Level Co-occurrence Matrix (GLCM)-based texture features [27] provide computationally efficient and interpretable representations. Handcrafted features with efficient feature selection and strong classifiers can also reach competitive performance, which also offers higher transparency and practical benefits for deployment [28], [29].

At the same time, the focus has changed from the complexity of models to the rigor of evaluation in medical AI [17], [18]. Reliable performance estimate requires validation techniques that reduce optimism bias and increase reproducibility. Therefore, nested cross-validation, independent holdout testing, and bootstrap-based confidence intervals are widely recommended as best practices for constructing clinically reliable machine learning models [14], [15], [16].

Despite the increasing number of studies on thermographic diabetic foot categorization, some key gaps still exist. First, little attention has been paid to the effect of feature dimension and representation compactness on model stability and discriminative performance in HDLSS contexts. Second, validation procedures lack consistency with numerous studies relying on single train-test splits or non-nested cross-validation, which could lead to optimistic and less reproducible results. Finally, despite the dominance of deep learning, very few systematic research have been performed to compare structured handcrafted feature pipelines with rigorous bias-controlled validation, preventing a fair comparison of their effectiveness in thermographic classification with limited data.

These restrictions combined imply current gains are likely being driven more by architectural complexity than a fundamental grasp of representation design and evaluation dependability. This poses an interesting question if more straightforward and more interpretable methods may reach similar levels of speed under strict validation standards and at the same time provide more transparency and resilience.

Thus, this study aims to answer the question of whether a well-structured handcrafted-feature framework accompanied by bias-controlled evaluation techniques may give reliable and generalizable classification of diabetic foot thermograms in limited-sample biomedical contexts. In particular, the question examined in this paper is: Can a strongly validated handcrafted representation that incorporates multi-descriptor features and systematic feature selection lead to stable and generalizable performance under HDLSS conditions?

The proposed approach addresses this question by favoring methodological rigor over architectural complexity, through the combination of multi-descriptor feature representation, the Recursive Feature Elimination (RFE) for the reduction of the feature space, the margin-based classification, and a strictly nested cross-validation scheme with embedded feature selection and bootstrap-based stability estimation.

The study has four primary contributions. First, it proposes a validation-centered approach that explicitly decouples model selection from performance estimation via a tiered validation strategy, minimizing optimism bias and boosting the dependability of performance estimations. Second, it rigorously investigates the trade-off between feature dimensionality and classification performance, providing empirical evidence on the impact of feature compactness on model stability in thermographic diabetic foot classification. Third, it shows that the discriminative information is spread out over complimentary gradient and texture descriptors and is not localised in a small collection of features, thereby providing insights into effective feature representation. Finally, we propose an interpretable and computationally efficient handmade pipeline that can reach competitive performance and be used in resource-constrained clinical situations.

In summary, this work highlights that dependable validation and effective feature representation are as critical as model complexity for establishing trustworthy and operationally deployable artificial intelligence systems in small-sample medical imaging.

2. METHOD

Problem Formulation and Framework Overview

The classification of diabetic foot thermograms is defined as a supervised binary classification job with a goal to classify diabetic and non-diabetic plantar thermograms based on handcrafted feature representations. We denote the training dataset as

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N} \tag{1}$$

Where $x_i \in \mathbb{R}^d$ indicates the feature vector taken from the i-th plantar thermogram, and $y_i \in \{0,1\}$ denotes the corresponding class label (0 = non-diabetic, 1 = diabetic). Learning objective: estimate a classifier

$$f: \mathbb{R}^d \to \{0,1\} \tag{2}$$

They are capable of effective classification of unknown thermograms and can generalise well in the high-dimensional low-sample-size (HDLSS) situation. This goal is attained by minimising a regularised empirical risk function.

$$\min_{f \in \mathcal{H}} \frac{1}{N} \sum_{i=1}^{N} \quad \ell(y_i, f(x_i)) + \Omega(f) \tag{3}$$

Where $\ell(\cdot)$is a loss function for classification and $\Omega(f)$ is a regularisation term to control the complexity of the model.

In recent works in diabetic thermography, deep convolutional neural network (CNN) architectures have been increasingly used for automated categorisation and risk assessment [6]. These techniques have demonstrated a promising performance by automatically learning the features from thermographic pictures. However, CNN-based models generally require larger annotated datasets, significant computational resources, and rigorous validation procedures to guarantee reliable generalisation [12]. The learning problem in this study is in the high-dimensional low-sample-size (HDLSS) regime since the publicly available plantar thermogram dataset used herein only contains 167 pictures. In this case, handcrafted feature representations combined with feature selection and margin-based classifiers still provide a methodologically suitable alternative because of their interpretability, computational efficiency, and robustness in limited-data settings [14]. We thus propose a hierarchical handmade representation and tiered validation to provide a methodological robustness and an impartial performance estimation.

The suggested framework is composed of five successive steps. After preprocessing and normalisation, the thermograms are turned into complimentary handcrafted feature representations of gradient, texture and structure information. Then, the Recursive Feature Elimination (RFE) method is utilised to select the most informative subset to discard the redundant features for classifier training. The framework is tested by layered cross-validation and bootstrap analysis to give a reliable estimate of generalisation performance with very low optimism bias.

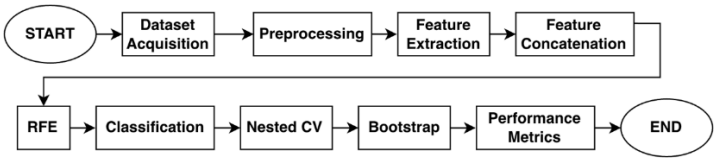
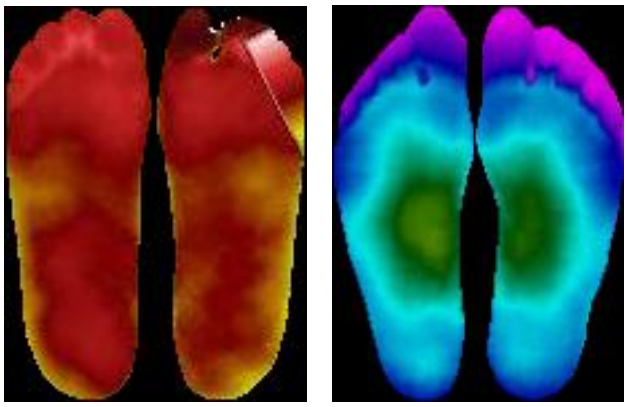


Figure 1. Proposed thermogram classification framework

The overall workflow of the proposed thermogram classification framework is illustrated in Figure 1. We acquire the dataset first by collecting plantar thermogram images from the publicly available database. After that, the images are preprocessed including converting them to greyscale, resizing them and standardizing them so that the input representations are uniform. Then, several hand-crafted descriptors are extracted to capture complementary gradient and texture characteristics. These descriptors are then concatenated to obtain a single feature vector, which is then refined using Recursive Feature Elimination (RFE) to remove redundant and weakly informative features. Then the selected features are used to train the classification models. We adopt nested cross-validation to obtain unbiased performance estimates, where feature selection and hyperparameter optimization are performed in the inner loop, and generalization assessment in the outer loop. Finally, we test the statistical stability of the model by bootstrap resampling. Finally, the effectiveness of classification is evaluated by computing the performance metrics such as accuracy, sensitivity, specificity, F1-score and AUC. The sequential procedure was designed to reduce selection bias and increase reproducibility in biomedical datasets with limited samples.

Dataset and Preprocessing

The studies were conducted on the Plantar Thermogram Database for the Study of Diabetic Foot Complications [20]. This is easily available and has 167 thermographic images (122 diabetic and 45 non-diabetic). Thermal imaging has been identified as a non-invasive modality for detection of temperature asymmetry associated with ulcer risk and inflammation [3, 5].



(a) (b)

Figure 2. Representative examples of plantar thermograms from the dataset: (a) Diabetic subject and (b) Non-diabetic subject

Representative thermograms from the data set used in this study are shown in Figure 2. The diabetic thermogram (Figure 2a) shows increased thermal asymmetry and heterogeneous temperature distribution in the plantar regions, which is often related to inflammation, neuropathy and altered vascular circumstances. In contrast, the thermogram of non-diabetics (Figure 2b) shows a more uniform and symmetric temperature distribution between the respective plantar sites. The visual variations are indicative of employing gradient and texture based feature descriptors giving complementary information to detect thermographic abnormalities associated with diabetic foot problems.

All images were converted to grey scale and resized to 128 × 128 pixels. No histogram equalisation, contrast enhancement or temperature normalisation were applied to preserve the original thermal distribution.

The stratified 80:20 train-test split was used to preserve class proportions. Because of the imbalance of classes, the Synthetic Minority Over-sampling Technique (SMOTE) [30] was only applied to the training data. SMOTE generates synthetic minority samples by interpolating between neighbouring instances in feature space, an improvement in class balance without resorting to simple duplication. Feature standardisation and SMOTE were applied strictly within the training partition to avoid information leakage.

Multi-Descriptor Feature Representation

Thermal patterns in diabetic foot imaging are characterised by structural gradients, local texture irregularities and spatial intensity correlations. Five descriptor families were incorporated to capture these complementary properties.

Histogram of Orientated Gradients (HOG) extracts structural edge information [25]. Gradient components are computed as:

$$G_x = I(x+1, y) - I(x-1, y), G_y = I(x, y+1) - I(x, y-1) \tag{4}$$

With magnitude and orientation:

$$M(x, y) = \sqrt{G_x^2 + G_y^2}, \theta(x, y) = \tan^{-1}(\frac{G_y}{G_x}) \tag{5}$$

The HOG descriptor extracted 1764 features per image with nine orientation bins and L2-Hys normalisation.

Local Binary Patterns (LBP) encode micro-textural variations robustly [26]:

$$LBP_{P,R}(g_c) = \sum_{p=0}^{P-1} \quad s(g_p - g_c)2^p \tag{6}$$

The Modified Block LBP is extended to multiple spatial scales, yielding 356 features.

Modified Block LBP (MBLBP) is an extension of the conventional LBP in which the comparisons of individual pixels are replaced by the comparisons between the average intensity values of local blocks. MBLBP does not directly estimate $g_p - g_c$, but computes mean intensities within pre-defined rectangular regions, which captures macro-textural structures and minimises sensitivity to pixel-level noise. Let μ_c be the average intensity of the central block and μ_p the average intensity of the p-th neighbouring block formally. The MBLBP operator is defined by

$$MBLBP_{P,R}(\mu_c) = \sum_{p=0}^{P-1} \quad s(\mu_p - \mu_c)2^p \tag{7}$$

Where $s(\cdot)$ is the same thresholding function as in Eq. (6). MBLBP summarizes intensity information over spatial neighborhoods and therefore captures larger scale thermal patterns that may reflect regional inflammation and temperature asymmetry in plantar thermograms. In this work, we proposed multi-scale block configurations to extract 356 MBLBP based features which complement the fine-grained texture description of traditional LBP.

The second order texture statistics are measured by Haralick texture features, which are derived from the Gray-Level Co-occurrence Matrix (GLCM) [27]. Thirteen features averaged across orientations.

Dominant temperature distributions are represented by centroids from k-means clustering (k = 8). All the descriptors were concatenated to a feature vector of 2150 dimensions.

Feature Selection and Classification

Handcrafted high dimensional vectors may be redundant or weakly informative, which can inflate model variance, deteriorate the stability of the margin, and increase the risk of overfitting, especially in the high dimensional low sample size (HDLSS) setting. In these cases, high feature dimensionality may hide discriminative structure and degrade generalisation performance. For this purpose, recursive feature elimination (RFE) was implemented using a linear SVM estimator [21]. In each iteration the features with the smallest absolute weight coefficients were eliminated, thus progressively keeping the most discriminative subset:

$$S_{t+1} = S_t \setminus \{j : | \, w_j \, | \ \text{minimal}\} \tag{8}$$

To study the trade-off between the feature dimensionality and the classification performance, three possible feature subsets were evaluated: 100, 300 and 500 features. Then, the classification models were trained

using the selected feature subsets. The main classifier was an SVM which determines the best separating hyperplane by solving

$$\min_{w,b,\xi} \frac{1}{2} || w ||^2 + C \sum_{i=1}^{N} \quad \xi_i \tag{9}$$

subject to:

$$y_i(w^T \phi(x_i) + b) \geq 1 - \xi_i \tag{10}$$

Where C controls the regularization trade-off [31]. The Radial Basis Function kernel is defined as:

$$K(x_i, x_j) = \exp(-\gamma \parallel x_i - x_j \parallel^2) \tag{11}$$

Where γ controls the effect of each training sample on the final decision boundary. For baseline comparison, a K-Nearest Neighbours (KNN) classifier was also implemented with k=5.

Nested Cross-Validation and Performance Estimation

Medical imaging datasets are particularly prone to optimism bias when feature selection and hyperparameter optimisation are done outside the validation phase, especially when the datasets are small [14]. We performed nested 5-fold cross-validation to provide an unbiased estimate of model performance. Nested cross-validation separates model optimisation and performance evaluation, which is commonly considered best practice for small-sample biomedical machine learning [15].

In this approach, the inner loop was utilized for Recursive Feature Elimination (RFE) and classifier hyperparameter adjustment, whereas the outer loop was employed solely for estimating generalization performance. In this method feature selection is completely embedded within the inner folds, avoiding information leaking from the validation data to the model selection process. The separation is designed to enable an unbiased and statistically accurate estimate of the predictive performance

The model performance was evaluated in terms of different complementing metrics. These are accuracy, precision, recall (sensitivity), specificity, F1-score, and the Area Under the Receiver Operating Characteristic Curve (AUC). Accuracy is the overall fraction of correct thermogram classifications. Precision is the number of correctly predicted diabetes cases to the total predicted diabetic cases . Recall (sensitivity) is the number of correctly predicted diabetic cases to the total actual diabetic cases . Specificity is the capacity to correctly identify those without diabetes. The F1-score is the harmonic mean of precision and recall and is a good measure to employ when the class distribution is uneven. AUC measures the model's capacity to discriminate between diabetes and non-diabetic thermograms across all feasible classification levels. AUC was chosen as the primary discrimination metric of all metrics because of its resistance to class imbalance and threshold independence [16]

Metric definitions follow standard formulations:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{12}$$

$$Recall = \frac{TP}{TP+FN}, Specificity = \frac{TN}{TN+FP} \tag{13}$$

Where TP, TN, FP and FN are true positive, true negative, false positive and false negative counts respectively. In practical terms, TP stands for diabetic thermograms that are correctly identified as diabetic, TN stands for non-diabetic thermograms that are correctly classified as non-diabetic, FP stands for non-diabetic thermograms that are incorrectly classified as diabetic, and FN stands for diabetic thermograms that are missed by the classifier. In screening applications, the reduction of FN is of particular importance as missed diabetic cases may delay preventive intervention.

To further assess statistical stability, bootstrap resampling (1000 iterations) was conducted on the final trained pipeline to estimate 95% confidence intervals for AUC. Bootstrap analysis provides an empirical approximation of performance variability and is particularly appropriate in limited-data biomedical studies.

Methodological Positioning

Methodologically, the suggested framework emphasizes the rigor of validation design and the efficiency of deployment-oriented aspects, instead of architectural originality. The main contribution is the systematic integration of different handcrafted descriptors in a bias-controlled tiered validation pipeline for small biomedical datasets.

Unlike other systems that focus on more sophisticated deep neural networks, our framework specifically considers two real-world limitations common to clinical thermography, namely restricted processing resources and small dataset size. Feature selection was only included in inner validation folds to prevent performance inflation and dimensionality-performance trade-offs were examined with several RFE subset sizes (100, 300 and 500 features). This enabled us to explicitly investigate the effect of representation compactness on generalization.

The methodology demonstrates that sophisticated feature representations, combined with rigorous validation, continue to be a feasible and competitive substitute for computationally intensive deep learning models in low-data thermographic screening scenarios, as long as computational efficiency, statistical robustness and interpretability are prioritized.

This work does not claim that handcrafted descriptors are superior to deep learning systems, but that, in constrained data regimes, they are methodologically appropriate. The emphasis is therefore on the validation discipline and deployment practicality that are critical to translational biomedical machine learning.

3. RESULTS AND DISCUSSIONS

A detailed evaluation of the proposed thermogram categorization system is presented in this section. Results are given in a way that respects the whole experimental flow, starting from the analysis of the feature dimensionality, passing through validation stability, independent evaluation, calibration assessment, feature behavior interpretation, compact model feasibility, baseline comparison and error characterization.

In this section, we critically review the experimental results in terms of representation richness, validation rigor, and deployment consequences, rather than simply reporting metrics in isolation.

Feature Dimensionality and RFE Analysis

Due to the feature extraction and concatenation process, every thermogram was represented as a hand-crafted feature vector of 2150 dimensions. The resulting feature matrix had shape $X \in R^{167 \times 2150}$. To eliminate duplication and improve the stability of generalization, Recursive Feature Elimination (RFE) was performed with candidate subset sizes of 100, 300 and 500 features

Table 1 indicates that the discrimination performance improves monotonically with the retention of more features, with the AUC improving from 0.871 for RFE-100 to 0.942 for RFE-500. This tendency shows that the discriminative information is spread among several complementing descriptors rather than in a restricted collection of features. The Youden optimised accuracy was also continuously high, while the default-threshold accuracy declined somewhat with increasing dimensionality. This suggests that the richer feature representations resulted in higher class separability, but that minor threshold calibration was required. In general, the results show that maintaining a wider number of informative features leads to better ranking results while not losing threshold-optimized classification accuracy.

Table 1. Holdout performance across rfe subset sizes

RFE Subset	Feature Count	Holdout AUC	Accuracy (Default)	Accuracy (Youden)
RFE-100	100	0.871	0.824	0.882
RFE-300	300	0.907	0.823	0.853
RFE-500	500	**0.942**	0.794	**0.882**

Feature Selection Stability under Nested Cross-Validation

We used nested 5-fold cross-validation to segregate classifier optimisation and performance evaluation in order to provide an unbiased estimate of model generalisation. Feature selection and hyperparameter optimisation were only done in the inner folds, whereas the outer folds were used for assessing the optimised models on unseen data. Table 2 summarises the performance attained throughout the five outer folds.

Table 2. Nested cross-validation results (outer 5-folds)

Fold	Selected Features	Accuracy	AUC
1	500	1.000	1.000
2	500	0.824	0.956
3	500	0.970	0.995
4	500	0.909	0.972
5	100	0.818	0.931
Mean ± SD	—	0.916 ± 0.071	0.959 ± 0.045

The suggested framework achieved a mean AUC of 0.959±0.045, demonstrating consistently good discriminative performance across the outer validation folds. Importantly, the 500-feature subset was picked in four out of five folds, demonstrating that maintaining a moderately rich feature representation yielded the most resilient performance during independent optimisation. The little standard deviation also demonstrates the performance is stable even with the small sample size. Furthermore, the near match between the nested cross-validation AUC (0.959) and the independent holdout AUC (0.942) implies that the suggested assessment technique adequately controlled optimism bias, supporting the dependability of the stated generalisation performance.

Independent Holdout Evaluation

We tested the final RFE-500 + SVM pipeline on the independent 20% holdout set, which was totally omitted from the nested cross-validation process to ensure an unbiased estimation of generalisation. The appropriate ROC curve is shown in Figure 3.

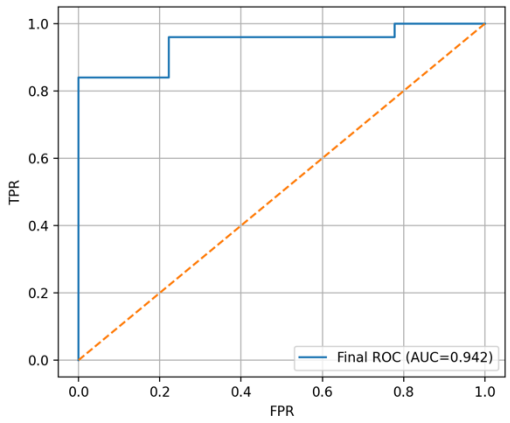


Figure 3. ROC curve of final RFE-500 + SVM model

The model achieved a holdout AUC of 0.942, see Figure. 3, which demonstrates good discriminative power on unseen thermograms. This level of performance suggests that the diabetic and non-diabetic groups are reliably distinguished over a wide variety of decision criteria. Importantly, the holdout AUC is comparable with the nested cross-validation estimate (mean AUC = 0.959) suggesting that the model performance is not caused by selection bias or overfitting during feature tuning.

The total accuracy of the model was 0.794 with the default probability threshold (0.5). However, as the screening-oriented applications are based on balanced detection and not on arbitrary cut-offs, the threshold adjustment was accomplished based on the Youden index. At the optimised operating point, sensitivity was 0.88 and specificity was 0.889, giving a total accuracy of 0.882 and an F1-score of about 0.917. The near symmetric sensitivity-specificity values further confirm the classifier's applicability under screening support scenarios by showing that the classifier's error behaviour is balanced in spite of the underlying class imbalance.

To further explore statistical stability, bootstrap resampling (n=1000) was performed on the holdout predictions.

Table 3. Bootstrap stability analysis

Metric	Value
Bootstrap Mean AUC	0.965
95% Confidence Interval	(0.925 – 0.993)

The bootstrap mean AUC (Table 3) is a little better than the single holdout estimate, but the 95% confidence range is narrow and below 0.92. This means that repeated resampling does not influence the discriminative performance of the model, and the good performance is not due to a favourable split of test data. The consistency between nested cross-validation, holdout evaluation and bootstrap estimation gives evidences

for the stability and reliability of the recommended RFE-500 framework in thermographic screening condition with limited samples.

Calibration Analysis

AUC and other discrimination measures assess the model's ability to differentiate between classes, not its ability to forecast probabilities that correspond to the true likelihood of the result. Model calibration was thus evaluated by comparing the anticipated probability to the observed result frequencies. The calibration curve of the final model is shown in Figure 4.

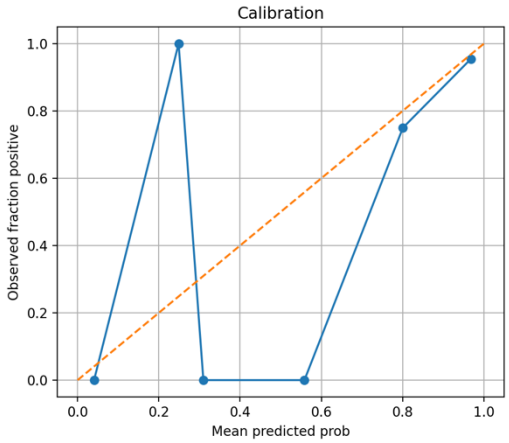


Figure 4. Calibration curve of final model

Figure 4 shows that the calibration curve is nearly superimposed on the diagonal reference line, demonstrating a strong agreement between the predicted probability and the observed event rates. The model had a Brier score of 0.065, translating to a low mean squared error between the predicted probability and the related class labels. This result shows good probability calibration with good discriminative performance . The little differences seen for the middle probability ranges are probably due to the small number of holdout samples rather than systemic calibration error. These results, along with strong discrimination performance (AUC=0.942), indicate that the proposed model gives reliable probability estimates and support the possible use of the model in screening applications where risk-based decision making is needed.

Feature Importance and Descriptor Contribution

We conducted a permutation significance analysis of the final RFE-500 + SVM model to assess the relative contribution of selected variables on classification performance. At the same time, descriptor level contribution was mapped by retaining features using RFE.

The RFE approach retained a large majority of HOG characteristics as seen in Table 4. This suggests that the gradient-based representations contribute significantly to the final decision boundary. The second largest collection of MBLBP features is the complementary discriminative information of block-level texture patterns.

The permutation importance analysis of the top ranking features further confirms this conclusion, as HOG bins often appear among the most important variables, while MBLBP features have secondary but consistent value. In contrast, LBP, K-means centroids and Haralick statistics rarely appear in the final selection, indicating their relatively lesser marginal contribution in the optimised feature subset.

Table 4. Descriptor composition of selected features (RFE-500)

Descriptor	Selected Count	Percentage
HOG	423	84.6%
LBP	1	0.2%
MBLBP	69	13.8%
K-Means	3	0.6%
Haralick	4	0.8%
Total	500	100%

Collectively, these results show that the structural gradient information recorded by HOG contributes dominantly to thermogram discrimination, while the texture-based descriptors fine-tune the categorisation border.

Baseline Comparison

To offer a relevant performance baseline, the proposed framework was compared to traditional classifiers trained on the same preprocessing pipeline, feature scaling approach and train--test split. This comparison separates the contribution of structured feature selection from the choice of classifier. Table 5 summarises the results.

Table 5. Baseline model comparison

Model	Feature Count	Holdout AUC	Accuracy (Youden)
Random baseline	-	0.500	-
KNN (k=5)	2150	0.873	0.824
SVM (full features)	2150	0.914	0.853
RFE-500 + SVM	500	0.942	0.882

As shown in Table 5, the suggested RFE-500 + SVM framework produced the best discriminative performance with an AUC of 0.942 and a Youden-optimized accuracy of 0.882. The application of RFE boosted AUC by 0.028 compared with the full-feature SVM and reduced the feature dimensionality from 2,150 to 500. These results show that removing redundant or weakly informative features can increase the classification performance in HDLSS scenarios, not only reduce the computational cost.

The KNN baseline had the worst performance among the trained classifiers, which indicates that distance-based classification is more susceptible to high-dimensional handcrafted feature representations. On the other hand, the better performance of the RFE-optimized SVM shows that the combination of structured feature selection and a margin-based classifier is a more effective method for thermographic diabetic foot classification under limited-data situations.

Error Pattern Analysis

Classification errors were evaluated on the independent holdout set using the Youden-optimized decision threshold. The proposed model exhibited a sensitivity of 0.880 and a specificity of 0.889, indicating a well-balanced compromise between accurately recognizing diabetes patients and correctly rejecting non-diabetic instances. Despite the intrinsic class imbalance in the dataset, the model was able to maintain similar performance on both classes.

Similarly, the comparable sensitivity and specificity values imply that the proposed framework provides a balanced discrimination without bias for any class. The sensitivity obtained in the clinical setting shows that most of the thermograms of the diabetics were effectively identified yet the resulting specificity reveals a tolerable false-positive rate for screening-oriented applications.

To sum up, the results indicate that the framework of RFE-500 + SVM provides a balanced classification performance and trustworthy discrimination, therefore encouraging its future implementation as a decision-support tool for thermographic diabetic foot screening. However, the proposed paradigm is designed to augment, not replace, clinical assessment and diagnostic decision-making.

Discussions

The experimental results are discussed in relation to the current literature on thermographic diabetic foot classification, with special emphasis on feature representation, validation rigour, model dependability and clinical applicability. Instead of the prediction performance, the discussion will be focused on how the suggested approach addresses the problems of learning from high-dimensional low sample size (HDLSS) biological data.

Experimental results show that manual feature engineering, structured feature selection and rigorous validation consistently leads to strong classification performance with a holdout AUC of 0.942, a nested cross validation AUC of 0.959+/-0.045 and good probability calibration (Brier score = 0.065). The findings suggest that with proper design for feature representation and model evaluation, accurate categorisation of thermograms can be obtained without the need for very complicated deep learning architectures.

More importantly, our analysis reveals that in limited-data medical imaging, the reliability of the model depends not only on the choice of the classifier, but also on the quality of the feature representation and the strength of the validation technique. This view moves the focus from the search for more and more complicated architectures to the development of robust, interpretable and reproducible machine learning frameworks for clinical decision assistance.

Representation Richness and Dimensional Stability

The dimensionality-performance investigation reveals that the thermographic discrimination is not activated by a minimal sparse collection of descriptors but by distributed structural-textural interactions. Increasing the number of preserved features from 100 to 500 features steadily improved the AUC, demonstrating that important pathogenic signals are geographically large and multi-scale.

The superiority of RFE-500 over RFE-100 suggests that diabetic plantar thermograms encode asymmetries related to coherent gradient fields of inflammation, not isolated pixel-level defects. Over-compression leads to loss of margin stability if the complementary descriptors are discarded, while keeping certain dimensions preserves the discriminative geometry without the variance inflation of the entire 2150 dimensional representation.

This finding accords with the bias-variance view: optimal feature reduction is not maximum compression but organised dimensionality optimisation preserving richness of informative representations.

Validation Rigor as a Determinant of Credibility

A major strength of this work is the triangulation of evaluation methodologies . The nested cross-validation, the independent holdout test and the bootstrap confidence interval estimate all converged to the same performance levels. The strong agreement between nested AUC (0.959) and holdout AUC (0.942) suggests very limited optimism bias, a common problem in small biomedical datasets when feature selection and evaluation are not performed in separate steps [14].

In the current architecture, optimisation and performance estimation are strictly segregated, since RFE and hyperparameter tuning are included only in the inner folds. The stability over the validation regimes suggests that the discrimination skill shown is the result of actual generalisation behaviour rather than favourable sampling effects.

The contribution is therefore as much methodological as it is empirical: rigorous assessment design has a large impact on the trustworthiness of reported outcomes in biological AI research.

Structural Gradient Dominance in Thermographic Discrimination

The feature importance analysis showed that the optimised representation was dominated by HOG descriptors, which contributed 84.6% of the selected features. MBLBP features offered supplementary texture information. The distribution shows that spatial temperature gradients are the main feature of diabetic foot thermograms, while local texture patterns are a supporting feature and not the main source of discriminative information.

This finding is congruent with the pathophysiology of diabetic foot, where inflammation and microvascular dysfunction are usually represented as regional thermal asymmetries, rather than individual pixel-level anomalies [3]. Gradient-based descriptors appear therefore well equipped to capture such therapeutically relevant thermal trends.

Notably, RFE preserved features from various descriptor families, rather than selecting a single predominant representation. This result implies that gradient and texture descriptors include complimentary information at different spatial scales, emphasising the importance of combining numerous handmade representations for the categorisation of thermographic diabetic foot.

Margin-Based Optimization versus Distance-Based Classification

The baseline comparison also provides further insight into the link of feature representation and classifier design. The proposed RFE-500 + SVM framework shows consistent improvement over the full-feature SVM and KNN baseline, demonstrating that the performance gain is not just attributable to the choice of classifier but also due to the fusion of structured feature selection and an appropriate learning strategy.

The RFE-optimized model performed better than the full-feature SVM despite using less than one-quarter of the original feature set. This finding implies that removing duplicated or weakly informative features improves the efficacy of the trained representation, especially in the high-dimensional low-sample-size setting.

The inferior performance of KNN also further indicates distance-based classifiers may not be as useful in high dimensional handcrafted feature representations as pairwise distances lose their relevance. In contrast, the higher performance of SVM implies that margin-based learning is more adapted to use the complementary gradient and texture information retained after feature selection [31]. Our results, in general, demonstrate that for successful biomedical image classification, it is important to align feature representation with classifier design rather than any individual component in isolation.

Calibration and Screening Suitability

Besides great discriminative performance, the proposed framework also showed good probability calibration as indicated in a low Brier score (0.065) and a calibration curve close to the diagonal reference line.

Overall, the model's risk estimates were reliable, as predicted probabilities were consistent with the observed frequencies of outcome occurrences.

The Youden-optimal threshold yielded identical values for the model's sensitivity (0.880) and specificity (0.889), indicating a balanced detection capacity despite the intrinsic class imbalance of the dataset. From a clinical standpoint, it is particularly important to find a balance between false-negative and false-positive predictions in screening applications where classification outputs may influence further referral or diagnostic assessment.

Both discrimination and calibration results show that the proposed framework has potential as a decision support tool for thermographic diabetic foot screening. It should be noted, however, that its objective is to assist in the first risk stratification and not to replace a comprehensive clinical assessment or final diagnosis.

Positioning within the Deep Learning Landscape

Deep convolutional neural networks (CNNs) have been the state-of-the-art solution for thermographic diabetic foot classification due to their ability to learn hierarchical feature representations directly from the imaging data [6]. However, their performance mostly depends on large annotated datasets and extensive computing resources, which are still not easily available in many biological applications [12].

Our results indicate that carefully designed handcrafted feature representations can still be a powerful option in limited-data settings. Complementary gradient and texture descriptors, organised feature selection and rigorous validation combined to yield a proposed framework that produced good discrimination while preserving interpretability and computing economy. These properties make handmade techniques especially desirable for resource limited clinical settings where transparency, reproducibility and flexibility of deployment might be as crucial as predicting performance.

In this study we argue that the representation should be selected according to the amount of data and the application, instead of considering handmade approaches as an alternative to deep learning. In small-sample biomedical imaging, methodological rigour and proper feature representation may be as important for model reliability as architectural complexity.

Comparative Performance with Prior Studies

To put the present findings into context, Table 6 compares the proposed framework with representative machine learning and deep learning studies on thermographic diabetic foot categorisation.

Table 6. Comparative experimental performance across thermographic studies

Study	Dataset Size (N)	Method	Validation	Accuracy	Sensitivity	Specificity	AUC
Panamonta et al. [8]	153	Deep Learning	5-fold CV	71.8%	81.2%	64.0%	–
Khandakar et al. [6]	122	ML-based	Train/Test	~93–95%	–	–	–
Proposed Study	167	RFE-500 + SVM	Nested CV + Holdout	88.2%	0.88	0.889	0.942

Table 6 shows that the holdout AUC of the proposed RFE-500 + SVM architecture is 0.942, indicating that it has significant discriminative performance in the range reported in contemporary publications. Some of the earlier papers, e.g., Panamonta et al. [8] and Khandakar et al. [6] show competitive performance, but the approaches used to evaluate them are considerably different in validation design and performance reporting. In particular, some studies apply single train-test splits or standard cross-validation without explicitly separating model selection and performance estimations, which can introduce optimistic bias.

The suggested framework emphasises evaluation rigour more than earlier research using nested cross-validation and independent holdout testing. Several previous research have found similar performance, but changes in validation procedures make direct comparisons problematic. The key contribution of this work is therefore not only the competitive discrimination performance but also a more reliable assessment of model generalisation for limited-data thermographic classification.

On the other hand, the present study employs a carefully constructed cross-validation system with embedded feature selection and independent holdout testing, in addition to bootstrap-based stability analysis. This validation method guarantees a good separation between model optimisation and model evaluation, thereby providing a more reliable estimation of the generalisation performance in small-sample biomedical settings.

Some deep learning algorithms exhibit great classification accuracy, but often require large datasets, great data augmentation or architecture-specific customisation to perform well. For limited data, e.g., the thermogram dataset used in this work (n=167 samples), high-capacity models may become unstable or overfit if they are not carefully checked. The results obtained in this work show that a structured handcrafted-feature representation, along with disciplined validation processes, can provide competitive discrimination performance with interpretability and computational efficiency.

In conclusion, the comparative study shows that differences in performance between studies are not only due to model complexity, but are strongly influenced by representation design and rigour of evaluation. Moreover, the consistency of findings obtained in the layered cross-validation, holdout testing and bootstrap estimates further support the robustness of the proposed system. Our results demonstrate that well-engineered feature representations with bias-controlled validation can be a trustworthy and dependable alternative to high-capacity deep learning models for screening-type tasks in small-sample medical imaging settings.

4. CONCLUSION

This paper examines if a rigorously verified handcrafted-feature framework can provide reliable diabetic foot thermogram classification in low-data biomedical contexts. The suggested pipeline consists of multi-descriptor feature extraction, RFE, SVM and layered cross-validation and bootstrap stability analysis. The experimental results using a publicly accessible thermogram dataset (N = 167) showed a good and stable performance with a holdout AUC of 0.942 and a nested cross-validation AUC of 0.959 ± 0.045, combined with a balanced sensitivity and specificity. The results demonstrate that well designed handcrafted representations supplemented by strong validation techniques can lead to reliable discrimination performance on high-dimensional, low-sample-size thermographic data.

The study also found that the discriminative information is distributed among complimentary gradient and texture descriptors and not in a narrow group of features. Furthermore, we observed that the greatest stable classification performance was achieved with moderate feature dimensionality, underscoring the important role of feature representation, dimensionality control, and rigorous validation in the development of robust machine learning models for biomedical screening applications.

Despite these promising outcomes, there are several limits that should be addressed. First, the study was conducted using a single public available thermogram dataset of 167 photos, which would limit the generalisability of the results to diverse populations and imaging situations. Second, although the suggested framework displayed competitive performance, a direct comparison with deep learning models under the same evaluation protocol was outside the scope of this study. Finally, the framework has not yet been tested in prospective clinical settings and its practical utility should therefore be regarded with due caution.

Future study should include testing the suggested framework on larger multi-center datasets, longitudinal temperature monitoring data and prospective clinical validation processes. Furthermore, hybrid methods that combine handcrafted descriptors with lightweight deep feature representations might further increase classification performance while keeping interpretability and deployment practicalities.

ACKNOWLEDGEMENTS

We are grateful for the financial support provided by Telkom University Research Grant, which enabled us to conduct this research project. The author would like to express gratitude to PPM Telkom University. This research was funded by Telkom University's Internal Basic Research, Period 2, 2024

CREDIT AUTHORSHIP CONTRIBUTION STATEMENT

Bedy Purnama: Conceptualization, Methodology, Software, Project administration. **Bayu Erfianto**: Software, Writting – original draft. **Linlin Lindayani**: Writing – review & editing, Validation.

DECLARATION OF COMPETING INTERESTS

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.

DATA AVAILABILITY

Data will be made available on request.

REFERENCES

[1] D. G. Armstrong, A. J. M. Boulton, and S. A. Bus, "Diabetic Foot Ulcers and Their Recurrence.," *N. Engl. J. Med.*, vol. 376, no. 24, pp. 2367–2375, Jun. 2017, doi: 10.1056/NEJMra1615439.

[2] B. B. Duncan, C. Stein, and A. Basit, "Edinburgh Research Explorer IDF Diabetes Atlas," *Glob. Reg. country-level diabetes Preval. Estim. 2021 Proj. 2045*, 2021.

[3] L. A. Lavery *et al.*, "Preventing diabetic foot ulcer recurrence in high-risk patients: use of temperature monitoring as a self-assessment tool.," *Diabetes Care*, vol. 30, no. 1, pp. 14–20, Jan. 2007, doi: 10.2337/dc06-1600.

[4] S. A. Bus *et al.*, "Guidelines on the prevention of foot ulcers in persons with diabetes (IWGDF 2023 update)," *Diabetes. Metab.*

Res. Rev., vol. 40, no. 3, Mar. 2024, doi: 10.1002/dmrr.3651.
[5] J. A. Thompson and R. Pop-Busui, "Advances in diabetic neuropathy," *Lancet Diabetes Endocrinol*, vol. 11, no. 5, 2023.
[6] A. Khandakar *et al.*, "A Novel Machine Learning Approach for Severity Classification of Diabetic Foot Complications Using Thermogram Images," *Sensors*, vol. 22, no. 11, p. 4249, Jun. 2022, doi: 10.3390/s22114249.
[7] N. Papanas and E. Maltezos, "The diabetic foot: a global threat and a huge challenge for Greece.," *Hippokratia*, vol. 13, no. 4, pp. 199–204, Oct. 2009.
[8] V. Panamonta *et al.*, "Plantar Thermogram Analysis Using Deep Learning for Diabetic Foot Risk Classification.," *J. Diabetes Sci. Technol.*, vol. 20, no. 3, pp. 810–814, May 2026, doi: 10.1177/19322968251316563.
[9] M. H. Alshayeji, S. ChandraBhasi Sindhu, and S. Abed, "Early detection of diabetic foot ulcers from thermal images using the bag of features technique," *Biomed. Signal Process. Control*, vol. 79, p. 104143, Jan. 2023, doi: 10.1016/j.bspc.2022.104143.
[10] M. Tan and Q. Le, "Efficient Net: Rethinking Model Scaling for Convolutional Neural Networks," in *Proceedings of the 36th International Conference on Machine Learning*, K. Chaudhuri and R. Salakhutdinov, Eds., in Proceedings of Machine Learning Research, vol. 97. PMLR, 2019, pp. 6105–6114. [Online]. Available: https://proceedings.mlr.press/v97/tan19a.html
[11] M. M. Rahman, "Ensemble deep learning for thermogram classification," *Sensors*, vol. 22, no. 4, 2022.
[12] J. Heaton, "Ian Goodfellow, Yoshua Bengio, and Aaron Courville: Deep learning: The MIT Press, 2016, 800 pp, ISBN: 0262035618," *Genet. Program. Evolvable Mach.*, vol. 19, Oct. 2017, doi: 10.1007/s10710-017-9314-z.
[13] I. H. Sarker, "Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions," Nov. 2021, *Springer*. doi: 10.1007/s42979-021-00815-1.
[14] S. Varma and R. Simon, "Bias in error estimation when using cross-validation for model selection," *BMC Bioinformatics*, vol. 7, no. 1, p. 91, 2006, doi: 10.1186/1471-2105-7-91.
[15] S. Raschka, "Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning," Nov. 2020, doi: https://doi.org/10.48550/arXiv.1811.12808.
[16] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, p. 6, 2020, doi: 10.1186/s12864-019-6413-7.
[17] M. Roberts *et al.*, "Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans," *Nat. Mach. Intell.*, vol. 3, no. 3, pp. 199–217, 2021, doi: 10.1038/s42256-021-00307-0.
[18] C. J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King, "Key challenges for delivering clinical impact with artificial intelligence," *BMC Med.*, vol. 17, no. 1, p. 195, Dec. 2019, doi: 10.1186/s12916-019-1426-2.
[19] H. Ibrahim *et al.*, "Reporting guidelines for clinical trials of artificial intelligence interventions: the SPIRIT-AI and CONSORT-AI guidelines," *Trials*, vol. 22, no. 1, p. 11, 2021, doi: 10.1186/s13063-020-04951-6.
[20] D. A. Hernandez-Contreras, H. Peregrina-Barreto, J. de J. Rangel-Magdaleno, and F. J. Renero-Carrillo, "Plantar Thermogram Database for the Study of Diabetic Foot Complications," *IEEE Access*, vol. 7, pp. 161296–161307, 2019, doi: 10.1109/ACCESS.2019.2951356.
[21] H. Nguyen, "Recursive feature elimination for biomedical data," *Comput. Biol. Med.*, vol. 132, 2021.
[22] A. Vabalas, "Machine learning in neuroimaging," *Neuroimage*, vol. 2019, 197AD.
[23] M. Kuhn and K. Johnson, "Feature Engineering and Selection for Predictive Models," in *CRC Press*, 2021.
[24] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*. in Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press, 2014. [Online]. Available: https://books.google.co.id/books?id=ttJkAwAAQBAJ
[25] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, IEEE, pp. 886–893. doi: 10.1109/CVPR.2005.177.
[26] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002, doi: 10.1109/TPAMI.2002.1017623.
[27] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural Features for Image Classification," *IEEE Trans. Syst. Man. Cybern.*, vol. SMC-3, no. 6, pp. 610–621, Nov. 1973, doi: 10.1109/TSMC.1973.4309314.
[28] Z. Chen, F. Xiao, F. Guo, and J. Yan, "Interpretable machine learning for building energy management: A state-of-the-art review," *Adv. Appl. Energy*, vol. 9, p. 100123, Feb. 2023, doi: 10.1016/j.adapen.2023.100123.
[29] X. Li, "Handcrafted and hybrid features for thermographic classification," *Pattern Recognit. Lett*, vol. 156, 2022.