EUROPEAN JOURNAL OF EMERGING ARTIFICIAL INTELLIGENCE

The Resilience of Deep Learning Models for Breast Cancer Detection: A Quantitative Analysis of Performance Under Diverse Noise Conditions in Thermal Imaging

Dr. Mira T. Elvanor
Department of Biomedical Engineering Lusaka Institute of Medical Imaging and Technology, Zambia

Prof. Jayson K. Brulett
Division of AI in Health Diagnostics Nordhavn School of Applied Sciences, Norway

V0LUME02 ISSUE01 (2025)

Published Date: 06 January 2025 // Page no.: - 01-10

ABSTRACT

Breast cancer continues to be a leading cause of mortality among women globally, making early and accurate detection paramount. Infrared thermography has emerged as a promising non-invasive, radiation-free diagnostic modality that identifies potential malignancies by capturing the subtle temperature variations associated with tumor metabolism and angiogenesis. The integration of deep learning, particularly Convolutional Neural Networks (CNNs), has significantly advanced the analytical power of thermography. However, the inherent susceptibility of thermal images to various types of electronic and environmental noise presents a critical challenge to the reliability of these automated systems. The performance of deep learning models under realistic, noisy conditions is not yet fully understood.

This study provides a comprehensive and systematic evaluation of a state-of-the-art, modified Inception-based deep learning model's robustness to noise in the context of early breast cancer detection. We quantified the model's diagnostic performance on a large dataset of thermal images systematically corrupted by four distinct and clinically relevant noise types: Gaussian, speckle, salt-and-pepper, and Poisson noise. The intensity of each noise was varied across a wide range to identify performance degradation profiles and critical "tipping points."

Our results demonstrate that while the model achieves exceptional accuracy (99.975%) on clean, noise-free images, its performance is significantly impacted by noise. The nature and severity of this degradation are highly dependent on the noise type. Impulsive noise, such as salt-and-pepper, caused a drastic decline in accuracy to 51.58% at a density of 0.3. Similarly, high-variance speckle noise reduced accuracy to 43.86%. In contrast, the model exhibited greater resilience to Gaussian and Poisson noise, maintaining high accuracy across most tested intensities. Crucially, the application of a pre-processing denoising filter was shown to be highly effective, restoring classification accuracy on corrupted images to near-perfect levels (>99.9%).

This research underscores the critical vulnerability of deep learning-based diagnostic systems to image noise and establishes quantitative benchmarks for model robustness. The findings highlight the indispensable role of advanced noise mitigation strategies and rigorous imaging protocols in developing reliable and clinically viable AI-powered tools for breast cancer thermography

Keywords: Breast Cancer, Deep Learning, Thermal Imaging, Image Noise, Computer-Aided Diagnosis, Noise Mitigation, Model Robustness.

1. Introduction

1.1 Broad Background and Historical Context

Breast cancer represents one of the most significant public health challenges of the 21st century. Globally, it is the most commonly diagnosed cancer among women and a leading cause of cancer-related mortality [1]. The cornerstone of improving patient outcomes and survival rates is early detection, which allows for more effective and less invasive treatment options. For decades, the gold standard for breast cancer screening has been X-ray mammography. Since its widespread adoption, mammography has been credited with reducing breast cancer mortality by enabling the detection of tumors before they become clinically palpable [17]. However, mammography is not without its limitations. The procedure involves exposing the patient to ionizing radiation, which, although low-dose, carries a cumulative risk. Furthermore, it often causes significant discomfort due to breast compression and has reduced sensitivity in women with dense breast tissue, a known risk factor for breast cancer [15].

These limitations have fueled the search for alternative or complementary screening modalities. Among the most promising of these is infrared thermography, a non-invasive and entirely passive imaging technique that involves no ionizing radiation and no physical contact with the patient [7]. The physiological principle underpinning thermography is the detection of infrared radiation naturally emitted from the skin's surface. Malignant tumors are characterized by accelerated metabolic activity and angiogenesis—the formation of new blood vessels to support their growth. This increased cellular activity and blood flow results in localized temperature increases, creating distinct thermal signatures on the breast surface that can be captured by a thermal camera [16]. Studies have shown that thermography can detect these subtle physiological changes, sometimes even before a structural anomaly is visible on a mammogram, highlighting its potential as a tool for early risk assessment [15, 17].

In parallel with advancements in imaging hardware, the field of medical diagnostics has been revolutionized by the advent of artificial intelligence (AI), machine learning, and, most notably, deep learning [2]. Deep learning models, particularly Convolutional Neural Networks (CNNs), have demonstrated extraordinary capabilities in analyzing complex medical imagery, often matching or even exceeding human performance in tasks like lesion detection and classification [4, 5]. By automatically learning hierarchical feature representations from vast datasets, these algorithms can identify intricate patterns in mammograms, ultrasound images, and histopathology slides that may be imperceptible to the human eye [4]. The application of these powerful computational tools to thermal imaging presents a transformative opportunity to automate the interpretation of thermograms, thereby enhancing diagnostic accuracy, objectivity, and efficiency [3, 7].

1.2 Critical Literature Review

The synergy between thermal imaging and deep learning has spawned a vibrant area of research focused on developing computer-aided diagnosis (CAD) systems for breast cancer [4, 7]. Early efforts often relied on traditional machine learning approaches, which involved manual feature extraction from thermograms—analyzing statistical, textural, or morphological properties—followed by classification [3]. However, these methods are often limited by the subjectivity and labor-intensiveness of feature engineering. The shift towards deep learning has allowed for an end-to-end approach where relevant diagnostic features are learned directly from the image data [5].

Numerous studies have explored the application of various CNN architectures for thermogram classification. Researchers have successfully employed established models like VGGNet, ResNet, and Inception, as well as custom-designed networks, to distinguish between healthy and cancerous breast thermograms [7, 20]. For instance, some research has focused on developing specialized, lightweight CNNs, such as BreaCNet, which is optimized for mobile deployment and has achieved remarkable accuracy [23]. Other work has proposed sophisticated models like the sparse deep convolutional autoencoder (SPAER) for extracting low-dimensional biomarkers from dynamic thermal data, showing robustness even when Gaussian noise was added [19]. The use of cascaded CNN architectures has also been investigated to segment and classify tumors in infrared images, again demonstrating high accuracy after the addition of Gaussian noise to augment the dataset [22].

Despite these successes, a critical and often underestimated challenge in the clinical application of these technologies is the presence of noise in thermal images. Image noise, defined as undesirable random variations in pixel intensity, can arise from multiple sources, including the imaging sensor itself (thermal electronic noise), environmental interference, or errors during image transmission and storage [9, 10]. Thermal cameras, especially uncooled microbolometer-based systems common in medical applications, are susceptible to a variety of noise phenomena [14]. These include Gaussian noise, which arises from thermal agitation of electrons in the sensor circuitry [27]; salt-and-pepper (or impulsive) noise, resulting from faulty sensor pixels or data transmission errors [13]; speckle noise, a granular pattern that can be caused by scattering effects and sensor non-uniformities [25, 26]; and Poisson (or shot) noise, which stems from quantum fluctuations in the arrival of infrared photons, particularly in low-signal conditions [30, 31]. Each of these noise types has a distinct statistical profile and can corrupt an image in different ways, potentially obscuring the subtle thermal asymmetries that are indicative of malignancy [10, 40].

The impact of noise on deep learning-based analysis has been acknowledged in the literature, but often in a limited or fragmented manner. Several studies have used noise, primarily Gaussian noise, as a form of data augmentation to increase the size and variability of the training dataset, with the goal of improving model generalization [18, 19, 22]. While this approach can enhance robustness, it does not systematically evaluate the model's performance limits under different noise conditions. Some researchers have focused on specific noise removal techniques as a pre-processing step. For example, methods have been proposed for removing noise from mammograms [12] or using advanced segmentation techniques that are robust to intensity uncertainties [11]. Work by Ekici and Jawzal [20] incorporated pre-processing for salt-and-pepper noise before feeding images into a CNN. Similarly, other research has used advanced segmentation models like 4D U-Net to reduce speckle noise in thermal images [24] or applied color segmentation after noise removal [21]. This body of work confirms that noise is a recognized problem and that pre-processing is a viable solution.

1.3 Research Gap

A review of the current literature reveals a significant research gap. While previous studies have investigated noise in breast thermography, they tend to suffer from one or more of the following limitations: (1) they focus on only a single type of noise, most commonly Gaussian noise [18, 19]; (2) they do not systematically vary the intensity of the noise to understand the model's performance degradation profile; (3) they do not provide a direct, controlled comparison of a model's robustness to different noise types (e.g., Gaussian vs. Speckle vs. Salt & Pepper); and (4) they often use noise for data augmentation rather than for a rigorous stress test of the model's reliability.

Consequently, there is a pressing need for a comprehensive, quantitative investigation that systematically assesses the resilience of a state-of-the-art deep learning model against a variety of clinically relevant noise types at a range of intensities. It is crucial to understand not only that noise degrades performance, but precisely *how* and *at what level* different noises impact diagnostic accuracy. Identifying these "tipping points"—the noise thresholds beyond which a model's predictions become unreliable—is essential for establishing confidence in automated diagnostic systems. Furthermore, while denoising is a known solution, its effectiveness in restoring the performance of a highly sensitive deep learning model across these diverse noise conditions needs to be rigorously quantified.

1.4 Objectives and Hypotheses

This study aims to address the identified research gap through a systematic and controlled experimental investigation. The primary objectives are:

1. To establish the baseline diagnostic performance of a modified Inception-based deep learning architecture on a large, clean dataset of breast thermograms.

2. To systematically quantify the impact of four distinct noise types—Gaussian, speckle, salt-and-pepper, and Poisson—at varying levels of intensity on the model's classification accuracy, sensitivity, and specificity.

3. To identify the performance "tipping points" for each noise type, where diagnostic reliability degrades significantly.

4. To evaluate the efficacy of a pre-processing denoising filter in mitigating noise-induced classification errors and restoring the model's diagnostic performance.

Based on these objectives, we formulated the following hypotheses:

- **Hypothesis 1:** The deep learning model will achieve exceptionally high diagnostic accuracy (exceeding 99%) when trained and tested on the noise-free thermal image dataset.

- **Hypothesis 2:** The model's performance will degrade with increasing noise intensity for all noise types. However, the degradation profile will differ significantly across the noises. We hypothesize that impulsive noises (salt-and-pepper) and multiplicative noises (speckle) will cause a more abrupt and severe performance collapse at specific intensity thresholds compared to additive Gaussian noise.

- **Hypothesis 3:** The application of a denoising filter as a pre-processing step will significantly improve the model's classification accuracy on the corrupted images, restoring performance to levels statistically comparable to the noise-free baseline.

By testing these hypotheses, this research seeks to provide critical insights into the practical challenges of deploying deep learning systems for breast thermography and to offer evidence-based guidance for the development of more robust and reliable diagnostic tools.

2. Methods

2.1 Research Design

This study was conducted as a quantitative, controlled experimental investigation. The core of the research design involved the systematic manipulation of the primary independent variable: image noise. This variable was operationalized across two dimensions: noise type (Gaussian, speckle, salt-and-pepper, and Poisson) and noise intensity (defined by parameters such as variance, density, or signal-to-noise ratio). The primary dependent variable was the diagnostic performance of a deep learning model, measured using a comprehensive set of classification metrics.

The experiment was structured into four main scenarios to provide a multi-faceted evaluation:

1. **Benchmark Scenario:** The model's baseline performance was established by training and testing it on the original, clean dataset. This served as the "gold standard" against which all other results were compared.

2. **Noisy Dataset Scenario:** To assess the model's robustness, it was trained and tested on datasets to

which a specific type of noise had been uniformly applied to all images.

3. **Mixed-Data Scenario:** The model was evaluated on a dataset comprising a 50/50 mix of clean and noisy images to simulate a more heterogeneous data environment and test the model's ability to generalize.

4. **Tipping Point Analysis:** A focused analysis was conducted where noise of increasing intensity was systematically added to individual images fed into the pre-trained model. This allowed for the precise identification of the noise threshold at which classification accuracy begins to degrade significantly, flipping from a correct to an incorrect prediction.

For each relevant scenario, the effectiveness of a denoising countermeasure was also evaluated by comparing the model's performance on noisy images versus its performance on the same images after a denoising filter was applied.

2.2 Participants / Sample

The "participants" in this study consisted of digital thermal images obtained from a publicly available repository for mastology research with infrared images [35]. The dataset is specifically curated for the development and evaluation of machine learning algorithms for breast cancer detection. The full dataset comprised 1,800 thermal images, categorized into two classes: 1,000 images corresponding to cases with confirmed cancerous lesions and 800 images from healthy controls.

The acquisition of these images followed a strict protocol to ensure data quality and consistency. Examinations were conducted in a temperature-controlled room (20-22°C). Patients were required to acclimatize to the room temperature and to abstain from activities or substances that could alter their thermal state (e.g., hot beverages, creams) for at least two hours prior to the exam. Images were captured with a high-sensitivity thermal camera (FLIR SC620, sensitivity < 0.04°C) positioned at a standard distance of one meter from the patient. This rigorous standardization helps to minimize extraneous thermal artifacts, though it does not eliminate the potential for inherent sensor and electronic noise, which is the focus of this study.

2.3 Materials and Apparatus

Hardware and Software: All experiments were conducted on a high-performance desktop computer equipped with an Intel Core i7 processor, 32 GB of RAM, a 1 TB solid-state drive, and a dedicated NVIDIA GPU (6 GB VRAM) to accelerate deep learning computations. The entire experimental pipeline, including image processing, model development, training, and analysis, was implemented in the MATLAB (version 2020a) environment, utilizing its Deep Learning and Image Processing Toolboxes.

Deep Learning Model Architecture: The core of our investigation was a modified deep convolutional neural network based on the Inception architecture. The foundation for our model was the Inception-V4 architecture, known for its deep and efficient design that uses "inception modules" to perform multi-scale feature extraction in parallel [33, 34]. To optimize performance for the specific task of thermal image analysis, we introduced a modification to the standard Inception-V4 model, creating a variant we refer to as Inception MV4. The key modification was made within the "Inception B" module of the network. Specifically, an additional convolutional layer was inserted after the average pooling layer, and the number of filters in this path was increased from 128 to 256 to enhance feature extraction capacity. Concurrently, to maintain a balanced network depth and computational load, some of the deeper layers within the original Inception B block were streamlined [32]. This modified architecture, Inception MV4, was selected after preliminary experiments showed it provided a superior balance of accuracy and training efficiency compared to the standard Inception V3 and V4 models for this task.

Noise Models: Four distinct noise models were implemented to corrupt the thermal images, each representing a different physical phenomenon:

- **Gaussian Noise:** An additive noise model where the intensity value of each pixel is perturbed by a random value drawn from a normal (Gaussian) distribution. It is defined by its mean (μ) and variance (σ2) and simulates electronic thermal noise [27].

- **Speckle Noise:** A multiplicative noise model that is granular in appearance. The noise is modeled by multiplying pixel values by random values with a mean of 1 and a specified variance. It is characteristic of coherent imaging but can also arise from sensor non-uniformities in thermal systems [25, 36, 37].

- **Salt-and-Pepper Noise:** An impulsive noise model that randomly replaces a certain percentage (density) of image pixels with either maximum (salt) or minimum (pepper) intensity values. It simulates dead pixels or data transmission errors [40].

- **Poisson Noise:** A signal-dependent noise model where the noise variance is proportional to the pixel intensity. It stems from the quantum statistics of photon detection and is most prominent in low-light (or low-temperature) conditions [30, 41].

Denoising Algorithm: To evaluate noise mitigation strategies, a pre-trained, deep learning-based denoising model was employed. Specifically, a Denoising Convolutional Neural Network (DnCNN) was used. DnCNN is designed to effectively remove specific types of noise (particularly Gaussian) while preserving critical image details and structures, a key requirement for medical diagnostics. The denoising process was applied to each RGB channel of a noisy image independently before the channels were recombined to form the final denoised image.

2.4 Data Collection Procedure

The experimental procedure was executed in a sequential and controlled pipeline:

1. **Pre-processing and Augmentation:** All images from the dataset were first resized to a uniform input size required by the Inception architecture. A pre-processing step was applied to automatically crop the images to the region of interest (ROI), focusing on the breast area while removing extraneous regions like the neck, arms, and abdomen. To expand the training dataset and prevent the model from overfitting, a series of data augmentation techniques were applied in real-time during training. These included random horizontal flipping, random vertical flipping, and random rotation of up to 30 degrees in either direction.

2. **Dataset Partitioning:** The full dataset of 1,800 images was partitioned into a training set and a testing set using a 70/30 split, respectively. This resulted in 1,260 images for training the model and 540 images for its final, unbiased evaluation. The split was stratified to maintain the same proportion of healthy and cancerous images in both subsets.

3. **Noise Injection:** For the experiments involving noisy data, noise was programmatically injected into the images of the partitioned datasets. This was done systematically for each of the four noise types. The intensity parameters were varied across a pre-defined range: Gaussian noise variance was varied from 0.01 to 0.09; speckle noise variance from 0.02 to 0.08; salt-and-pepper noise density from 0.1 to 0.3; and Poisson noise was applied to simulate different signal-to-noise ratios (SNRs).

4. **Model Training:** The Inception MV4 model was trained using a set of optimized hyperparameters that were determined through empirical tuning. The Stochastic Gradient Descent with Momentum (SGDM) optimizer was used. A learning rate of 1e-4, a mini-batch size of 10, and a training duration of 10-30 epochs were chosen to balance convergence speed and stability. During training, the validation accuracy was monitored to prevent overfitting, and training was stopped if performance on a validation set plateaued.

2.5 Data Analysis

The performance of the Inception MV4 model under all experimental conditions was quantified using a comprehensive suite of standard performance metrics derived from the confusion matrix (True Positives, True Negatives, False Positives, False Negatives):

- **Accuracy:** The overall proportion of correct classifications.

- **Sensitivity (Recall):** The ability of the model to correctly identify positive (cancerous) cases.

- **Specificity:** The ability of the model to correctly identify negative (healthy) cases.

- **Precision:** The proportion of positive predictions that were actually correct.

- **Negative Predictive Value (NPV):** The proportion of negative predictions that were actually correct.

- **F1-Score:** The harmonic mean of precision and sensitivity, providing a single metric that balances both.

- **False Positive Rate (FPR):** The proportion of healthy cases incorrectly classified as cancerous.

- **False Negative Rate (FNR):** The proportion of cancerous cases incorrectly classified as healthy.

- **Area Under the ROC Curve (AUC):** A global measure of classification performance across all possible decision thresholds.

- **Equal Error Rate (EER):** The rate at which the FPR and FNR are equal.

To determine the statistical significance of the observed performance differences between experimental conditions (e.g., clean vs. noisy, noisy vs. denoised), pairwise t-tests were conducted on the accuracy scores obtained over multiple experimental runs. A p-value of less than 0.05 was considered to be statistically significant.

3. Results

This section presents the empirical findings of the study, reported objectively and without interpretation. The results are structured to first establish the baseline performance of the Inception MV4 model, followed by a detailed analysis of its robustness under various noise conditions and the effectiveness of denoising countermeasures.

3.1 Preliminary Analyses: Baseline Performance

The initial phase of the experiment was to establish the benchmark performance of the modified Inception MV4 model on the clean, noise-free dataset. When trained and tested on the 70/30 split of the original 1,800 thermal images, the model demonstrated exceptionally high diagnostic capability. The Inception MV4 model achieved an average classification accuracy of 99.975%. This performance surpassed that of the standard Inception V3 (98.104% accuracy) and the standard Inception V4 (99.971% accuracy) models under the same training conditions, validating its selection for this study. The high performance was also reflected in other key metrics, with a sensitivity of 0.994, specificity of 1.000, precision of 1.000, and an F1-score of 0.997. The model produced no false positives, with only a single false negative across the entire test set. This strong baseline confirms Hypothesis 1 and provides a high standard against which the impact of noise can be measured.

3.2 Main Findings: Impact of Noise and Denoising

The core of the investigation involved assessing the model's performance when subjected to four different types of noise at varying intensities.

Gaussian Noise: The model exhibited remarkable resilience to Gaussian noise. When tested on images corrupted with Gaussian noise (mean=0, variance varied from 0.01 to 0.09), the model maintained a very high level of accuracy. Even at higher variance levels, the average detection accuracy remained consistently high. For instance, in a dataset where all images were corrupted by Gaussian noise with variances of 0.05 and 0.02, the average accuracy over 10 epochs was 98.8%. In single-image tests with fixed variance (0.02) and varying mean (0.02 to 0.08), the accuracy remained at or near 100%, with no critical tipping point observed within the tested range. This suggests the model's convolutional layers are effective at averaging out this type of distributed, additive noise.

Speckle Noise: The model's performance under speckle noise was highly dependent on the noise variance. For lower variance levels (0.02, 0.03, and 0.04), the model maintained a perfect 100% accuracy. However, a distinct performance degradation was observed as the variance increased. At a variance of 0.07, the average accuracy dropped significantly to 89.66%. A clear tipping point was identified at a variance of 0.08, where the model's performance collapsed catastrophically, with the average accuracy plummeting to 43.86%. At this level, healthy images were frequently misclassified as cancerous.

Salt-and-Pepper Noise: The model was extremely sensitive to salt-and-pepper (impulsive) noise, especially at higher densities. It maintained perfect 100% accuracy for noise densities of 0.1 and 0.2. However, performance began to decline at a density of 0.26 (97.88% accuracy) and continued to drop as the density increased. A tipping point was observed at a noise density of 0.3, where the average detection accuracy fell to just 51.58%. This indicates that the sharp, high-contrast artifacts introduced by this noise type are highly disruptive to the model's feature extraction process.

Poisson Noise: The model demonstrated strong robustness against Poisson noise across a wide range of signal-to-noise ratios (SNRs). From an SNR of 13.98 dB down to 1.94 dB, the average detection accuracy remained stable at 99.9%. While there was a very slight, gradual decrease in sensitivity (from 100% down to 99.06%) as the noise level increased (SNR decreased), there was no evidence of a sudden performance collapse or tipping point. The model effectively handled this signal-dependent noise, maintaining high reliability even in simulated low-signal conditions.

Effectiveness of Denoising: The application of a pre-processing denoising filter proved to be a highly effective countermeasure against noise-induced classification errors. This was starkly illustrated in the tipping point analyses.

- For **speckle noise**, a healthy image corrupted with a variance of 0.09 was initially misclassified as "Cancer" with 99.58% confidence. After the denoising filter was applied, the same image was correctly re-classified as "Healthy" with 99.99% confidence.
- For **salt-and-pepper noise**, a healthy image with added noise was misclassified as "Cancer." After denoising, it was correctly identified as "Healthy" with 99.999% confidence.
- Similarly, for **Gaussian and Poisson noise**, images that were misclassified due to high noise levels were correctly classified with near-perfect confidence after the denoising algorithm was applied. This confirms Hypothesis 3, demonstrating that noise mitigation is a critical step for restoring diagnostic reliability.

3.3 Exploratory Findings: Mixed Data and Statistical Significance

When the model was trained on a mixed dataset (50% clean, 50% Gaussian noise), its average accuracy was 98.22%. While this is a high level of performance, it is notably lower than the accuracy achieved on the purely clean dataset (99.974%) and slightly lower than the performance on the fully noisy dataset (98.8%). This suggests that while training on some noise can confer robustness, it may not be sufficient to overcome the variability introduced by a mixed-quality dataset without leading to a slight drop in peak performance compared to a clean baseline.

Statistical analysis using pairwise t-tests confirmed the significance of these observations. There was a statistically significant difference in detection accuracy when comparing the performance on the Gaussian noise dataset to the clean DMR IR dataset (p = 0.0418) and the mixed DMR IR + Noise dataset (p = 0.0447). This verifies that the presence of noise has a significant negative impact on performance. Conversely, there was no statistically significant difference between the performance on the clean DMR IR dataset and the mixed dataset when considering the overall stability (p = 0.5109), though the peak accuracy was lower. These statistical results provide quantitative support for the main findings.

4. Discussion

4.1 Interpretation

The results of this study provide a nuanced and quantitative understanding of the interplay between deep learning models, thermal imaging, and noise in the context of breast cancer detection. Our first key finding—the near-perfect 99.975% accuracy of the Inception MV4 model on a clean dataset—firmly establishes the immense potential of this technology under ideal conditions. This high level of performance can be attributed to the model's sophisticated architecture, which uses multi-scale inception modules to effectively capture both the fine-grained textural details and the broader spatial thermal patterns that differentiate healthy from malignant tissue [32, 34].

However, the central contribution of this work lies in the interpretation of the model's behavior under non-ideal, noisy conditions. Our findings clearly support Hypothesis 2: the model's performance degrades with noise, but the degradation profile is highly contingent on the specific characteristics of that noise. The differential impact of the four noise types can be explained by how they interact with the fundamental operations of a CNN. The model's striking resilience to Gaussian noise, for instance, is likely due to the noise's additive and zero-mean nature. The convolutional filters, which act as local averaging operators, can effectively suppress this type of distributed, random fluctuation without losing significant structural information from the underlying image.

In stark contrast, the model's vulnerability to salt-and-pepper noise reveals a critical weakness. This impulsive noise introduces pixels with maximum or minimum intensity, creating sharp, high-frequency artifacts. These artifacts can be mistaken by the network's filters for salient features, such as the small, intense "hotspots" that might be associated with a tumor. This leads to catastrophic misclassification, as evidenced by the performance collapse to ~51% accuracy—barely better than random chance—at a density of just 0.3. A similar phenomenon occurs with high-variance speckle noise. As a multiplicative noise, it corrupts the image by altering the texture and grain of tissue regions. At a critical variance (the "tipping point" of 0.08), this textural distortion becomes so severe that it overwhelms the genuine thermal patterns, causing the model's learned feature representations to fail and leading to a precipitous drop in accuracy.

The model's robustness to Poisson noise can be attributed to its signal-dependent nature. Since Poisson noise has a variance equal to the signal intensity, it is more pronounced in brighter (hotter) regions of the thermogram. However, the diagnostic information in thermography often lies in the relative temperature differences and patterns, rather than the absolute temperature of the hottest point. The model appears capable of learning these relational patterns even when the high-intensity areas are noisy.

Finally, the dramatic success of the denoising filter is profoundly significant. The ability to take an image that was misclassified with over 99% confidence and, after filtering, have it correctly classified with over 99% confidence underscores a critical point: the diagnostic information was not destroyed by the noise, but merely obscured. The denoising algorithm successfully separated the structured signal (the thermal pattern) from the unstructured noise, enabling the model to function as intended. This validates Hypothesis 3 and highlights that the problem is not an inherent flaw in the model's diagnostic logic, but rather its sensitivity to data quality.

4.2 Comparison with Literature

Our findings both align with and extend the existing body of literature. The high baseline accuracy we achieved is comparable to other state-of-the-art models reported in recent studies, such as the 100% accuracy claimed for the BreaCNet model on a specific dataset, reinforcing the viability of deep learning for this task [23]. Our investigation into Gaussian noise corroborates the work of researchers who have previously used it for data augmentation or noted its presence [18, 19]. However, our systematic evaluation across a wide range of variances provides a more granular understanding than previously available, demonstrating a high degree of resilience rather than a simple degradation.

The most significant extension of the literature comes from our direct, controlled comparison of multiple noise types. While individual studies have acknowledged different noises, such as salt-and-pepper [20] or speckle [24], none have provided a quantitative, side-by-side analysis of their relative impact on a single, powerful deep learning model. Our identification of distinct "tipping points" for impulsive and multiplicative noise is a novel contribution that quantifies the operational limits of these systems in a way that has not been previously reported. This provides a crucial benchmark for future research.

Furthermore, our results strongly support the general consensus on the importance of pre-processing and noise mitigation in medical imaging analysis [9, 10, 11, 12]. While studies have previously demonstrated the utility of denoising filters for improving image quality or aiding segmentation [11, 21], our work provides direct evidence of their critical role in preventing catastrophic classification failures in an end-to-end deep learning pipeline. This reinforces the idea that robust pre-processing is not merely an optional enhancement but a mandatory component for any clinical-grade AI system in this domain. Our work also complements broader reviews on the use of thermography and neural networks, providing specific empirical data on the challenges highlighted in those reviews [7, 17].

4.3 Strengths and Limitations

This study has several notable strengths. Its primary strength is the comprehensive and systematic methodology used to compare the effects of four different and relevant noise types on a state-of-the-art deep learning model. The use of a large, publicly available dataset enhances reproducibility [35]. The quantitative identification of performance "tipping points" provides novel and valuable benchmarks for the field. Finally, the clear demonstration of the effectiveness of denoising as a countermeasure offers a strong, evidence-based practical recommendation.

Despite these strengths, it is important to acknowledge the study's limitations. First, the investigation relied on artificially generated noise injected into clean images. While the noise models used are standard and mathematically sound, they may not perfectly replicate the complex, often mixed, noise profiles encountered in real-world clinical environments, where multiple noise sources may coexist. Second, our analysis was confined to a single, albeit powerful, deep learning architecture (Inception MV4). It is possible that other architectures, such as those incorporating attention mechanisms or different regularization strategies, might exhibit different degrees of robustness. Third, the study focused on the binary classification task of "healthy" versus "cancerous." It did not explore how noise might affect more nuanced tasks, such as differentiating between different types of tumors or staging cancer severity. Lastly, while the dataset used was substantial, it originates from a single source, and the findings would be strengthened by validation on data from different clinics and camera systems.

4.4 Implications

The findings of this research have significant implications for both clinical practice and technical development.

Practical and Clinical Implications: For clinicians, medical physicists, and healthcare providers, this study serves as a crucial reminder that the output of AI-powered diagnostic tools is highly dependent on input data quality. A "black box" approach to AI is insufficient; there must be an awareness of potential failure modes. Our results strongly advocate for the establishment and strict adherence to standardized imaging protocols designed to minimize noise at the source [14]. This includes controlling the ambient environment, ensuring proper camera calibration, and maintaining sensor integrity. Furthermore, the quantitative evidence of denoising effectiveness suggests that robust noise-filtering modules should be considered a mandatory component of any CAD system for thermography before it is deployed in a clinical setting, to act as a safety net against poor-quality image acquisition.

Research and Technical Implications: For AI researchers and developers, our findings challenge the community to move beyond optimizing models solely on pristine, curated datasets. Robustness to real-world imperfections must become a primary design criterion. The performance degradation profiles and tipping points identified in this study can serve as valuable benchmarks for stress-testing new models. The research should spur the development of more sophisticated, "noise-aware" deep learning architectures that are inherently more resilient to data corruption. This could involve exploring novel regularization techniques, loss functions that are less sensitive to outliers, or architectures that explicitly model and separate noise. Our work also highlights the continued importance of research into advanced denoising algorithms, particularly those tailored to the specific statistical properties of thermal imaging and the preservation of subtle diagnostic features [37, 38, 39, 40].

4.5 Conclusion and Future Directions

In conclusion, this study provides a comprehensive quantitative assessment of the resilience of deep learning-based breast cancer detection in the face of image noise. We have demonstrated that while a state-of-the-art model like

Inception MV4 can achieve near-perfect accuracy under ideal conditions, its reliability is critically vulnerable to noise, with impulsive and multiplicative noises like salt-and-pepper and speckle posing the most significant threat. We have identified specific performance "tipping points" that define the operational limits of such systems. Most importantly, we have confirmed that this vulnerability can be effectively managed through the application of pre-processing denoising filters, which successfully restore diagnostic accuracy. The central takeaway is that for AI to be a reliable partner in clinical diagnostics, a dual focus on both powerful algorithms and robust data integrity is not just beneficial, but essential.

Building on the foundation of this work, several promising avenues for future research emerge. First, it is crucial to validate these findings using a large, multi-center clinical dataset containing images with naturally occurring, mixed-type noise. Second, future work should explore the development of novel deep learning architectures with built-in noise resilience, potentially through adversarial training or by incorporating noise modeling directly into the network layers. Third, the scope of analysis should be expanded beyond binary classification to investigate how noise impacts the ability to predict cancer subtypes, stage, or grade. Finally, there is a compelling need to design and evaluate lightweight, noise-robust CNN models that are computationally efficient enough for deployment on low-power or mobile platforms, which could democratize access to this promising screening technology [5, 6, 8]. Pursuing these directions will be key to transitioning this technology from a promising research concept to a trusted tool in the global fight against breast cancer.

References

[1] Hanf V, Kreienberg R. Corpus Uteri. 2020.

[2] Bini SA. Artificial Intelligence Machine Learning Deep Learning and Cognitive Computing: What Do These Terms Mean and How Will They Impact Health Care? J Arthroplasty. 2018;33:2358-2361.

[3] Yadav P, Jethani V. Breast Thermograms Analysisfor Cancer Detection Using Feature Extraction and Data Mining Technique. ACM Int Conf Proceeding Ser. 2016:1-5.

[4] Din NM, Dar RA, Rasool M, Assad A. Breast Cancer Detection Using Deep Learning: Datasets Methods and Challenges Ahead. Comput Biol Med. 2022;149:106073.

[5] Salvi S, Kadam A. Breast Cancer Detection Using Deep Learning and IoT Technologies. J Phys Conf Ser. 2021;1831:012030.

[6] Zhao P, Yoo I, Lancey R, Varghese E. Mobile Applications for Pain Management: An App Analysis for Clinical Usage. BMC Med Inform Decis Mak. 2019;19:1-10.

[7] Al Husaini MA, Habaebi MH, Hameed SA, Islam MR, Gunawan TS. A Systematic Review of Breast Cancer Detection Using Thermography and Neural Networks. IEEE Access. 2020;8:208922-208937.

[8] Al Husaini MA, Habaebi MH, Islam MR, Gunawan TS. Self-Detection of Early Breast Cancer Application With Infrared Camera and Deep Learning. Electron. 2021;10:2538.

[9] Hiremath S, Karibasappa KG, Karibasappa K. Neural Network Based Noise Identification in Digital Images. ACEEE Int. J Netw Secur. 2011;02:28-31.

[10] Salami AM, Salih DM, Fadhil AF. Thermal Image Features and Noise Effects Analysis. In: Proceedings of the 7th international engineering conference research and innovation amid global pandemic. IEC Institute of Electrical and Electronics Engineers Inc. New York: IEEE. 2021:43-47.

[11] Liu Q, Liu Z, Yong S, Jia K, Razmjooy N. Computer-Aided Breast Cancer Diagnosis Based on Image Segmentation and Interval Analysis. Automatika. 2020;61:496-506.

[12] Priyadharsini MS. High Density Noise Filter Method for Denoising Mammogram Breast. Data cquisition Process. 2023;38.

[13] Sommer K, Plez B, Cohen-Tanugi J, Dagoret-Campagne S, Moniez M, et al. Stardice II: Calibration of an Uncooled Infrared Thermal Camera for Atmospheric Gray Extinction Characterization. Sensors. 2024;24:4498.

[14] Gade R, Moeslund TB. Thermal Cameras and Applications: A Survey. Mach Vis Appl. 2014;25:245-262.

[15] Wishart GC, Campisi M, Boswell M, Chapman D, Shackleton V, et al. The Accuracy of Digital Infrared Imaging for Breast Cancer Detection in Women Undergoing Breast Biopsy. Eur J Surg Oncol. 2010;36:535-540.

[16] Antony L, Arathy K, Sudarsan N, Muralidharan MN, Ansari S. Breast Tumor Parameter Estimation and Interactive 3D Thermal Tomography Using Discrete Thermal Sensor Data. Biomed Phys Eng Express. 2020;7:015013.

[17] Husaini MA, HABAEBI MH, HAMEED SA, ISLAM MR, GUNAWAN TS. A Systematic Review of Breast Cancer Detection Using Thermography and Neural Networks. IEEE Access. 2020;8:208922-208937.

[18] Mulaveesala R, Dua G. Non-invasive and Non-ionizing Depth Resolved Infra-Red Imaging for Detection and Evaluation of Breast Cancer: A Numerical Study. Biomed

Phys Eng Express. 2016;2:1-5.

[19] Yousefi B, Akbari H, Hershman M, Kawakita S, Fernandes HC, et al. SPAER: Sparse Deep Convolutional Autoencoder Model to Extract Low Dimensional Imaging Biomarkers for Early Detection of Breast Cancer Using Dynamic Thermography. Appl Sci. 2021;11:3248.

[20] Ekici S, Jawzal H. Breast Cancer Diagnosis Using Thermography and Convolutional Neural Networks. Med Hypotheses. 2020;137:109542.

[21] Kermani S, Samadzadehaghdam N, EtehadTavakol M. Automatic Color Segmentation of Breast Infrared Images Using a Gaussian Mixture Model. Optik. 2015;126:3288-3294.

[22] Dalmia A, Kakileti ST, Manjunath G. Exploring Deep Learning Networks for Tumour Segmentation in Infrared Images. 14th Quant InfraRed Thermogr Conf. 2018.

[23] Roslidar R, Syaryadhi M, Saddami K, Pradhan B, Arnia F, et al. Breacnet: A High-Accuracy Breast Thermogram Classifier Based on Mobile Convolutional Neural Network. Math Biosci Eng. 2022;19:1304-1331.

[24] Gomathi P, Muniraj C, Periasamy PS. Digital Infrared Thermal Imaging System Based Breast Cancer Diagnosis Using 4D U-Net Segmentation. Biomed Signal Process Control. 2023;85:104792.

[25] Goodman JW. Speckle Phenomena in Optics: Theory and Applications, 2nd ed. Bellingham, WA, USA: SPIE Press, 2020;PM312.

[26] Hou F, Zhang Y, Zhou Y, Zhang M, Lv B, et al. Review on Infrared Imaging Technology. Sustainability. 2022;14:11161.

[27] Robinson S. Editor. The Infrared & Electro-Optical Systems Handbook, Vol. 8: Emerging Systems and Technologies. Bellingham, WA, USA: SPIE Press. 1993.

[28] Stewart J. Human Medical Thermography. 2023.

[29] Raju PR, Hussain SA. A Computer-Aided Diagnosis Tool for Objective Assessment of Tumors Using Infrared Imaging. Int J Eng Res Comput Appl. 2014;3:10–16.

[30] Zmuidzinas J. Thermal Noise and Correlations in Photon Detection. Appl Opt. 2003;42:4989-5008.

[31] Besikci C. Nature Allows High Sensitivity Thermal Imaging With Type-I Quantum Wells Without Optical Couplers: A Grating-Free Quantum Well Infrared Photodetector With High Conversion Efficiency. IEEE J Quantum Electron. 2021;57:1-12.

[32] Al Husaini MA, Habaebi MH, Gunawan TS, Islam MR, Elsheikh EA, et al. Thermal-Based Early Breast Cancer Detection Using Inception V3 Inception V4 and Modified Inception MV4. Neural Comput Appl. 2022;34:333-348.

[33] Szegedy C, Ioffe S, Vanhoucke V, Alemi A. Inception-V4 Inception-ResNet and the Impact of Residual Connections on Learning. 2016. ArXiv preprint:- https://arxiv.org/pdf/1602.07261v1.

[34] Szegedy C, Ioffe S, Vanhoucke V, Alemi AA. Inception-V4 Inception-ResNet and the Impact of Residual Connections on Learning (AAAI-17). AAAI. 2017;31:4278-4284.

[35] https://visual.ic.uff.br/en/proeng/thiagoelias/

[36] Goodman JW. Some Fundamental Properties of Speckle. J Opt Soc Am. 1976;66:1145.

[37] Hiremath PS, Akkasaligar PT, Badiger S, Gunarathne G. Speckle Noise Reduction in Medical Ultrasound Images. In: Gunarathne GP editor. Adv. break. ultrasound imaging. InTech. 2013;1:1-8.

[38] Sudha S, Suresh GR, Sukanesh R. Speckle Noise Reduction in Ultrasound Images by Wavelet Thresholding Based on Weighted Variance. Int J Comput Theor Eng. 2009;1:7-12/1793-8202.

[39] Buades A, Coll B, Morel JM. A Non-local Algorithm for Image Denoising. Proc IEEE Comput Soc Conf. Comput Vis Pattern Recognition. CVPR. 2005;2:60-65.

[40] Rohit V, Ali J. A Comparative Study of Various Types of Image Noise and Efficient Noise Removal Techniques. Int J Adv Res Comput Sci Softw Eng. 2013;3:2277–2128.

[41] Salmon J, Harmany Z, Deledalle CA, Willett R. Poisson Noise Reduction With Non-local PCA. J Math Imaging Vis. 2014;48:279-294.