Image Anal Stereol 2025;44:111-129 doi: 105566/ias.3561
Original Research Paper

RobustDeiT: NOISE-ROBUST VISION TRANSFORMERS FOR MEDICAL IMAGE CLASSIFICATION

MEHDI TAASSORI
Institute of Cyber-Physical Systems, John von Neumann Faculty of Informatics, Obuda University, Budapest, Hungary
e-mail: taassori.mehdi@uni-obuda.hu
(Received February 21, 2025; revised June 27, 2025; accepted June 27, 2025)

ABSTRACT

Effective classification of medical images is vital for accurate diagnosis and treatment, but noisy datasets remain a significant challenge, obscuring critical features and leading to unreliable predictions. To address this, we propose RobustDeiT, a noise-robust architecture based on the Data-efficient Image Transformer (DeiT), tailored for medical image classification in noisy environments. By integrating a multi-stage preprocessing pipeline, our approach systematically reduces noise, enhances contrast, and highlights fine details, ensuring the preservation of essential features. Advanced denoising methods, contrast enhancement with Contrast Limited Adaptive Histogram Equalization, and sharpening via unsharp masking collectively improve image quality, enabling the model to extract meaningful patterns. Extensive evaluations demonstrate that RobustDeiT achieves superior performance across diverse metrics, establishing its effectiveness in handling noisy medical imaging datasets and paving the way for reliable and accurate classification in real-world scenarios.

Keywords: Classification, Data-efficient Image Transformer, Medical imaging, Noise Robustness.

INTRODUCTION

Medical image classification is a cornerstone of modern healthcare, serving as a key enabler for accurate diagnosis and effective treatment planning. Despite its critical importance, the task is often hindered by the pervasive presence of noise in medical imaging datasets, which can significantly impair the performance of classification models. Noise obscures essential diagnostic features, introducing ambiguity and reducing the reliability of predictions. Addressing this challenge requires a delicate balance, mitigating noise effectively while preserving critical image details. Failure to manage this balance can result in over-smoothed images, where vital diagnostic features are blurred or lost, undermining the accuracy and robustness of classification models. Consequently, the design of classifiers that can perform reliably on noisy datasets has become a pressing need in the field.

To overcome these challenges, we present a novel architecture rooted in the Data-efficient Image Transformer (DeiT), tailored specifically for the classification of noisy medical images. Our approach incorporates an advanced preprocessing pipeline, combining techniques such as edge-preserving filters, Gaussian and median filtering, contrast enhancement via CLAHE, and sharpening through unsharp masking. These methods collectively mitigate the effects of noise while retaining critical diagnostic features essential for accurate classification. By leveraging the DeiT framework, our model utilizes the strengths of transformers to achieve robust and reliable performance in medical image classification. Designed with noise-resilience at its core, our approach ensures dependable classification even in the presence of noisy datasets.

Medical image classification involves the use of advanced computational models to identify patterns and features within images, allowing for the categorization of medical conditions. Recent advancements in deep learning have demonstrated significant improvements in classification accuracy. These models have shown great potential in automating the diagnostic process, reducing human error, and enabling faster decision-making in clinical settings. In (Ling *et al*., 2024), the authors propose a multi-task attention network (MTANet) to address both medical image segmentation and classification efficiently. Their model incorporates a reverse addition attention module for segmentation and an attention bottleneck module for classification, fusing image and clinical features. This reflects the trend of enhancing feature extraction in medical image classification,

aligning with our focus on improving classification performance in noisy datasets using Vision Transformers.

Noise in medical images can significantly hinder the performance of classification models. Noise in medical images, whether from errors in labeling or from issues during image capture, can reduce the quality of data used for training. This can lead to overfitting, reduced classification accuracy, and unreliable model predictions. Researchers have explored various methods to mitigate the adverse effects of noise, focusing on label correction, dual-network learning, and uncertainty estimation techniques to improve model robustness in the presence of noisy data. In (Penso *et al*., 2024), the authors address the challenge of calibrating neural networks for medical image classification in the presence of label noise. They propose a noise-robust calibration procedure that estimates the noise level in the labels and integrates it into the training process. By incorporating the noise level into the network's accuracy estimation, the method ensures reliable calibration results, even when using noisy, unreliable labels. In (Liu *et al*., 2021), the authors propose a noise-tolerant medical image classification framework, Co-Correcting, to address the challenge of label noise, which often affects the accuracy of classifiers in medical image analysis. Their approach integrates dual-network mutual learning, label probability estimation, and curriculum label correcting to enhance classification accuracy, even under varying levels of noise. The framework demonstrates superior performance across multiple medical image datasets, showcasing its effectiveness in improving the robustness of deep learning models in noisy environments. In (Ju *et al*., 2022), the authors explore the challenges of label noise in medical image datasets, specifically focusing on two types: disagreement label noise from inconsistent expert opinions and single-target label noise from biased aggregation. They propose an uncertainty estimation-based framework to manage these issues and introduce a boosting-based curriculum training approach for robust learning. Their method is validated across various medical datasets, demonstrating its effectiveness in handling noise and improving classification performance.

In medical image classification, dealing with noisy data is a significant challenge, as noise can corrupt crucial features and degrade model performance. To address this, several robust architectures have been proposed, specifically designed to handle noisy datasets more effectively. These architectures integrate innovative methods to filter out noise, enhance feature extraction, and maintain classification accuracy, even in the presence of substantial data corruption. In (Zhu *et al*., 2021), the authors present a noise-robust learning method for histopathology image classification. Their approach introduces a novel easy/hard/noisy (EHN) detection model that distinguishes between informative hard samples and harmful noisy ones based on training history. By integrating this model into self-training architecture, they gradually correct noisy labels and suppress noise during training. This method effectively handles noisy labels without relying on a clean subset, making it suitable for real-world noisy datasets. In (Xue *et al*., 2022), the authors tackle the issue of noisy-labeled data in medical image classification by introducing a collaborative training paradigm. Their method combines global and local representation learning to improve robustness against noisy labels. A self-ensemble model with a noisy label filter is employed to differentiate between clean and noisy samples, while a collaborative training strategy ensures that imperfect labels do not compromise the model's performance. This approach is particularly relevant to addressing the challenges posed by label noise in medical image classification, which aligns with our focus on developing noise-robust architectures. In (Li *et al*., 2021), the authors tackle the challenge of noise sensitivity in convolutional neural networks (CNNs) by incorporating wavelet transforms. Their proposed method, WaveCNets, replaces traditional down-sampling techniques like max-pooling with discrete wavelet transforms (DWT). This allows the model to separate the important low-frequency features, which carry key information about the object, from the high-frequency components, which often contain noise. By discarding the high-frequency components, WaveCNets improve noise robustness and classification accuracy, particularly in noisy image datasets.

In our exploration of denoising techniques, we delve into several approaches that have shown promise in enhancing image quality. Each method plays a distinct role in preserving essential features while mitigating noise, ultimately contributing to improved classification outcomes. By examining these established techniques, we can better understand their efficacy and applicability in the context of medical image processing. Several studies have explored the application of edge-preserving filters in enhancing image quality for better classification outcomes. In (Yang *et al*., 2021), the authors introduce a global edge-preserving filter based on soft clustering, utilizing a restricted Gaussian mixture model to enhance image quality. Their approach effectively suppresses intensity shift artifacts and handles halo effects common in local filters. The proposed method allows for flexible control over smoothing levels, while maintaining low computational complexity. In (Wang *et al*., 2020), the authors address the fundamental problem of image denoising, emphasizing the need to

preserve significant geometric features, such as edges and textures, while filtering out noise. They introduce an edge detection function based on the Gaussian filtering operator and analyze the characteristics of the fractional derivative operator. Building on this, they establish a spatially adaptive fractional edge-preserving denoising model within a variational framework, discussing the existence and uniqueness of its solution and deriving the nonlinear fractional Euler-Lagrange equation. This approach represents a fractional order extension of traditional variational methods. In (Zhou *et al*., 2020), the authors tackle the issue of preserving edge structures in salient object detection, a crucial preprocessing step in various computer vision tasks. They introduce the Hierarchical and Interactive Refinement Network (HIRN), which aims to counteract the blurring effects of downsampling operations, such as pooling and striding, on edge detection. The proposed network features a multistage and dual-path structure that estimates salient edges and regions from both low-level and high-level feature maps. This approach enhances the accuracy of predicted regions by improving weak edge responses while refining the semantic quality of edge predictions. Additionally, they present an edge-guided inference algorithm to further refine the output regions based on the predicted edges.

The Gaussian filter is widely utilized in image denoising due to its effectiveness in reducing random noise while preserving essential features like edges and textures. Its ability to apply a weighted average based on the Gaussian distribution allows for smoothening while minimizing distortion, making it a fundamental tool in various image processing applications. In (Zhu and Ng, 2020), the authors address the challenge of mixed noise denoising, specifically focusing on images affected by both Gaussian and impulse noise. They propose two structured dictionary learning models that combine fidelity and regularization terms to recover corrupted images. By employing proximal alternating minimization methods, the study emphasizes the importance of accurately fitting image patches while utilizing sparse coding to effectively mitigate the impact of noise, highlighting a significant advancement in the field of image restoration.

Median filtering is a widely used technique for image denoising that effectively preserves edges while removing noise, making it particularly suitable for applications where maintaining important features is crucial. In (Taassori and Vizvári, 2024), a novel hybrid approach is presented, combining multiple noise reduction strategies to enhance the quality of medical images. This approach begins with an adaptive Kalman filter for initial noise attenuation and is followed by post-processing steps that include a non-local means (NLM) method and a median filter. The median filter plays a critical role in further refining the denoised images by effectively suppressing residual noise while maintaining the integrity of important diagnostic features. Weighted median (WM) filters are often employed for tasks requiring enhanced noise suppression while preserving edges. In (Mishiba, 2023), the authors propose an efficient real-time WM filter that avoids traditional histogram construction challenges, achieving high-quality denoising performance.

CLAHE (Contrast Limited Adaptive Histogram Equalization) is widely used in image processing to enhance contrast, particularly in medical images. It is effective in improving the visibility of subtle features by adjusting the contrast locally in different regions of an image, making it an essential tool for images with varying lighting conditions. This method is particularly beneficial for enhancing medical images, where precise details are critical for accurate diagnoses. In (Chang *et al*., 2018), the authors propose an automatic contrast-limited adaptive histogram equalization (CLAHE) method for image contrast enhancement. This approach automatically sets the clip point based on the textureness of image blocks and incorporates dual gamma correction to enhance contrast while maintaining naturalness. The method effectively redistributes histograms in each block, enhancing luminance, particularly in dark areas, while minimizing over-enhancement artifacts. Automatic contrast-limited adaptive histogram equalization (CLAHE) is a widely used technique for enhancing image contrast. It works by adjusting the contrast in local regions of an image, which helps improve visibility, especially in low-light conditions. Recent studies have explored various enhancements to CLAHE to further optimize its performance in challenging environments, including nighttime settings where visibility is critical for applications such as autonomous driving (Chen *et al*., 2023).

In image processing, sharpening is a crucial technique aimed at enhancing the clarity and detail of images. By increasing the contrast between adjacent pixels, sharpening helps to bring out important features, making images appear more defined and focused. This process is particularly beneficial in medical imaging, where precise details can significantly impact diagnosis and analysis. Various methods exist for sharpening, including unsharp masking and high-pass filtering, each offering unique advantages for different applications. In (Ye & Ma, 2018), the authors propose a highly adaptive unsharp masking method known as blurriness-guided

unsharp masking (BUM). This method utilizes estimated local blurriness to perform pixel-wise enhancement, addressing the challenges of over-enhancement in sharp regions and noise enhancement in blurred areas. The enhancement strength is adjusted based on the local blurriness, and a mapping process generates a scaling matrix from the blurriness map. Additionally, the study emphasizes the importance of the layer-decomposition filter used for creating base and detail layers, focusing on preventing artifacts through the choice between edge-preserving and non-edge-preserving filters. In addition to denoising techniques, image sharpening plays a critical role in enhancing visual details. Unsharp masking (UM) has been widely adopted for this purpose. Recent studies have introduced new approaches to enhance sharpening performance while avoiding issues such as over- or under-enhancement. For instance, combining unsharp masking with histogram equalization has shown improved control over enhancement levels, maximizing image information and ensuring a balance in brightness distribution (Kansal, Purwar & Tripathi, 2018).

Transformers have revolutionized the field of machine learning by introducing a novel approach to handling sequential data, such as images and text, using self-attention mechanisms. Unlike traditional models that process data in a linear or hierarchical manner, transformers can simultaneously capture relationships between different parts of an input sequence, regardless of their distance. The attention mechanism plays a pivotal role in this, allowing the model to focus on relevant parts of the input when making predictions, significantly improving performance on tasks like classification, especially when dealing with complex patterns, such as in medical imaging.

In (Vaswani, 2017), the authors propose the Transformer architecture, which leverages self-attention mechanisms and eliminates the need for recurrence or convolution. By eliminating recurrence and convolution, the Transformer architecture allows for more parallelization and significantly faster training. This architecture leverages self-attention to capture global dependencies in sequences, making it particularly effective in various sequence processing tasks, such as machine translation. Its ability to handle long-range dependencies efficiently has made it a foundation in deep learning, especially for tasks requiring high performance and scalability. In (Dosovitskiy, 2020), the authors introduce Vision Transformers (ViTs) as a novel approach to image recognition, eliminating the need for convolutional neural networks (CNNs). Instead, they apply transformers directly to sequences of image patches, demonstrating their effectiveness in image classification tasks. This marks a shift from traditional CNN-based methods to a more flexible architecture, highlighting the potential of transformers in the domain of computer vision. The work emphasizes that transformers can be applied successfully to image-based tasks, paving the way for further advancements in vision-related applications.

In this work, the architectural foundation is built upon the Data-efficient Image Transformer (DeiT), which is well-suited for image classification tasks. DeiT offers an efficient approach to Vision Transformer (ViT) models by optimizing training processes and reducing the dependence on large datasets, making it an appropriate choice for medical image analysis. Its lightweight design and robust performance allow for improved generalization on noisy datasets, which aligns with the challenges addressed in this study. The reliability of deep learning models in medical diagnosis remains a significant concern, particularly considering potential adversarial attacks that could lead to severe consequences. To address these challenges, recent studies have explored hybrid architectures that combine the strengths of Convolutional Neural Networks (CNNs) and Transformers. For instance, one approach proposes a robust CNN-Transformer hybrid model, which leverages the locality of CNNs alongside the global connectivity of Vision Transformers. This model enhances computational efficiency through an optimized attention mechanism and aims to learn smoother decision boundaries (Manzari *et al*., 2023). Recent advancements in generative adversarial models have demonstrated the effectiveness of convolutional neural networks (CNNs) in medical image synthesis tasks. However, CNNs' local processing capabilities can hinder the learning of contextual features. To address this, researchers in (Dalmaz *et al*., 2022) propose the novel approach ResViT, which combines the contextual sensitivity of vision transformers with the precision of convolutional operators. Its generator incorporates aggregated residual transformer (ART) blocks, promoting diversity in representations while distilling relevant information. The model also includes a weight sharing strategy to reduce computational load, making it versatile for various modality. In (Song *et al*., 2023), the authors address the limitations of convolutional neural network-based methods in image dehazing by proposing DehazeFormer. They highlight that the popular Swin Transformer has key designs unsuitable for dehazing tasks. To improve performance, DehazeFormer incorporates modifications such as a new normalization layer, adjusted activation functions, and enhanced spatial information aggregation. Recent advancements in medical imaging have increasingly leveraged Vision Transformers (ViTs) due to their ability to capture long-range dependencies and global context, which offer clear

advantages over traditional convolutional neural networks (CNNs). The review by (Azad *et al*., 2023) presents an encyclopedic examination of the applications of Transformers in medical imaging, covering tasks such as classification, segmentation, and detection while highlighting the strengths and weaknesses of various strategies. Additionally, (Shamshad *et al*., 2023) provides a comprehensive survey of Transformer architectures and their applications in medical imaging, discussing key challenges and promising future directions. These studies underscore the transformative potential of ViTs in enhancing medical image analysis, paving the way for further exploration in this domain. In (Touvron *et al*., 2021), the authors introduce an innovative framework for training data-efficient vision transformers without the need for extensive datasets or significant computational resources. They propose a teacher-student strategy specifically designed for transformers, employing a distillation token to enhance the learning process. This approach not only facilitates effective knowledge transfer but also demonstrates the potential of vision transformers in various image understanding tasks.

Despite significant advancements in medical image classification, the presence of noise in real-world datasets remains a major challenge that undermines the reliability and accuracy of classification models. While much of the existing research focuses on improving classification performance under ideal conditions, few methods effectively address the impact of noise on medical images. This highlights the critical importance of our work, which introduces a robust, noise-resilient model based on the DeiT framework. By incorporating advanced preprocessing techniques and leveraging the strengths of transformers, our approach enhances both the performance and reliability of medical image classification systems. This work has the potential to substantially improve clinical decision-making, enabling more accurate and dependable diagnoses in noisy, real-world environments.

Challenges in Noisy Medical Image Classification

Noisy medical image classification presents significant challenges, as noise obscures critical diagnostic details, leading to reduced classifier accuracy and reliability. The loss of essential information, such as blurred edges or obscured lesions, makes it difficult for models to distinguish between healthy and affected tissues, severely compromising diagnostic outcomes. Noise-induced misclassification further exacerbates the issue, where distorted features can cause incorrect predictions, increasing the risk of wrong diagnoses. Additionally, data imbalance, including the rarity of certain conditions, compounds the problem, as limited examples of underrepresented classes hinder a classifier's ability to learn distinguishing features. Noise, along with data collection biases and technological limitations, further degrades performance, as inadequate imaging quality and computational constraints limit the efficacy of classification algorithms. Addressing these challenges is essential for developing robust classifiers capable of reliable clinical applications.

METHODS

We realized that noisy datasets, particularly in medical image classification, introduce significant challenges that impact classifier performance. The presence of noise can obscure important features within the images, leading to misclassifications and reduced diagnostic accuracy. This issue becomes especially critical in medical applications, where precise classification is essential for patient outcomes. Understanding these challenges led to the creation of a robust architecture designed to handle the complexities of noisy datasets.

Edge Preserving Filter

Edge-preserving filters smooth images while preserving important features like edges and textures. These filters reduce noise without blurring edges, which is crucial for applications like medical image processing, where structural details are vital for accurate diagnosis. This makes them valuable for enhancing noisy medical images while maintaining essential boundaries. Our approach uses a recursive filter as an edge-preserving method, chosen for its effectiveness in improving image quality and preserving key features for noisy medical image classification. The Edge-Preserving Filter algorithm is outlined in Algorithm 1, which details the process of noise reduction while maintaining edge integrity.

Algorithm 1 Edge-Preserving Filter

1: **Input:** Image (in three R, G, and B channels), sigma_s (spatial smoothing parameter), sigma_r (range smoothing parameter)
2: **Output:** edge_preserved image
3: **for** channel in [R, G, B]:
4: **Initialize** filtered_image with the original channel values
5: # Horizontal Pass (Left to Right)
6: **for** each row in channel:
7: **Initialize** previous_filtered_value to the first pixel value in the row.
8: **for** each pixel from left to right:
9: gradient = abs (current_pixel - previous_pixel)
10: weight = exp (-gradient / sigma_r)
11: filtered_pixel = weight × current_pixel + (1 - weight) × previous_filtered_value
12: update previous_filtered_value = filtered_pixel
13: # Horizontal Pass (Right to left)
14: **for** each row in channel:
15: **Initialize** previous_filtered_value to the last pixel value in the row.
16: **for** each pixel from right to left:
17: gradient = abs (current_pixel - previous_pixel)
18: weight = exp (-gradient / sigma_r)
19: filtered_pixel = weight × current_pixel + (1 - weight) × previous_filtered_value
20: update previous_filtered_value = filtered_pixel
29: # Vertical Pass (top to bottom)
30: **for** each row in channel:
31: **Initialize** previous_filtered_value to the first pixel value in the col.
32: **for** each pixel from top to bottom:
33: gradient = abs (current_pixel - previous_pixel)
34: weight = exp (-gradient / sigma_r)
35: filtered_pixel = weight × current_pixel + (1 - weight) × previous_filtered_value
36: update previous_filtered_value = filtered_pixel
37: # Vertical Pass (Bottom to Top)
38: **for** each row in channel:
39: **Initialize** previous_filtered_value to the last pixel value in the col.
40: **for** each pixel from bottom to top:
41: gradient = abs (current_pixel - previous_pixel)
42: weight = exp (-gradient / sigma_r)
43: filtered_pixel = weight × current_pixel + (1 - weight) × previous_filtered_value
44: update previous_filtered_value = filtered_pixel
45: Combined filtered_image channels (R, G, B) to produce final image

Gaussian Filter

Combining multiple denoising methods can significantly enhance the overall quality of medical images by leveraging the strengths of each technique. For instance, using an edge-preserving filter in conjunction with a Gaussian filter allows for effective noise reduction while maintaining important structural details. This hybrid approach minimizes the risk of losing critical features during the denoising process, resulting in images that are clearer and more informative. Furthermore, combining

methods can lead to improved robustness against various types of noise, ultimately facilitating more accurate analyses and classifications in medical imaging tasks.

Following the application of the edge-preserving filter, a Gaussian filter is utilized to further enhance the quality of the denoised medical images. The Gaussian filter is widely recognized for its effectiveness in reducing noise while preserving essential image features. In medical image processing, noise reduction is critical for improving the visibility of important structures and details within the images. By applying the Gaussian filter, high-frequency noise is effectively mitigated, resulting in a smoother image that retains critical information necessary for accurate analysis. This enhancement is particularly beneficial for the subsequent classification processes, where clear and well-defined features are essential for reliable outcomes. Moreover, the Gaussian filter provides a controlled method for smoothing images, as the standard deviation parameter allows for adjustments in the degree of blurring. This flexibility enables optimization based on the specific characteristics of the medical images being analyzed, ensuring that important details remain discernible while unwanted noise is minimized. Overall, the application of the Gaussian filter significantly contributes to improving the quality of medical images, thereby facilitating more accurate analyses and enhancing the performance of classification algorithms. In Algorithm 2, the process of generating the kernel for the Gaussian filter is detailed. Algorithm 3 presents the Gaussian filter, highlighting its key steps and parameters for spatial smoothing and noise reduction.

Algorithm 2 Generating Filter Kernel

Input: Kernel Size k, Standard Deviation (σ)
Output: Kernel (k × k)
center = (k - 1) / 2
sum = 0
for x in (0, k):
  for y in range (0, k):
    dx = x - center
    dy = y - center
    kernel[x][y] $= \frac{1}{2\pi\sigma^2} e^{-\frac{dx^2+dy^2}{2\sigma^2}}$
    sum += kernel[x][y]
normalize the kernel
for x in (0, k):
  for y in range (0, k):
    Kernel[x][y] /= sum
return kernel (k × k)

Algorithm 3 Gaussian Filter

Input: Image (rows × cols), Gaussian Kernel (k × k)
Output: Filterred_Image (rows × cols)
Radius = k × k/2
for each pixel (i,j) in Image :
    Sum = 0
    for x in (-Radius, Radius+1):
      for y in (-Radius, Radius+1):
        if (i+x < rows and j+y < cols):
          Sum += image[i+x][j+y] × kernel[x+Radius][y+Radius]
    Filtered_image[i][j] = sum
return Filtered_image

Median Filter

The median filter is a widely used non-linear filter in image processing. Instead of averaging pixel values, it replaces each pixel with the median value of its surrounding pixels, ensuring that the noise is minimized while preserving important edges and details. This makes the median filter especially useful in scenarios where maintaining edge sharpness is crucial, as it efficiently removes noise without excessively blurring the image.

The combination of denoising methods, such as edge-preserving, Gaussian, and median filters, leverages the strengths of each technique. This approach improves noise reduction while maintaining critical features like edges and structural details. The sequential application of these filters provides enhanced robustness against various noise types, better feature preservation, and ultimately results in clearer images, which is essential for accurate analysis. The algorithm for the Median Filter is illustrated in Algorithm 4. This diagram outlines the procedure and parameters used to apply the filter, which effectively reduces Gaussian noise while preserving the image's structural details. The process systematically calculates the median value within a defined kernel to achieve noise reduction.

Algorithm 4 Median Filter

```
Input: Image (rows × cols), Kernel size (k × k)
Output: Filterred_Image (rows × cols)
Radius = k × k/2
for each pixel (i,j) in Image :
        # an empty to store neighbors
        neighbors = []
        for x in (-Radius, Radius+1):
            for y in (-Radius, Radius+1):
                if (i+x < rows and j+y < cols):
                    Neighbots.append (image [i+x][j+y])
        # sort neighbors and find the median
        Filtered_image [i][j] = neighbors.sort.median
return Filtered_image
```

Advantages of Combining Denoising Method

In image processing, no single denoising method effectively removes all types of noise while preserving crucial image details. Combining multiple techniques leverages their strengths, resulting in a more balanced and effective noise reduction approach (Taassori and Vizvári, 2024), (Taassori, 2024). This synergy enhances noise removal by targeting various noise types, such as impulse noise with median filtering and high-frequency noise with Gaussian filtering. Additionally, it helps preserve important image features, preventing excessive blurring while retaining edges and textures. By increasing robustness, combined methods adapt better to diverse noise conditions, ensuring consistent image quality. Furthermore, this approach enhances versatility, making it suitable for various image types and tasks.

Contrast Limited Adaptive Histogram Equalization (CLAHE)

Contrast Limited Adaptive Histogram Equalization (CLAHE) is a powerful technique for improving image contrast, particularly in images with varying lighting conditions. Unlike traditional histogram equalization, which can produce over-enhancement and amplify noise, CLAHE operates on small, localized regions of the image, known as tiles. This localized approach allows for better contrast enhancement while preserving important details and avoiding the introduction of artifacts.

CLAHE enhances contrast by applying a contrast-limiting algorithm to prevent over-amplification of specific intensity levels, ensuring uniform enhancement without introducing artifacts. This is especially useful for medical image classification, as it helps highlight

subtle features critical for accurate diagnosis, improving classification model performance.

The CLAHE process begins with a denoised input image, which is converted to LAB color space to separate the luminance (L) channel from the color channels (A and B). Contrast is enhanced independently in the L channel by dividing it into small tiles, computing histograms for each, and applying a clip limit to prevent noise amplification. The enhanced tiles are merged into a single luminance channel, which is combined with the original A and B channels and converted back to RGB, preserving natural color while improving contrast. CLAHE enhances local contrast, making it a valuable tool for medical imaging. The CLAHE algorithm is detailed in Algorithm 5, outlining the steps for contrast enhancement while avoiding noise over-amplification.

Algorithm 5 Contrast Limited Adaptive Histogram Equalization

1: **Input:** Image, Clip_Limit, Tile_Grid_Size
2: **Output:** Color_Corrected_Image
3: **if** Image is in RGB format:
4: Convert Image to LAB color space
5: Split the LAB image into L, A, B channels
6: L, A, B = SplitChannels(LAB_Image)
7: **Initialize:** Clip_Limit and Tile_Grid_Size for CLAHE
8: **for** each tile in the L channel:
9: Compute the histogram of pixel intensities in the tile
10: **if** histogram exceeds Clip_Limit:
11: Redistribute the excess pixels evenly across bins
12: Equalize the histogram of the tile
13: Recombine CLAHE-enhanced L channel with A and B to form LAB_Image
14: LAB_Image = MergeChannels(L_CLAHE, A, B)
15: Convert LAB_Image back to RGB color space
16: Color_Corrected_Image = ConvertToRGB(LAB_Image)
17: **return** Color_Corrected_Image

Sharpening

This contribution focuses on enhancing the clarity and detail of the image, which is essential for accurate interpretation in medical image classification tasks. By refining image features and emphasizing important details, this step significantly improves the quality of the processed images. Sharpening enhances the visibility of fine details, making them more prominent and aiding in the interpretation and analysis of medical images. It also improves edge clarity, highlighting critical structures and contributing to more precise diagnoses. Additionally, the sharpening process is designed to enhance important features without amplifying noise, ensuring the final image maintains high integrity. This contribution plays a key role in preparing medical images for further analysis and classification, resulting in better diagnostic outcomes.

In our proposed method, sharpening was performed using the unsharp mask technique. First, we generated a Gaussian Blur of the input image using the Gaussian filter. The blurring process involves convolving the image with a Gaussian kernel defined by the formula:

$$Gaussian\_Blur(x,y) = \sum_{u=-K}^{k}\sum_{v=-k}^{k}\frac{1}{2\pi\sigma^2}e^{-\frac{u^2+v^2}{2\sigma^2}} \cdot Input\_Image(x-u,y-v) \quad (1)$$

The kernel size is controlled by K, and the parameter σ determines the spread of the Gaussian filter, which influences the amount of blurring applied to the image.

Once the Gaussian Blur is computed, the sharpened image is obtained by emphasizing the difference

between the input image and the Gaussian Blur**,** which can be expressed using the following formula:

$$\begin{aligned} Sharpened\_Image[i,j] \\ = (1 + Weight) \\ * Input\_Image[i,j] \\ - (Weight \\ * Gaussian\_Blur[i,j]) \end{aligned} \quad (2)$$

The weight parameter controls the strength of the sharpening effect, with higher values of weight increasing the enhancement of edges and fine details. This approach, commonly known as unsharp masking, emphasizes high-frequency components by amplifying the difference between the original and blurred images, resulting in a sharper appearance.

Unsharp masking offers several benefits for medical imaging, particularly in edge enhancement. It sharpens edges for clearer distinction between adjacent areas without over-enhancing non-edge regions, ensuring accurate fine details. The scaling factor can be adjusted for controlled sharpness, meeting specific enhancement needs. Additionally, it allows selective noise management, enhancing important details while minimizing noise to preserve critical information. Customizable parameters reduce artifacts, enabling practitioners to fine-tune the effect and maintain essential image characteristics while improving detail visibility. Algorithm 6 presents the sharpening process, which utilizes the kernel generated in Algorithm 2.

Algorithm 6 Sharpening

1: **Input:** Image (row × col), Kernel_Size, Sigma, Sharpening_Weight
2: **Output:** Sharpened_Image
3: **Compute a blurred version of the image**
4: **for** each pixel (i,j) in Image (row × col):
5: **for** i in (0, row):
6: **for** j in (0, col):
7: Sum = 0
8: **for** x in (0,k) do # kernel (k × k):
9: **for** y in (0,k):
10: **if** (0 <= i+x-k/2 < row and 0< j+y-k/2 < col):
11: Sum += image [i+x-k/2][j+y-k/2] × kernel[x][y]
12: Gaussian_blur[i,j] = sum
13: **Calculate the sharpened image**
14: for each pixel (i, j) in Image (row × col):
15: Compute the weighted combination of original and blurred images:
16: Sharpened_Image[i, j] = (1 + Weight) × Image[i, j] − (Weight × Gaussian_Blur[i, j])
17: **if** Sharpened_Image values exceed allowed range:
18: Clip values to valid intensity range (e.g., 0 to 255)
19: **return** Sharpened_Image

Normalization

Normalization is a critical step in the image processing pipeline that ensures the pixel intensity values are scaled to a specified range, enhancing the overall visual quality of the image. In this method, the normalized image is adjusted to fit within the range of 0 to 255, which is standard for 8-bit images. In the proposed method, normalization is performed to adjust the pixel intensity values of the sharpened image. The normalization process can be mathematically represented by the following formula:

$$I_{norm}(x,y) = \frac{I(x,y) - I_{min}}{I_{max} - I_{min}} \times (New_{max} - New_{min}) + New_{min} \quad (3)$$

Where $I_{norm}(x,y)$ is the normalized pixel value at position (x,y), $I(x,y)$ is the original pixel value at position (x,y), I_{min} is the minimum pixel value in the

image, I_{max} is the maximum pixel value in the image, New_{max} and New_{min} define the desired output range, typically 0 and 255 for 8-bit images.

Normalization in the proposed approach enhances contrast, improves feature visibility, and optimizes pixel value utilization, benefiting low-contrast images. It also standardizes images for consistency, facilitating effective processing and analysis. This step is crucial for improving classification performance. The normalization process, detailed in Algorithm 7, scales pixel values to a specified range, ensuring uniform intensity levels and better visual quality.

Algorithm 7 Normalization

1: **Input:** Image

2: **Output:** Normalized_Image

3: **Initialize** input image (Image)

4: Set the **target minimum intensity value** as New_Min_Value

5: Set the **target maximum intensity value** as New_Max_Value

6: Apply normalization

7: Min_Value and Max_Value are the minimum and maximum pixel values in the image, while New_Min_Value and New_Max_Value define the target intensity range

8: **for** pixel (i,j) in Image:

9: Normalize the pixel value to the target range [New_Min_Value, New_Max_Value] using the formula:

$$\text{Normalized\_Image}[\text{i},\text{j}] = \frac{(\text{Image}[\text{i},\text{j}] - \text{Min\_Value})}{\text{Max\_Value} - \text{Min\_Value}} \times (\text{New\_Max\_Value} - \text{New\_Min\_Value}) + \text{New\_Min\_Value}$$

10: **return** Normalized_Image

Fine-Tuning Vision Transformer for Medical Image Classification

Unlike traditional convolutional neural networks (CNNs), which extract features hierarchically, ViT treats images as sequences of patches, using self-attention mechanisms to capture global relationships. This approach is particularly effective in medical image classification, where subtle, dispersed patterns are critical for diagnosis. The self-attention mechanism enables ViT to focus on important areas, improving performance even in complex images with small or varying textures.

ViTs also offer scalability, efficiently handling large datasets and adapting to growing data and computational resources. Their transfer learning capabilities, where models pretrained on large datasets are fine-tuned for specific medical tasks, further enhance their effectiveness, even with limited labeled data. Additionally, ViTs generate attention maps, improving interpretability and supporting clinical decision-making, which builds trust among healthcare professionals.

In ViT, images are divided into n patches, represented as vectors in the input matrix, which the self-attention mechanism processes to capture complex relationships within the visual data.

$$X = [x_1, x_2, \dots, x_n] \in R^{n \times d} \tag{4}$$

where n is the number of patches and d is the dimension of each patch representation. The model learns three separate weight matrices, W^Q, W^K, and W^V, to generate the query Q, key K, and value V matrices from X as follows:

$$Q = XW^Q, K = XW^K, V = XW^V \tag{5}$$

The attention scores are computed using the dot product of the query and key matrices, scaled by the square root of the key dimension $\sqrt{d_k}$, followed by applying the softmax function:

$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{6}$$

The output of this self-attention mechanism combines the value vectors weighted by the attention scores, enabling the model to selectively focus on the most relevant parts of the input. This capability enhances the ViT's ability to understand and interpret visual information effectively, making it a powerful tool for various image classification tasks.

Data-Efficient Image Transformers (DeiT)

The Data-Efficient Image Transformer (DeiT) is an advanced Vision Transformer (ViT) designed to optimize training while minimizing the need for large datasets. Unlike traditional ViTs, which require extensive data for effective training, DeiT achieves competitive performance with fewer samples, making it ideal for scenarios with limited labeled data.

A key innovation in DeiT is the teacher-student distillation framework, where the model benefits from the guidance of a more powerful model, typically a CNN, acting as the teacher. This mechanism improves knowledge transfer, allowing DeiT to learn robust features and generalize well, even with limited data. This is particularly valuable in medical imaging, where labeled data is scarce and costly to obtain.

In medical image classification, DeiT improves learning from smaller datasets, optimizing the transformer architecture without the need for extensive data. This is particularly useful for medical tasks with limited labeled data. Fine-tuning a pretrained DeiT model enhances efficiency, allowing the model to extract meaningful patterns while maintaining accuracy and reducing data requirements.

DeiT is designed to optimize training with smaller datasets, addressing challenges in medical image classification. It excels with limited data, benefiting from knowledge distillation, where a CNN guides the model's learning, improving generalization and performance. Additionally, DeiT requires fewer computational resources and less time, making it highly efficient for medical image tasks with scarce labeled data.

Fig. 1 provides an overview of the proposed RobustDeiT approach, illustrating each stage, from preprocessing (edge-preserving filtering and Gaussian blurring) to sharpening and normalization. This workflow demonstrates how the integrated operations enhance image quality, improving classification accuracy in noisy datasets.

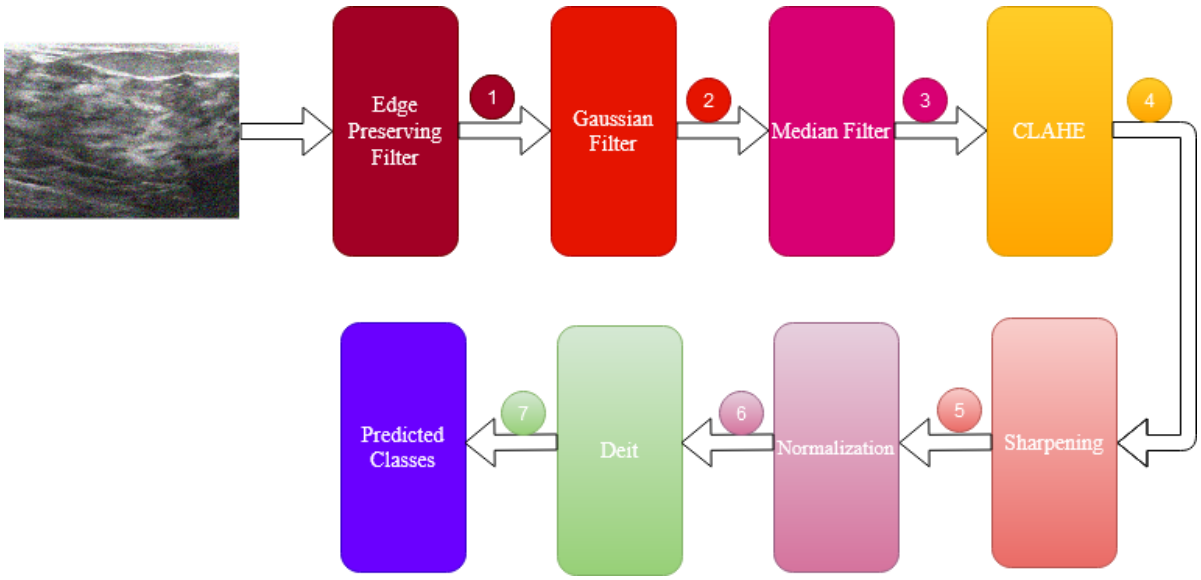

Fig. 1: *Pipeline of RobustDeiT*

RESULTS

This study aims to develop a robust classification framework for medical images within noisy datasets, enhancing diagnostic accuracy and reliability. Using a Data-efficient Image Transformer (DeiT) and a multi-stage preprocessing pipeline, the proposed method improves image quality by reducing noise, enhancing contrast, and standardizing intensity, thereby facilitating accurate feature extraction and classification in challenging clinical environments. The dataset includes breast ultrasound images collected in 2018 from 600 female patients between the ages of 25 and 75. It consists of 780 images with an average resolution of 500 × 500

pixels, stored in PNG format. The images are categorized into three classes: 487 labeled as benign, 210 as malignant, and 133 as normal (Al-Dhabyani *et al*., 2020).

Gaussian noise is applied to the dataset at standard deviations of 15, 25, 35, and 45 to simulate realistic imaging conditions. By incorporating Gaussian noise, we evaluate the robustness of the proposed classification framework under different noise levels, reflecting challenges faced in practical clinical settings.

The dataset is split into three parts: a training set, a validation set, and a test set. The training set is used to train the model, while the validation set is used to tune the model's hyperparameters and assess its performance during training. Finally, the test set is reserved for evaluating the model's performance on unseen data, ensuring that the results are reliable and generalizable.

The output of each stage of the proposed method is shown in Fig. 2, illustrating the sequential transformations applied to noisy images. The pipeline includes edge-preserving filtering, Gaussian blur, median filtering, CLAHE for contrast enhancement, and sharpening. Each stage contributes uniquely to image refinement: the initial stages (edge-preserving filtering, Gaussian blur, and median filtering) effectively suppress noise while maintaining essential structural details, minimizing the risk of over-smoothing. In contrast, the later stages (CLAHE and sharpening) enhance contrast and highlight finer details, ensuring key features become more prominent. The figure underscores the cumulative benefits of this sequential pipeline, revealing how each stage complements the others. It also demonstrates the potential drawbacks of skipping stages, such as amplified noise or diminished feature clarity. These results emphasize the necessity of applying all stages collectively for optimal image preparation and improved classification outcomes.

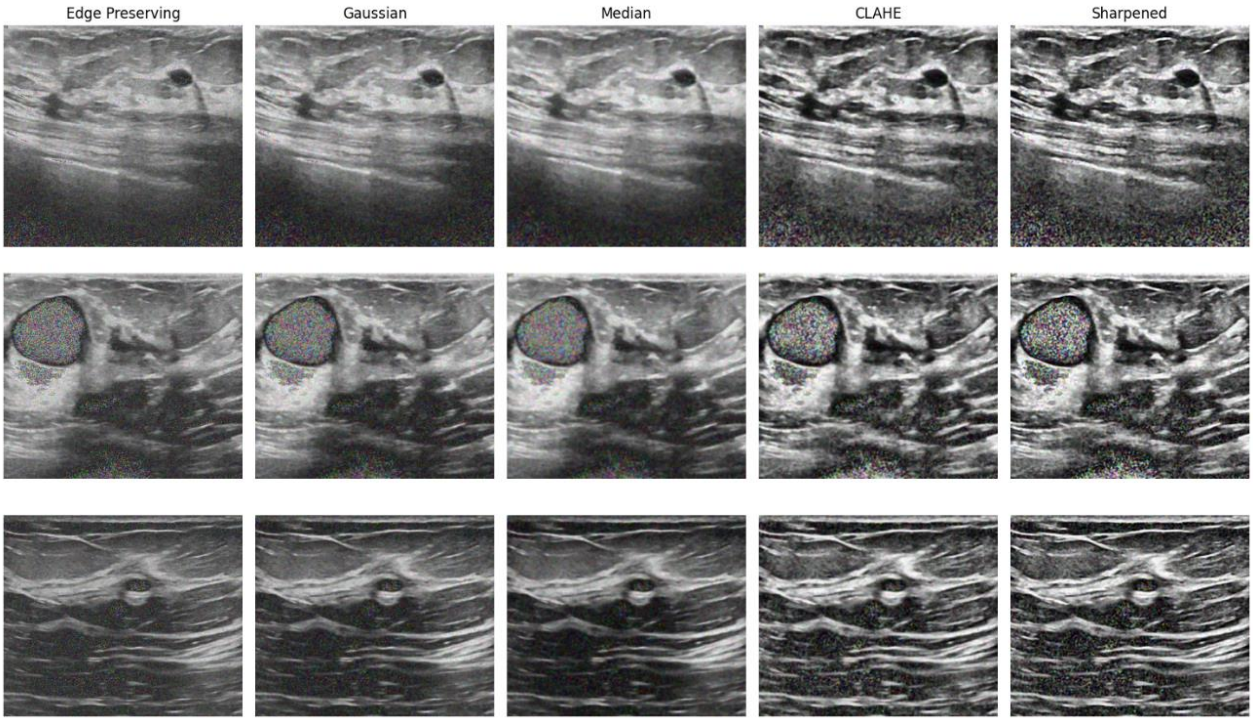

Fig. 2: *Progressive Image Refinement through Sequential Processing Stages*

To evaluate the classification model's performance, we employ four essential metrics: accuracy, precision, recall, and F1 score. Their corresponding formulas are presented in Equations (7), (8), (9), and (10), respectively.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

$$Recall = \frac{TP}{TP + FN} \quad (9)$$

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (10)$$

Where TP (True Positives) represents the number of correctly classified positive instances, TN (True Negatives) denotes the correctly classified negative instances, FP (False Positives) refers to negative instances mistakenly predicted as positive, and FN (False Negatives) corresponds to positive instances incorrectly predicted as negative.

Table 1: *Model Performance Metrics with Gaussian Noise ($\sigma = 15$)*

Metric		VGGNet	GoogLeNet	ResNet	RobustDeiT
Accuracy		0.72	0.74	0.79	0.85
Precision	Benign	0.69	0.73	0.84	0.88
	Malignant	0.74	0.81	0.72	0.76
	Normal	1.00	0.73	0.78	0.95
Recall	Benign	0.92	0.89	0.80	0.86
	Malignant	0.45	0.55	0.84	0.81
	Normal	0.45	0.55	0.70	0.90
F1-Score	Benign	0.79	0.80	0.82	0.87
	Malignant	0.56	0.65	0.78	0.78
	Normal	0.62	0.63	0.74	0.92

Table 2: *Improvement of RobustDeiT against Baseline Architectures ($\sigma = 15$)*

Metric		VGGNet (%)	GoogLeNet (%)	ResNet (%)
Accuracy		18.06	14.86	7.59
Precision	Benign	27.54	20.55	4.76
	Malignant	2.70	-6.17	5.56
	Normal	-5.00	30.14	21.79
Recall	Benign	-6.52	-3.37	7.50
	Malignant	80.00	47.27	-3.57
	Normal	100.00	63.64	28.57
F1-Score	Benign	10.13	8.75	6.10
	Malignant	39.29	20.00	0.00
	Normal	48.39	46.03	24.32

Table 3: *Model Performance Metrics with Gaussian Noise ($\sigma = 25$)*

Metric		VGGNet	GoogLeNet	ResNet	RobustDeiT
Accuracy		0.79	0.72	0.78	0.86
Precision	Benign	0.76	0.72	0.78	0.86
	Malignant	0.89	0.69	0.86	0.86
	Normal	0.78	0.77	0.70	0.89
Recall	Benign	0.92	0.85	0.89	0.91
	Malignant	0.55	0.58	0.58	0.77
	Normal	0.70	0.50	0.70	0.85
F1-Score	Benign	0.84	0.78	0.83	0.88
	Malignant	0.68	0.63	0.69	0.81
	Normal	0.74	0.61	0.70	0.87

Table 4: *Improvement of RobustDeiT against Baseline Architectures (σ = 25)*

Metric		VGGNet (%)	GoogLeNet (%)	ResNet (%)
Accuracy		8.86	19.44	10.26
Precision	Benign	13.16	19.44	10.26
	Malignant	-3.37	24.64	0.00
	Normal	14.10	15.58	27.14
Recall	Benign	-1.09	7.06	2.25
	Malignant	40.00	32.76	32.76
	Normal	21.43	70.00	21.43
F1-Score	Benign	4.76	12.82	6.02
	Malignant	19.12	28.57	17.39
	Normal	17.57	42.62	24.29

Table 5: *Model Performance Metrics with Gaussian Noise (σ = 35)*

Metric		VGGNet	GoogLeNet	ResNet	RobustDeiT
Accuracy		0.68	0.73	0.77	0.83
Precision	Benign	0.72	0.72	0.80	0.84
	Malignant	0.56	0.88	0.85	0.77
	Normal	0.65	0.60	0.63	0.89
Recall	Benign	0.76	0.88	0.85	0.86
	Malignant	0.45	0.48	0.55	0.74
	Normal	0.75	0.60	0.85	0.85
F1-Score	Benign	0.74	0.79	0.82	0.85
	Malignant	0.50	0.62	0.67	0.75
	Normal	0.70	0.60	0.72	0.87

Table 6: *Improvement of RobustDeiT against Baseline Architectures (σ = 35)*

Metric		VGGNet (%)	GoogLeNet (%)	ResNet (%)
Accuracy		22.06	13.70	7.79
Precision	Benign	16.67	16.67	5.00
	Malignant	37.50	-12.50	-9.41
	Normal	36.92	48.33	41.27
Recall	Benign	13.16	-2.27	1.18
	Malignant	64.44	54.17	34.55
	Normal	13.33	41.67	0.00
F1-Score	Benign	14.86	7.59	3.66
	Malignant	50.00	20.97	11.94
	Normal	24.29	45.00	20.83

Table 7: *Model Performance Metrics with Gaussian Noise (σ = 45)*

Metric		VGGNet	GoogLeNet	ResNet	RobustDeiT
Accuracy		0.66	0.68	0.73	0.80
Precision	Benign	0.68	0.70	0.75	0.83
	Malignant	0.62	0.71	0.79	0.83
	Normal	0.60	0.53	0.60	0.72
Recall	Benign	0.83	0.83	0.83	0.86
	Malignant	0.52	0.48	0.48	0.61
	Normal	0.30	0.45	0.75	0.90
F1-Score	Benign	0.75	0.76	0.79	0.84
	Malignant	0.56	0.58	0.60	0.70
	Normal	0.40	0.49	0.67	0.80

Table 8: *Improvement of RobustDeiT against Baseline Architectures (σ = 45)*

Metric		VGGNet (%)	GoogLeNet (%)	ResNet (%)
Accuracy		21.21	17.65	9.59
Precision	Benign	22.06	18.57	10.67
	Malignant	33.87	16.90	5.06
	Normal	20.00	35.85	20.00
Recall	Benign	3.61	3.61	3.61
	Malignant	17.31	27.08	27.08
	Normal	200.00	100.00	20.00
F1-Score	Benign	12.00	10.53	6.33
	Malignant	25.00	20.69	16.67
	Normal	100.00	63.27	19.40

The proposed RobustDeiT model demonstrated strong performance across all noise levels, achieving the highest accuracy compared to the other models. At σ = 15, RobustDeiT achieved an accuracy of 0.85, and while the accuracy slightly decreased with higher noise levels, it maintained a high accuracy of 0.80 at σ = 45. In comparison, other models showed greater drops in performance. This consistent accuracy across varying noise levels demonstrates RobustDeiT's ability to classify data correctly even under noisy conditions, making it more reliable than the other models.

In terms of precision, the proposed RobustDeiT model consistently outperformed the other models across all classes and noise levels. At a noise level of σ = 15, RobustDeiT achieved a high precision of 0.88 for the Benign class, 0.76 for the Malignant class, and 0.95 for the Normal class. As noise levels increased, RobustDeiT maintained strong precision values, with only a slight decrease observed at higher noise levels, achieving 0.83, 0.83, and 0.72 for the Benign, Malignant, and Normal classes, respectively, at σ = 45. The precision results further emphasize RobustDeiT's effectiveness in minimizing false positives, ensuring that its predictions are consistently accurate for the target classes even in the presence of noise.

Regarding recall, the proposed RobustDeiT model demonstrated resilience across increasing noise levels, maintaining relatively high recall values for all classes. At a noise level of σ = 15, RobustDeiT achieved recall rates of 0.86 for the Benign class, 0.81 for the Malignant class, and 0.90 for the Normal class, outperforming the other models. Even as the noise increased to σ = 45, RobustDeiT sustained competitive recall scores, reaching 0.86 for Benign, 0.61 for Malignant, and 0.90 for Normal, highlighting its robustness in capturing relevant patterns in the presence of noise. This consistent recall performance indicates that RobustDeiT effectively minimizes missed detections compared to other models under varied noise conditions.

Regarding the F1-score, the proposed RobustDeiT model exhibited strong and consistent performance across varying noise levels. At a noise level of σ = 15, RobustDeiT achieved F1-scores of 0.87 for Benign, 0.78 for Malignant, and 0.92 for Normal, outperforming all other models. As noise increased to σ = 45, RobustDeiT maintained competitive F1-scores, with values of 0.85 for Benign, 0.70 for Malignant, and 0.80 for Normal. These results demonstrate RobustDeiT's ability to balance precision and recall effectively, minimizing both false positives and false negatives. Even under more challenging noise conditions, RobustDeiT continued to deliver reliable, high-quality predictions across all classes, confirming its robustness and effectiveness in handling noisy data.

While some specific metrics display negative improvements for the proposed RobustDeiT model compared to baseline architectures, these are limited to certain instances and are outweighed by notable gains across other metrics. For example, while precision may have declined slightly for certain classes, such as malignant, the model achieves compensating gains in precision for benign and normal cases, alongside substantial improvements in recall and F1-score, as seen in Tables 2, 4, 6, and 8. These gains collectively result in higher overall accuracy and robustness, even under increased Gaussian noise levels.

The performance metrics across different noise levels, previously summarized in tables, are also depicted in Fig. 3 using line plots. These plots provide a clear visual comparison of the models' performance trends, including accuracy, precision, recall, and F1-score for each class, as the noise level increases. The vertical axis represents the metrics, while the horizontal axis indicates the noise levels. By illustrating the results in this format, Fig. 3 facilitates a more intuitive understanding of how each model responds to varying levels of noise, highlighting RobustDeiT's overall superior performance across metrics and noise levels compared to the other

models. This visualization emphasizes the robustness and reliability of the proposed approach under noisy conditions.

Moreover, the incremental improvements at higher noise levels illustrate the proposed model's resilience against noise, achieving consistently higher accuracy than baseline models. The overall advantage across key metrics, including accuracy, precision, recall, and F1-score, indicates that the proposed RobustDeiT model offers a balanced improvement, excelling in noisy conditions where traditional architectures like VGGNet, GoogLeNet, and ResNet exhibit performance drops. This overall performance confirms that the proposed model is not only competitive but also robust across a range of challenging conditions, thereby underscoring its potential effectiveness for practical applications in noisy environments.

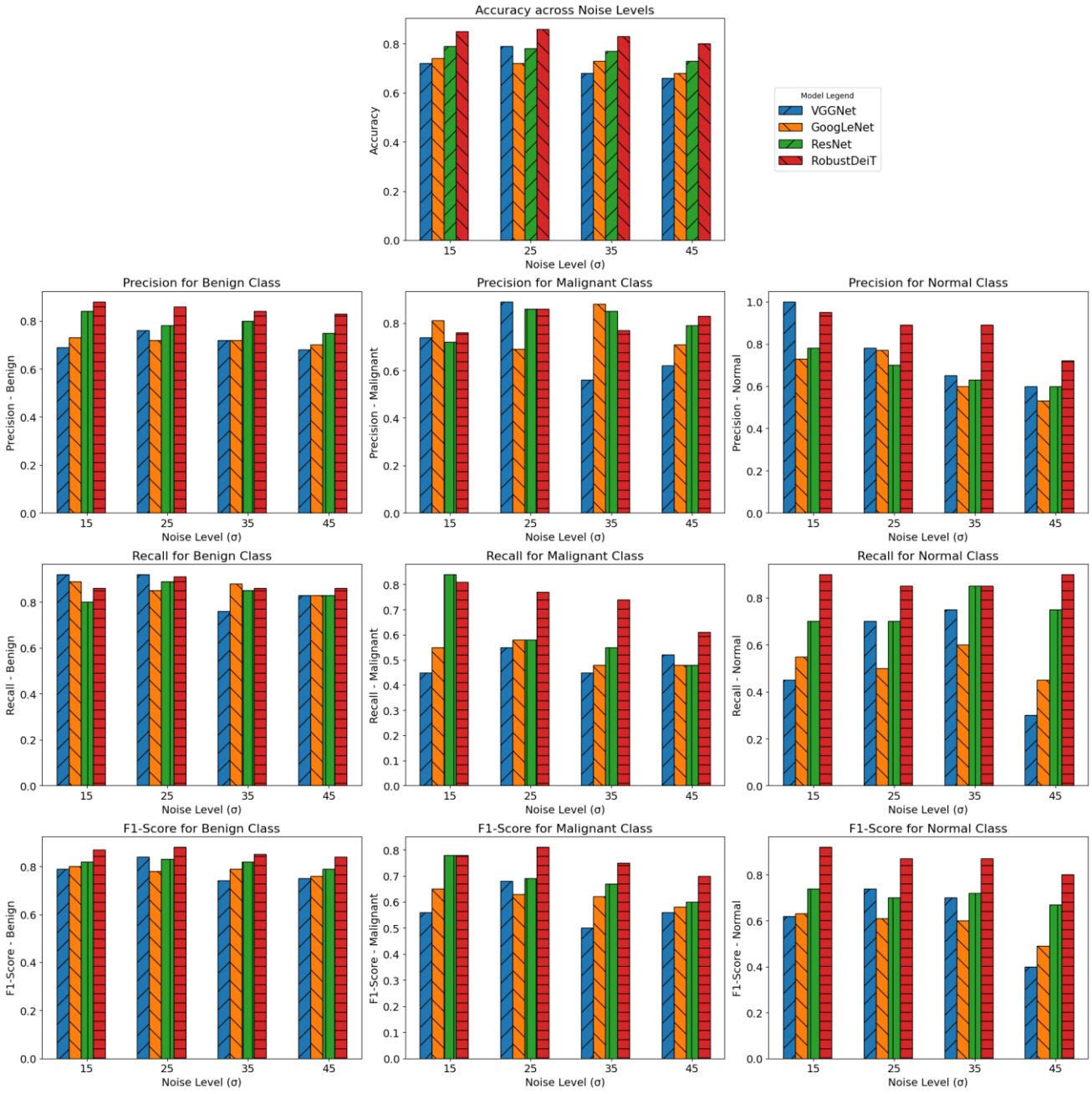

Fig. 3: *Models' Performance across Noise Levels*

DISCUSSION

While the current study evaluates RobustDeiT exclusively on breast ultrasound images, the design of the proposed method does not rely on any specific imaging modality or organ characteristics. By focusing on noise reduction and feature enhancement at a more general image-processing level, RobustDeiT is inherently applicable across various medical imaging techniques, including CT, MRI, and others.

Different modalities introduce distinct noise patterns and artifacts, for example, speckle noise in ultrasound versus motion artifacts in MRI. Similarly, anatomical differences between organs and disease-specific features may pose unique challenges. Our approach aims to be robust against such variations by enhancing image quality and extracting meaningful patterns.

False negative predictions, where a disease is present but not detected by the model, pose a critical risk in medical diagnosis. In the context of ultrasound and other imaging modalities, low contrast, poor image quality, or subtle lesions can contribute to missed detections. RobustDeiT aims to reduce false negatives by enhancing image clarity and emphasizing relevant features, even in noisy conditions. However, no model can entirely eliminate such errors, particularly in cases with atypical presentation or borderline visual cues. Future evaluation on more diverse and clinically challenging datasets will be essential to assess and further minimize false negative risks in real-world applications.

CONCLUSION

In this study, we have developed novel classification architecture for medical images, utilizing the Data-efficient Image Transformer (DeiT) to handle noisy datasets effectively. Our approach integrates a comprehensive preprocessing pipeline that includes advanced denoising techniques, color correction, sharpening, and normalization. The combination of edge-preserving filters, Gaussian blur, and median filtering significantly improves image quality by reducing noise while maintaining essential details. Contrast Limited Adaptive Histogram Equalization (CLAHE) enhances visual clarity and feature extraction, while the unsharp mask technique sharpens edges to make features more discernible. Normalization ensures consistent brightness and contrast across the dataset, further enhancing classification reliability. The integration of these preprocessing techniques with the DeiT model has demonstrated significant improvements in handling noisy medical images, leading to more accurate and reliable classification outcomes. Our findings underscore the effectiveness of this approach in advancing medical image analysis, providing a robust solution for challenging, noisy environments. Future work could explore additional optimizations and adaptations of this framework to further enhance performance and applicability in diverse medical imaging contexts.

REFERENCES

Al-Dhabyani W, Gomaa M, Khaled H, Fahmy A (2020). Dataset of breast ultrasound images. *Data Brief* 28:104863.

Azad R, Kazerouni A, Heidari M, Aghdam EK, Molaei A, Jia Y, Merhof D (2023). Advances in medical image analysis with vision transformers: a comprehensive review. *Med Image Anal* 103000.

Chang Y, Jung C, Ke P, Song H, Hwang J (2018). Automatic contrast-limited adaptive histogram equalization with dual gamma correction. *IEEE Access* 6:11782–92.

Chen RC, Dewi C, Zhuang YC, Chen JK (2023). Contrast limited adaptive histogram equalization for recognizing road marking at night based on YOLO models. *IEEE Access*.

Dalmaz O, Yurt M, Çukur T (2022). ResViT: residual vision transformers for multimodal medical image synthesis. *IEEE Trans Med Imaging* 41(10):2598–2614.

Dosovitskiy A (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv Prepr. arXiv:2010.11929*.

Ju L, Wang X, Wang L, Mahapatra D, Zhao X, Zhou Q, Ge Z (2022). Improving medical images classification with label noise using dual-uncertainty estimation. *IEEE Trans. Med Imaging* 41(6):1533–46.

Kansal S, Purwar S, Tripathi RK (2018). Image contrast enhancement using unsharp masking and histogram equalization. *Multimed Tools Appl.* 77:26919–38.

Li Q, Shen L, Guo S, Lai Z (2021). WaveCNet: Wavelet integrated CNNs to suppress aliasing effect for noise-robust image classification. *IEEE Trans. Image Process.* 30:7074–89.

Ling Y, Wang Y, Dai W, Yu J, Liang P, Kong D (2024). MTANet: Multi-task attention network for automatic medical image segmentation and classification. *IEEE Trans Med Imaging* 43:674-85.

Liu J, Li R, Sun C (2021). Co-correcting: noise-tolerant medical image classification via mutual label correction. *IEEE Trans Med Imaging* 40:3580–92.

Manzari ON, Ahmadabadi H, Kashiani H, Shokouhi SB, Ayatollahi A (2023). MedViT: a robust vision transformer for generalized medical image classification. *Comput Biol Med* 157:106791.

Mishiba K (2023). Fast guided median filter. *IEEE Trans Image Process* 32:737–49.

Penso C, Frenkel L, Goldberger J (2024). Confidence calibration of a medical imaging classification system that is robust to label noise. *IEEE Trans Med Imaging* 43:2050-60.

Shamshad F, Khan S, Zamir SW, Khan MH, Hayat M, Khan FS, Fu H (2023). Transformers in medical imaging: A survey. *Med Image Anal* 88:102802.

Song Y, He Z, Qian H, Du X (2023). Vision transformers for single image dehazing. *IEEE Trans. Image Process* 32:1927–41.

Taassori M (2024). Enhanced wavelet-based medical image denoising with Bayesian-optimized bilateral filtering. *Sensors* 24:6849.

Taassori M, Vizvári B (2024). Enhancing medical image denoising: A hybrid approach incorporating adaptive Kalman filter and non-local means with Latin square optimization. *Electronics* 13:2640.

Touvron H, Cord M, Douze M, Massa F, Sablayrolles A, Jégou H (2021). Training data-efficient image transformers & distillation through attention. In: *Int. Conf. Mach. Learn.*, pp. 10347–57. PMLR.

Vaswani A (2017). Attention is all you need. *Adv. Neural Inf. Process. Syst.* 30.

Wang D, Nieto JJ, Li X, Li Y (2020). A spatially adaptive edge-preserving denoising method based on fractional-order variational PDEs. *IEEE Access* 8:163115–28.

Xue C, Yu L, Chen P, Dou Q, Heng PA (2022). Robust medical image classification from noisy labeled data with global and local representation guided co-training. *IEEE Trans. Med. Imaging* 41:1371–82.

Yang Y, Hui H, Zeng L, Zhao Y, Zhan Y, Yan T (2021). Edge-preserving image filtering based on soft clustering. *IEEE Trans Circuits Syst Video Technol* 32:4150–62.

Ye W, Ma KK (2018). Blurriness-guided unsharp masking. *IEEE Trans. Image Process* 27(9):4465–77.

Zhou S, Wang J, Wang L, Zhang J, Wang F, Huang D, Zheng N (2020). Hierarchical and interactive refinement network for edge-preserving salient object detection. *IEEE Trans. Image Process* 30:1–14.

Zhu C, Chen W, Peng T, Wang Y, Jin M (2021). Hard sample aware noise robust learning for histopathology image classification. *IEEE Trans. Med. Imaging* 41:881–94.

Zhu H, Ng MK (2020). Structured dictionary learning for image denoising under mixed Gaussian and impulse noise. *IEEE Trans Image Process* 29:6680–93.