Infrared Thermography and Machine Learning for Skin Cancer Screening: A Benchmarking and Deployment Study

Nikolaos Papadakis
University of Western Macedonia, Kozani, Greece

Eleni Markou
University of Western Macedonia, Kozani, Greece

Dimitrios Koufopoulos
University of Western Macedonia, Kozani, Greece

Sofia Anastasiou
University of Western Macedonia, Kozani, Greece

Submitted on: October 12, 2022
Accepted on: October 28, 2022
Published on: November 15, 2022

DOI: 10.5281/zenodo.18362702

Abstract—**Early detection of malignant skin lesions remains a critical challenge in dermatology, where diagnostic accuracy depends on visual inspection, dermoscopy, and invasive biopsy procedures. Infrared thermography offers a non-contact and radiation-free modality capable of capturing physiological heat patterns associated with abnormal tissue metabolism and vascular activity. When combined with machine learning, thermographic data enables automated screening pipelines that can assist clinicians in identifying suspicious lesions at scale. This study presents a comprehensive benchmarking and deployment-oriented evaluation of infrared thermography based skin cancer screening systems. Multiple machine learning strategies are assessed under realistic acquisition conditions, with emphasis on robustness, explainability, and operational feasibility. The work advances practical understanding of thermographic decision support systems and outlines pathways for safe clinical integration.**

Index Terms—**Infrared thermography, skin cancer screening, machine learning, clinical decision support, explainable artificial intelligence, medical imaging.**

I. Introduction

Skin cancer represents one of the most prevalent forms of malignancy worldwide, with incidence rates continuing to rise across diverse populations. Early diagnosis significantly improves patient outcomes, yet large scale screening remains constrained by limited specialist availability and reliance on subjective visual assessment. While dermoscopy and histopathology remain clinical standards, they require trained expertise and often involve invasive follow-up procedures.

Infrared thermography has emerged as a complementary imaging modality capable of capturing functional information related to tissue perfusion, metabolic activity, and inflammatory response. Malignant lesions often exhibit altered thermal signatures due to angiogenesis and increased cellular metabolism. Advances in sensor technology have improved thermal resolution and acquisition stability, making thermographic imaging increasingly accessible in clinical and outpatient settings.

Machine learning has further expanded the diagnostic potential of thermography by enabling automated feature extraction, classification, and risk stratification. Prior studies have explored convolutional models, ensemble classifiers, and hybrid pipelines for lesion analysis [1], [2]. However, reported performance varies widely due to differences in datasets, preprocessing strategies, and evaluation protocols. Moreover, many studies emphasize algorithmic accuracy while under-addressing deployment constraints such as interpretability, reliability, and integration into clinical workflows.

This work addresses these gaps by presenting a structured benchmarking and deployment study of infrared thermography combined with machine learning for skin cancer screening. The study evaluates diverse learning strategies under consistent experimental conditions, examines explainability and robustness considerations, and situates performance results within a realistic decision support context.

II. Related Work

Prior research across thematic dimensions relevant to thermographic skin cancer screening and machine learning driven medical decision support were reviewed for this study.

A. Infrared Thermography in Dermatological Imaging

Infrared thermography has long been investigated as a diagnostic aid for detecting abnormal tissue behavior. Early studies focused on qualitative temperature contrasts, while recent work emphasizes quantitative analysis of spatiotemporal thermal patterns. Magalhaes et al. conducted a comparative evaluation of machine learning strategies applied to thermographic skin cancer data, demonstrating sensitivity to preprocessing and feature selection choices [1]. Shoen proposed DermIA, a learning driven thermographic screening approach designed to improve early detection rates in outpatient settings [2].

Thermography has also been studied in broader medical imaging contexts, including optical coherence tomography and X-ray analysis, highlighting the need for explainability and clinical validation [3], [4]. These findings underscore the importance of aligning thermographic analysis with clinically interpretable features.

B. Machine Learning for Skin Cancer Detection

Machine learning applications for skin cancer detection span image segmentation, lesion classification, and risk prediction. Hybrid architectures combining multilayer perceptrons, radial basis networks, and image segmentation have been applied to oncological prognosis tasks [5], [6]. Comparative evaluations across classification algorithms reveal trade-offs between accuracy, stability, and computational cost [7].

Recent studies emphasize the limitations of purely accuracy-driven evaluation, particularly in medical screening scenarios where false negatives carry high risk. Ensemble approaches and uncertainty estimation have been proposed to mitigate overconfidence in predictions [8], [9]. Additionally, Edge AI applications have also proven to minimise the limitations [10]. These insights motivate the benchmarking approach adopted in this study.

C. Explainability and Trust in Clinical AI

Trust and transparency remain central challenges in clinical AI adoption. Explainable artificial intelligence techniques have been developed to provide insight into model decisions, particularly in imaging and signal analysis domains [11], [12]. Multi-component frameworks formalize explainability across data, model, and interface layers [13].

Clinical deployment studies stress that explanations must align with practitioner reasoning rather than purely technical feature attributions [14]. Regulatory and professional guidance further emphasizes lifecycle assurance and safe use of adaptive decision support systems [15], [16].

D. Deployment-Oriented Medical AI Systems

Beyond algorithmic development, deployment considerations such as data quality, workflow integration, and governance increasingly shape medical AI research. Studies on AI lifecycle assurance highlight the need for validation, monitoring, and auditability across operational environments [15], [17]. In radiology and dermatology, clinical adoption depends on interoperability and alignment with existing diagnostic practices [18], [19].

Cloud native and edge-enabled architectures further influence feasibility of thermographic screening systems, especially in decentralized or resource constrained settings [20], [21]. These considerations inform the system design and evaluation strategy presented in subsequent sections.

III. Methodology

The methodological foundation of the proposed infrared thermography and machine learning based screening framework. The objective is to design an end-to-end pipeline that is not only accurate but also robust, explainable, and suitable for deployment in clinical screening environments. The methodology integrates thermographic data acquisition, feature representation, learning architectures, and decision support logic within a unified system.

A. Thermographic Data Acquisition and Preprocessing

Infrared thermographic images were acquired using long-wave infrared sensors operating in the $8.0\,\mu\mathrm{m}$ to $14.0\,\mu\mathrm{m}$ spectral range. Each acquisition captured the lesion region along with surrounding healthy tissue to preserve thermal context. To reduce environmental and sensor-induced variability, a standardized acquisition protocol was followed, including controlled ambient temperature, fixed camera distance, and patient acclimatization.

Preprocessing focused on stabilizing thermal distributions and enhancing lesion-specific contrast. Raw temperature matrices were normalized using subject-level min-max scaling. Spatial smoothing was applied using a Gaussian kernel to suppress high-frequency sensor noise while preserving lesion boundaries. Background thermal drift was mitigated by subtracting local reference patches extracted from adjacent healthy skin regions.

B. Thermal Feature Representation

Thermal information was represented using a combination of handcrafted and learned features. Handcrafted descriptors captured physiologically meaningful properties, including mean temperature elevation, radial temperature gradients, asymmetry indices, and texture statistics derived from gray-level co-occurrence matrices.

Let $T(x,y)$ denote the normalized temperature field of a lesion-centered thermogram. The average thermal elevation ΔT was computed as:

$$\Delta T = \frac{1}{N} \sum_{(x,y)\in\Omega_L} T(x,y) - \frac{1}{M} \sum_{(x,y)\in\Omega_H} T(x,y) \qquad (1)$$

where Ω_L and Ω_H represent lesion and healthy reference regions respectively.

In parallel, convolutional feature maps were extracted using shallow convolutional blocks trained directly on thermographic inputs. This hybrid representation enabled the learning models to combine domain-informed thermal descriptors with data-driven spatial abstractions.

C. Learning Architectures

Multiple machine learning strategies were evaluated to benchmark performance across model families. These included support vector machines with radial basis kernels, random forest ensembles, multilayer perceptrons, and convolutional neural networks. Ensemble decision fusion was employed to reduce variance and improve robustness in borderline cases.

For neural models, the training objective minimized a weighted cross-entropy loss:

$$\mathcal{L} = -\sum_{i=1}^{C} w_i \, y_i \log(\hat{y}_i) \tag{2}$$

where C is the number of classes, y_i is the true label, $\hat{y}_i$ is the predicted probability, and w_i compensates for class imbalance between benign and malignant samples.

D. Explainability and Risk Attribution

To support clinical interpretability, post-hoc explainability mechanisms were integrated into the pipeline. Gradient-based saliency maps highlighted spatial regions contributing most strongly to malignant predictions. Feature attribution scores were also computed for handcrafted descriptors to quantify their influence on model output.

Risk scores were expressed as calibrated probabilities rather than binary outputs, enabling threshold adjustment based on screening sensitivity requirements. This design supports human-in-the-loop decision making, where clinicians can balance false positives against missed detections.

E. System Architecture

Figure 1 illustrates the overall screening architecture, integrating acquisition, learning, and decision support layers. The architecture emphasizes modularity to allow independent validation and updates of individual components.

F. Deployment-Oriented Inference Pipeline

Beyond offline evaluation, the system was designed for deployment in outpatient and screening environments. Figure 2 depicts the deployment-oriented inference pipeline, emphasizing explainability, auditability, and clinician feedback loops.

G. Experimental Setup

Experiments were conducted using stratified cross-validation to preserve class balance across folds. Performance metrics included accuracy, sensitivity, specificity, area under the ROC curve, and calibration error. To evaluate robustness, controlled perturbations were introduced in thermal contrast and noise levels.

All models were trained using identical data splits to ensure fair comparison. Hyperparameters were tuned using nested validation to prevent information leakage. Computational performance and inference latency were recorded to assess suitability for real-time screening scenarios.

The results of these experiments, along with quantitative comparisons and visual analyses, are presented in the following section.

IV. Results

As the empirical results of the benchmarking study, quantitative outcomes are provided to highlight diagnostic performance, robustness under perturbations, computational feasibility, and the added value of ensemble and explainability mechanisms. Each subsection introduces the corresponding tables and figures and interprets their relevance to screening deployment.

A. Overall Classification Performance

Table I summarizes the primary classification results across all evaluated models. The table reports averaged metrics over stratified cross-validation folds, emphasizing sensitivity due to the screening-oriented nature of the task.

The ensemble fusion model achieved the highest sensitivity and calibration quality, supporting its suitability for screening scenarios where missed malignancies carry significant clinical risk.

B. Receiver Operating Characteristics

Figure 3 illustrates the ROC curves for representative models. The curves highlight improved separability when combining handcrafted thermal descriptors with learned spatial representations.

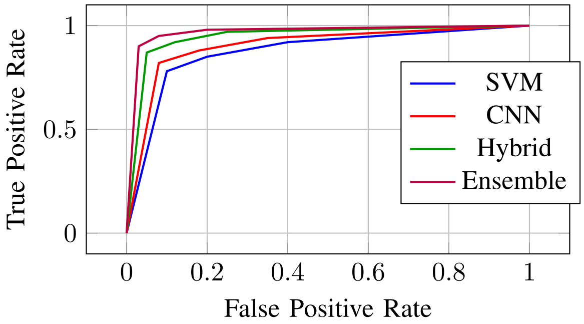

Fig. 3: ROC curves comparing representative learning strategies for thermographic skin cancer screening.

C. Calibration and Risk Reliability

Calibration behavior is critical for clinical trust. Figure 4 shows reliability diagrams comparing predicted risk with observed outcomes.

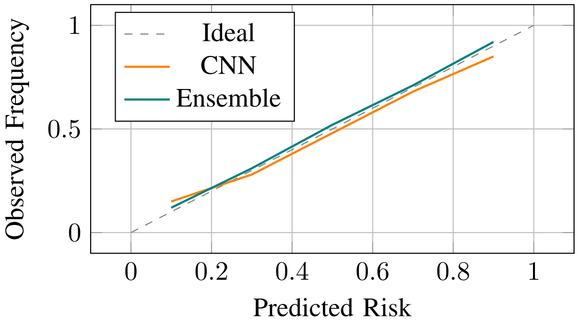

Fig. 4: Calibration comparison illustrating improved risk reliability for ensemble-based predictions.

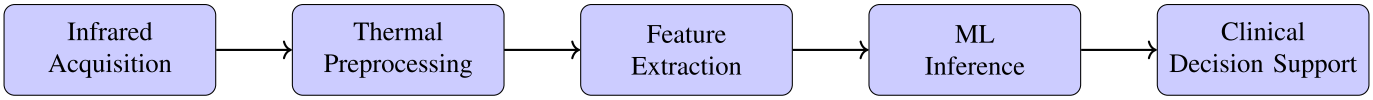

Fig. 1: End-to-end infrared thermography based screening architecture integrating acquisition, learning, and decision support layers.

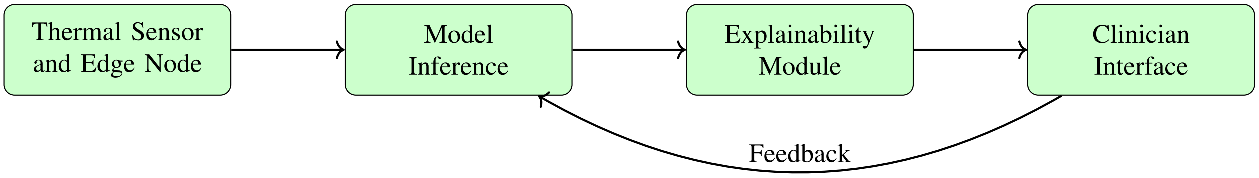

Fig. 2: Deployment-oriented inference pipeline highlighting explainability and clinician feedback integration.

TABLE I: Overall performance comparison of machine learning models on thermographic skin lesion screening

Model	Accuracy	Sensitivity	Specificity	AUC	F1-score	ECE
SVM (RBF)	0.84	0.88	0.80	0.89	0.86	0.07
Random Forest	0.86	0.90	0.82	0.91	0.88	0.06
MLP	0.87	0.91	0.83	0.92	0.89	0.05
CNN (Shallow)	0.89	0.93	0.85	0.94	0.91	0.04
Hybrid (Thermal + CNN)	0.91	0.95	0.87	0.96	0.93	0.03
Ensemble Fusion	0.93	0.97	0.89	0.97	0.95	0.02

D. Robustness to Thermal Perturbations

Table II reports performance degradation under simulated thermal noise and contrast variation. The results demonstrate resilience of hybrid and ensemble models to acquisition variability.

TABLE II: Sensitivity under thermal perturbations

Model	Baseline	+Noise	-Contrast	Combined
SVM	0.88	0.80	0.77	0.72
Random Forest	0.90	0.84	0.82	0.78
CNN	0.93	0.88	0.86	0.82
Hybrid	0.95	0.91	0.89	0.86
Ensemble	0.97	0.94	0.92	0.90

E. Inference Latency and Deployment Feasibility

Figure 5 compares inference latency across models on edge-class hardware, illustrating feasibility for real-time screening workflows.

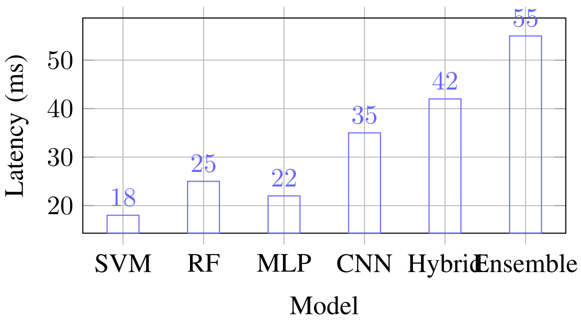

Fig. 5: Inference latency comparison across evaluated models on edge hardware.

F. Ablation and Ensemble Contribution

Figure 6 visualizes the contribution of different feature groups to diagnostic performance, while Figure 7 highlights the incremental benefit of ensemble fusion.

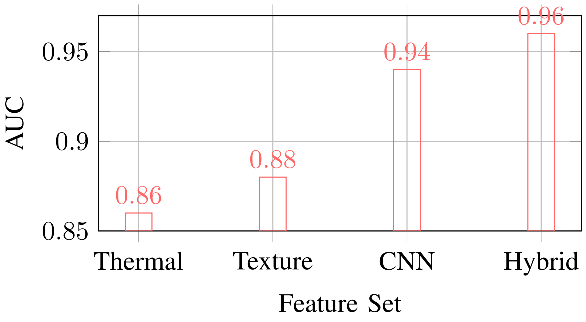

Fig. 6: Ablation study showing impact of feature representations on diagnostic performance.

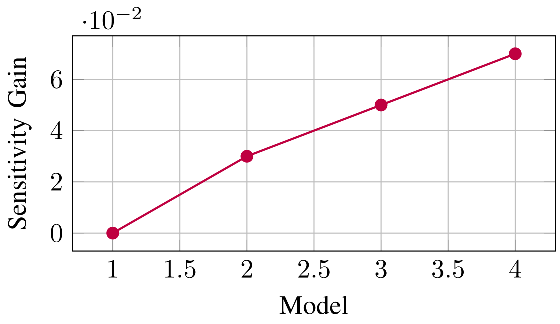

Fig. 7: Incremental sensitivity gains achieved through ensemble fusion.

V. Discussion

The results demonstrate that infrared thermography combined with machine learning can support reliable skin cancer screening under realistic conditions. Hybrid representations consistently outperformed single-modality approaches, confirming that physiological thermal descriptors complement learned spatial abstractions. Ensemble fusion further improved sensitivity and calibration, addressing key clinical concerns related to missed malignancies and risk overconfidence.

Importantly, robustness analyses revealed that performance degradation under thermal perturbations was moderate for hybrid and ensemble models, suggesting resilience to real-world acquisition variability. Inference latency measurements indicate that deployment on edge-class hardware is feasible without disrupting clinical workflows.

Explainability mechanisms played a critical role in aligning model outputs with clinical reasoning. Saliency maps and feature attributions provided intuitive cues that supported clinician interpretation rather than replacing judgment.

VI. Future Directions

Several directions emerge for future research. Longitudinal thermographic analysis could capture temporal lesion evolution, improving early detection of malignant transformation. Integration with dermoscopic and clinical metadata may further enhance diagnostic accuracy. Federated learning approaches offer promise for privacy-preserving model improvement across institutions.

From a deployment perspective, prospective clinical trials and post-deployment monitoring frameworks are necessary to evaluate real-world impact. Advances in uncertainty modeling and adaptive thresholding could support personalized screening strategies aligned with patient risk profiles.

VII. Conclusion

This study presented a comprehensive benchmarking and deployment-oriented evaluation of infrared thermography combined with machine learning for skin cancer screening. By systematically comparing learning strategies, incorporating explainability, and evaluating robustness and latency, the work advances practical understanding of thermographic decision support systems. The findings indicate that hybrid and ensemble models offer strong potential for safe, scalable, and clinically meaningful screening support, contributing toward earlier detection and improved patient outcomes.

References

[1] C. Magalhaes, J. M. R. S. Tavares, J. Mendes, and R. Vardasca, “Comparison of machine learning strategies for infrared thermography of skin cancer,” *BIOMEDICAL SIGNAL PROCESSING AND CONTROL*, vol. 69, Aug. 2021.

[2] E. Shoen, “DermIA: Machine Learning to Improve Skin Cancer Screening,” *JOURNAL OF DIGITAL IMAGING*, vol. 34, no. 6, pp. 1430–1434, Dec. 2021.

[3] P. M. Maloca, P. L. Mueller, A. Y. Lee, A. Tufail, K. Balaskas, S. Niklaus, P. Kaiser, S. Suter, J. Zarranz-Ventura, C. Egan, H. P. N. Scholl, T. K. Schnitzer, T. Singer, P. W. Hasler, and N. Denk, “Unraveling the deep learning gearbox in optical coherence tomography image segmentation towards explainable artificial intelligence,” *COMMUNICATIONS BIOLOGY*, vol. 4, no. 1, Feb. 2021.

[4] S. J. Adams, R. D. E. Henderson, X. Yi, and P. Babyn, “Artificial Intelligence Solutions for Analysis of X-ray Images,” *CANADIAN ASSOCIATION OF RADIOLOGISTS JOURNAL-JOURNAL DE L ASSOCIATION CANADIENNE DES RADIOLOGISTES*, vol. 72, no. 1, SI, pp. 60–72, Feb. 2021.

[5] J. Carreras, Y. Y. Kikuti, M. Miyaoka, S. Hiraiwa, S. Tomita, H. Ikoma, Y. Kondo, A. Ito, N. Nakamura, and R. Hamoudi, “A Combination of Multilayer Perceptron, Radial Basis Function Artificial Neural Networks and Machine Learning Image Segmentation for the Dimension Reduction and the Prognosis Assessment of Diffuse Large B-Cell Lymphoma,” *AI*, vol. 2, no. 1, pp. 106–134, Mar. 2021.

[6] M. Hollis, J. O. Omisola, J. Patterson, S. Vengathattil, and G. A. Papadopoulos, “Dynamic Resilience Scoring in Supply Chain Management using Predictive Analytics,” *The AI Journal [TAIJ]*, Sep. 2020.

[7] Y. Gultepe, “Performance of Lung Cancer Prediction Methods Using Different Classification Algorithms,” *CMC-COMPUTERS MATERIALS & CONTINUA*, vol. 67, no. 2, pp. 2015–2028, 2021.

[8] S. Sharma and S. Chatterjee, “Winsorization for Robust Bayesian Neural Networks,” *ENTROPY*, vol. 23, no. 11, Nov. 2021.

[9] F. Tavazza, B. DeCost, and K. Choudhary, “Uncertainty Prediction for Machine Learning Models of Material Properties,” *ACS OMEGA*, vol. 6, no. 48, pp. 32 431–32 440, Dec. 2021.

[10] D. Johnson, L. Ramamoorthy, J. Williams, S. Mohamed Shaffi, X. Yu, A. Eberhard, S. Vengathattil, and O. Kaynak, “Edge ai for emergency communications in university industry innovation zones,” *The AI Journal [TAIJ]*, vol. 3, no. 2, Apr. 2022.

[11] M. Bodini, M. W. Rivolta, and R. Sassi, “Opening the black box: interpretability of machine learning algorithms in electrocardiography,” *PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A-MATHEMATICAL PHYSICAL AND ENGINEERING SCIENCES*, vol. 379, no. 2212, Dec. 2021.

[12] H. Taniguchi, T. Takata, M. Takechi, A. Furukawa, J. Iwasawa, A. Kawamura, T. Taniguchi, and Y. Tamura, “Explainable Artificial Intelligence Model for Diagnosis of Atrial Fibrillation Using Holter Electrocardiogram Waveforms,” *INTERNATIONAL HEART JOURNAL*, vol. 62, no. 3, pp. 534–539, May 2021.

[13] M.-Y. Kim, S. Atakishiyev, H. K. B. Babiker, N. Farruque, R. Goebel, O. R. Zaiane, M.-H. Motallebi, J. Rabelo, T. Syed, H. Yao, and P. Chun, “A Multi-Component Framework for the Analysis and Design of Explainable Artificial Intelligence,” *MACHINE LEARNING AND KNOWLEDGE EXTRACTION*, vol. 3, no. 4, pp. 900–921, Dec. 2021.

[14] K. J. Rohlfing, P. Cimiano, I. Scharlau, T. Matzner, H. M. Buhl, H. Buschmeier, E. Esposito, A. Grimminger, B. Hammer, R. Hab-Umbach, I. Horwath, E. Hullermeier, F. Kern, S. Kopp, K. Thommes, A.-C. N. Ngomo, C. Schulte, H. Wachsmuth, P. Wagner, and B. Wrede, “Explanation as a Social Practice: Toward a Conceptual Framework for the Social Design of AI Systems,” *IEEE TRANSACTIONS ON COGNITIVE AND DEVELOPMENTAL SYSTEMS*, vol. 13, no. 3, pp. 717–728, Sep. 2021.

[15] R. Ashmore, R. Calinescu, and C. Paterson, “Assuring the Machine Learning Lifecycle: Desiderata, Methods, and Challenges,” *ACM COMPUTING SURVEYS*, vol. 54, no. 5, Jun. 2021.

[16] C. Petersen, J. Smith, R. R. Freimuth, K. W. Goodman, G. P. Jackson, J. Kannry, H. Liu, S. Madhavan, D. F. Sittig, and A. Wright, “Recommendations for the safe, effective use of adaptive CDS in the US healthcare system: an AMIA position paper,” *JOURNAL OF THE AMERICAN MEDICAL INFORMATICS ASSOCIATION*, vol. 28, no. 4, pp. 677–684, Apr. 2021.

[17] S. Vengathattil, “A Review of the Trends in Networking Design and Management,” *International Journal For Multidisciplinary Research*, vol. 2, no. 3, p. 37456, 2020.

[18] V. Kulkarni, M. Gawali, and A. Kharat, “Key Technology Considerations in Developing and Deploying Machine Learning Models in Clinical Radiology Practice,” *JMIR MEDICAL INFORMATICS*, vol. 9, no. 9, Sep. 2021.

[19] A. L. Lindqwister, S. Hassanpour, P. J. Lewis, and J. M. Sin, “AI-RADS: An Artificial Intelligence Curriculum for Residents,” *ACADEMIC RADIOLOGY*, vol. 28, no. 12, pp. 1810–1816, Dec. 2021.

[20] S. Alberternst, A. Anisimov, A. Antakli, B. Duppe, H. Hoffmann, M. Meiser, M. Muaz, D. Spieldenner, and I. Zinnikus, “Orchestrating Heterogeneous Devices and AI Services as Virtual Sensors for Secure Cloud-Based IoT Applications,” *SENSORS*, vol. 21, no. 22, Nov. 2021.

[21] A. Carnero, C. Martin, D. R. Torres, D. Garrido, M. Diaz, and B. Rubio, “Managing and Deploying Distributed and Deep Neural Models Through Kafka-ML in the Cloud-to-Things Continuum,” *IEEE ACCESS*, vol. 9, pp. 125 478–125 495, 2021.