Breast Cancer Detection in Mammography Images Using Neighborhood Attention Transformer and Shearlet Transform

Mahdi Firouzbakht

Department of Computer Engineering Amirkabir University of Technology
Tehran, Iran
mahdi.firuzbakht@aut.ac.ir

Maryam Amirmazlaghani *

Department of Computer Engineering Amirkabir University of Technology
Tehran, Iran
mazlaghani@aut.ac.ir

*Corresponding Author: mazlaghani@aut.ac.ir (Maryam Amirmazlaghani)

Abstract

Breast cancer is a leading cause of cancer-related deaths among women. Advances in early diagnosis and treatment, particularly through screening mammography, have reduced mortality rates by enabling the detection of small tumors. Recently, artificial intelligence (AI) and advanced computer vision models have further improved breast cancer detection and diagnosis. In this research, we have developed a novel model for detecting breast cancer in mammography images by extracting rich and suitable features. Our model utilizes the Neighborhood Attention Transformer, which enhances local feature processing by focusing on neighborhood attention alongside global and long-range features. This is crucial for analyzing masses within and at the boundaries. Additionally, we incorporate the Shearlet Transform to enhance feature extraction by capturing frequency-domain features, essential for precise edge and texture analysis in mammographic images. The Shearlet Transform's ability to manage anisotropic features and its strong localization in both spatial and frequency domains makes it particularly effective. Denoising is another key aspect, as mammograms often contain noise from imaging conditions and devices. To address this, our model applies Shearlet-based adaptive shrinkage denoising, significantly improving feature extraction. By combining the energy of Shearlet subbands with features from previous techniques, our model simplifies feature representation, highlights key patterns, and remains robust to noise and transformations. Our proposed model has achieved impressive results on the Curated Breast Imaging Subset of the Digital Database for Screening Mammography (CBIS-DDSM) dataset, with F1, Area Under the Curve (AUC), and Kappa scores of 76.8%, 84.5%, and 50.9%, respectively, outperforming other models.

Keywords

Mammogram; Breast Cancer Classification; Neighborhood Attention Transformer; Shearlet Transform; Adaptive Denoising; Computer Vision.

1. Introduction

What is Breast Cancer? Breast cancer is a condition where cancer cells grow in the tissues of the breast. There are several different types of breast cancer that affect both men and women. Breast cancer is a common issue that affects many American women. Approximately 13% of women, or about 1 in 8, will be diagnosed with breast cancer during their lifetime. It is the most common cancer among women, excluding skin cancer, and is expected to account for about 30% of new cancer cases in women. Early detection of around 64% of cases in the early and localized stages increases the effectiveness of treatment. In 2024, it is estimated that 297,790 cases of invasive breast cancer and 55,720 cases of non-invasive breast cancer will be diagnosed in American women. In the same year, it is estimated that 43,700 people will die from breast cancer [1]. Early detection of breast cancer leads to reduced mortality from this disease [2]. The mortality rate from breast cancer among women decreased by 40% from 1989 to 2016. This improvement is attributed to advances in early diagnosis [3].

Annually, the American Cancer Society provides estimates for new cancer cases and deaths in the United States. These data are extracted from central cancer registries that collect incidence information and from mortality data sourced from the National Center for Health Statistics. For 2023, it was predicted that the United States would see about 1,958,310 new cancer cases and 609,820 cancer-related deaths. Additionally, a comprehensive chart is available that breaks down these statistics by cancer type and gender [4]. Various breast cancer imaging methods used in computer-aided detection (CAD) systems include mammography, ultrasound, magnetic resonance imaging (MRI), and thermography. These imaging methods have been extensively reviewed in numerous scientific articles, demonstrating positive results in the identification and evaluation of breast cancer. Mammography is recognized as the leading and widely used method in breast cancer imaging. Mammography images are typically provided as detailed X-ray images of the breast, capturing the internal structure and tissue density. The process involves compressing the breast between two plates to spread the tissue for clearer images. Radiologists then examine these images for any signs of abnormalities, such as lumps or calcifications, which could indicate breast cancer or other conditions [5]. The integration of CAD systems has been undertaken to enhance the accuracy of breast cancer diagnosis based on mammogram images. Additionally, alongside these imaging techniques, BI-RADS [6] (Breast Imaging Reporting and Data System) serves as a standardized system for interpreting and reporting breast imaging results, particularly mammography. This system categorizes findings into seven numbered categories to assess the likelihood of breast cancer and guides healthcare providers in determining the next steps for patient evaluation and management.

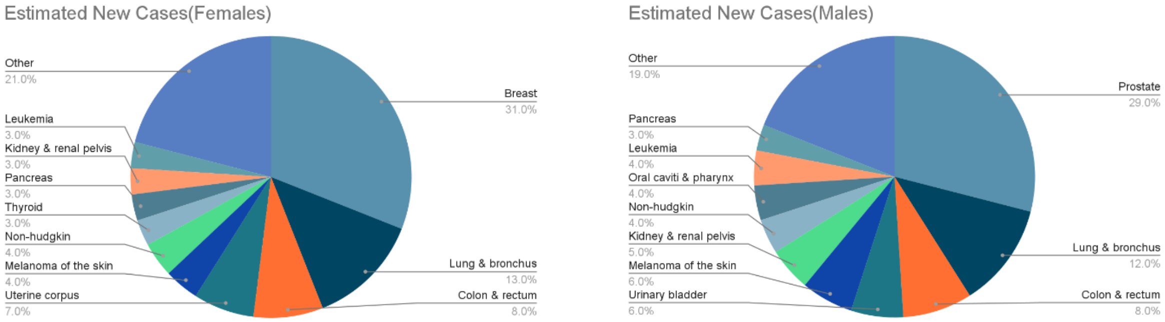


Figure 1. The Top Ten Types of Cancer for Predicted New Cases and Deaths by Gender in the United States for 2023 [4].

This field is of high sensitivity both in terms of the number of cases involved and the number of fatalities. It is very clear in the Figure 1 that the breast cancer in female gender has a very important role in new estimated cancer cases. In recent years, given the importance of early cancer detection [2] and the integration of imaging methods, especially mammogram images with computer-aided detection systems, significant advancements have been made in this area. However, we must keep in mind that there are many differences between normal and natural images with mammography images. Natural images and mammograms differ significantly when using deep learning. Natural images are diverse and include everyday scenes, whereas mammograms are high-resolution medical images focusing on breast tissue details. Natural image datasets are large and varied, while mammogram datasets are smaller, highly imbalanced with fewer positive (abnormal) cases, and require expert annotation. Preprocessing and augmentation techniques for mammograms must preserve medical features, unlike the more flexible methods used for natural images. By considering these differences, we can examine the presented systems.

Over the years, various techniques have been employed to aid healthcare systems in diagnosing cancer. These systems have progressed day by day, significantly improving in accuracy, speed, efficiency, reliability, and robustness. Initially, traditional CAD systems were developed. These systems did not use artificial intelligence and did not have a highly positive impact in assisting radiologists [7]. One of the drawbacks of these methods was the high number of false positives [8]. High false positive rates lead to stress, unnecessary radiological tests, high treatment costs, and various examinations [9]. Therefore, there has been a need for newer, more accurate methods. During this time, AI-based methods, especially machine learning techniques, have emerged. In particular, deep learning methods have made significant advancements in this field. In neural network and deep learning methods, mammography image data are input into neural networks, resulting in various outputs such as classification, segmentation, and detection. These methods have been used in recent years to examine and prevent cancer in its early stages [10].

Convolutional networks are one of these types of networks that have been extensively used in recent years, addressing the issues of traditional systems [10]. These methods have made substantial advancements each year, gradually approaching a stage where they can practically assist specialists in improving human health conditions. Recent advancements even enable these methods to accurately detect very small lesions in images, significantly aiding in early disease diagnosis [11]. In medical image processing with convolutional models, there are two main categories. One category uses the entire mammogram image as the input image, while the other category uses patches of these images. The reason for using patches, according to Shen et al. [12], is that only a small part of the image contains the lesion, and this small lesion is considered the determinant of the class of the entire image, making training very challenging. Therefore, the patching method has been proposed. However, it should be noted that many existing datasets do not contain patches and have not been patched specifically by radiologists. To create patches of images using artificial intelligence, we first need to train specific models like You Only Look Once (YOLO) [13] on annotated data and then use this model to create patches of mammography images. This method has issues regarding time, cost, accuracy, and reliability, making it difficult to use with confidence. Additionally, using this method may result in the loss of some auxiliary image information, which can increase the false positive rate [14]. Furthermore, in this method, the size and resolution of each patch are important for the final result, requiring many experiments to achieve the best possible outcome, which demands significant time and cost [15]. One of the main features of CNNs is their localized receptive fields. Each neuron in a CNN layer has a local receptive field determined by the size of the convolutional filters (kernels) applied to the input data. This local receptive field allows CNNs to find and use local features, but due to this characteristic, these models perform poorly in learning long-range information, limiting their capabilities in vision tasks [16].

Following these advancements, transformer-based methods have been introduced to address the issues of CNN methods, especially in long-range information. The Vision Transformer model, introduced by Dosovitskiy et al. [17] and inspired by the transformer model introduced by Vaswani et al. [18], was the first transformer-based model. In this model, data are input into the model as patches of the original image. Additionally, due to the self-attention module in this model, the relationships between patches are well-learned [17]. One of the main issues with this model is that, despite its understanding and learning of long-range and global information, it does not learn local information well and cannot consider the neighboring relationships of each pixel. In this regard, Neighborhood Attention Transformer (NAT) can be considered a suitable example of transformer-based models, as it can understand and learn local and neighboring information like CNN models while also correctly understanding global and long-range relationships.

Mammography images are created using radio waves, which may contain information in the frequency domain that few medical image processing methods consider. Additionally, these images may contain noise due to environmental conditions and the devices used for imaging, which can affect the final results of models. This issue is significant enough that noise can completely disrupt the performance of models [19]. Despite the extensive work done in this area, the numerous challenges associated with these types of images and imaging techniques keep the field open for achieving more stable results.

In this paper, we will present a new model called DualShearNAT for breast cancer detection. The DualShearNAT is a transformer-based model that employs neighborhood-attention instead of self-attention. This approach combines global and local processing capabilities. We also use the shearlet transform for feature extraction, leveraging its ability to extract unique information. Performance is enhanced with shearlet-transformed image energy, and image quality is improved with noise removal techniques based on the shearlet transform. Additionally, we employ some techniques to address dataset imbalance.

In this study, we aim to address the shortcomings of previous methods by introducing our new model. Previous approaches have several drawbacks, such as failing to simultaneously consider long-range and local features, as well as the interactions between these features. Additionally, they do not account for frequency domain features and inadequately remove noise from mammographic images, which is crucial for achieving accurate results. Moreover, there is an imbalance in the datasets used for mammographic images, necessitating a more suitable method beyond augmentation to address this issue. We have identified these problems in our study and have made efforts to resolve them. Our main contributions are as follows:

- in the processing of medical images, especially breast cancer images, the local relationship of each part of the image and each pixel with its neighboring pixels is highly significant. In this paper, we demonstrate that capturing these relationships using the NAT-based model produces better results in breast cancer classification compared to other similar models in ImageNet tasks.
- Furthermore, we will demonstrate the shearlet transform can extract information that convolutional and transformer models cannot. By combining these features with the features extracted from the original image, we can create a more powerful and robust model. Additionally, we use attention-based feature fusion to emphasize more important features.
- This paper also examines the impact of noise removal from medical images, particularly breast cancer images, and highlights the importance of denoising these images. We use the shearlet transform for noise removal, as it can extract critical features that play a significant role in image formation, enabling us to identify and eliminate noise generated during imaging.
- This paper also examines the impacts of dataset imbalance on the final model outcome. To address this issue, several different strategies have been proposed. One of the best solutions presented in this study is the use of Focal Loss, which has proven effective in balancing the dataset and improving the results significantly.
- In the results section, we will conduct a detailed and step-by-step analysis of the paper. We will precisely explain the effect of adding or removing each element and module to our model through extensive experiments.

In the following, we will first review previous works in Section 2. Section 3 will cover the description of the proposed model. Finally, in the last section, we will share the results of the experiments conducted on our model and conclude the paper.

2. Related Works

In recent years, numerous studies have been conducted addressing various challenges in mammography imaging. There are two main challenges in classifying these images: 1) distinguishing between tumor-bearing images and normal images, and 2) determining whether tumor-bearing images are benign or malignant. Several existing datasets, such as MIAS [20], InBreast [21], and VinDr [22], include three classes: normal, benign tumor, and malignant tumor, making them suitable for both tasks. However, datasets like CBIS-DDSM [23] only contain tumor-bearing images, categorized into benign and malignant tumor classes. Many studies on these datasets have yielded significant results.

Despite the tremendous advancements and extraordinary achievements of deep learning methods, the merits of traditional methods should not be overlooked. These older methods offer better interpretability compared to deep learning methods. Additionally, they are more efficient in terms of execution speed and time. They also perform significantly better with small datasets than deep learning methods. Therefore, we will review some traditional methods. In the study [24], Local Binary Pattern (LBP) [25], thresholding, and Support Vector Machine (SVM) [26] were used for classification. Despite employing traditional methods on the DDSM and MIAS datasets, an accuracy of 84% was achieved. In the study [27], a combination of Gabor wavelet [28], Principal Component Analysis (PCA) [29], and SVM was utilized. In this study, the Gabor wavelet was used for filtered mammography images. Directional features were then extracted in various orientations and frequencies. In the next step, data dimensions were reduced using PCA, and in the final step, data were classified using SVM. The MIAS dataset was used in this study. In the study [30], Gray Level Co-Occurrence Matrix (GLCM) was used for feature extraction. In this model, four features were initially extracted from each ROI of each image using GLCM. Then, image classification was performed using Radial Basis Function Neural Network (RBFNN). This model, which used the MIAS dataset, achieved an accuracy of 93.98% for distinguishing normal and abnormal images and 94.29% for determining whether the tumor is benign or malignant.

Despite the strength of traditional methods, over time, the field of medical image processing has benefited from convolution-based methods. In the study [31], a model was presented for distinguishing between benign and malignant tumor images. This research consists of two sequential parts. In the first part, images are preprocessed using various techniques and then fed into a convolutional model for feature extraction and classification. Both ROI images and whole images are used in this study. The research achieved an AUC of 98% for the InBreast dataset and 97.4% for the MIAS dataset.

In the study [32], researchers used four different ResNet [33] models to achieve the best results for the MIAS dataset, with the best accuracy being 93.83%. In the study [34], researchers combined k-means clustering, RNN architecture, CNN architecture, random forest technique, and boosting technique for classification. This study, which classified normal, benign, and malignant images, achieved an accuracy of 96% for the DDSM dataset and 95% for the MIAS dataset. In the study [35], researchers used a convolutional structure to create their model. Alongside the MIAS dataset, a private and unofficial dataset was also used to further improve their model, achieving a 99% AUC score.

Older studies, such as [36], also used deep learning methods on the DDSM dataset to differentiate between benign and malignant tumor images. In this study [36], a convolutional model and other techniques such as the grid search algorithm [37] were used. In more recent studies, like [38], an ensemble-based model with spatial and channel attention was used. This model consists of two parts. The first part uses various augmentation techniques to enhance input data. The second part uses a parallel structure with a modified SE-ResNet50 [39] and InceptionV3 [40] as the backbone for feature extraction. This research achieved an accuracy of 96.5% for the DDSM dataset and 95.33% for the MIAS dataset.

In the study [41], multiple convolutional networks were used in an ensemble approach. This study combined the MIAS and CBIS-DDSM datasets with another dataset to form a composite dataset with two classes: benign and malignant tumors. The final result of this model showed an accuracy of 99.76%. In the study [42], various convolution-based methods were used to determine which conv-based model yields better results, resulting in an accuracy of 67.08%. In the paper [43], a parallel process was used. In the CBIS-DDSM dataset, each breast has at least two images: one from the side and one from the top. This study processed both images in parallel using a pre-trained EfficientNet and achieved two results. The AUC for this study was 80.33% with the default dataset division and 84.83% with a modified division. In the study [15], the Densenet-121 model was used. This research aimed to achieve the best results by considering different patch sizes and resolution sizes, ultimately achieving an AUC of 80.9%.

In recent years, following the remarkable success of transformer-based and attention-based methods, the field of medical image processing, particularly breast medical image processing, has also adopted these approaches. In the study [44], a model based on VIT was used. This study stated that VIT-based models, due to their exceptional power in handling long-range dependencies, can easily overcome the challenges of previous computer vision methods. The InBreast dataset was used in this study, achieving an accuracy of 96.48%. In the study [45], a new type of data augmentation called AGE, which stands for attention-guided erasing, was used in medical image processing. This paper utilized a VIT-based model and examined the VinDr dataset, achieving an F1 score of 59.1%. Without AGE, this score would have been between 55.94% and 56.91%.

In the study [46], a hybrid approach was used. In this research, data undergo preprocessing stages and are then fed into a deep convolutional model for deep feature extraction. The extracted features are then input into a VIT-based network for classification. The DDSM dataset was used in this study, achieving an accuracy of 95.80% for classifying normal, malignant, and benign data. The study [47] employed a local and global structure. In this study, the data consists of four images: two images from the side and two from the top of the breast. Initially, each image is separately input into a VIT-based structure. Then, the output features of each image are concatenated and input into another VIT-based model. This study was conducted on private datasets, resulting in an AUC of 81.8%. The study [48] used a Swin-based structure, which, by altering the structure and using cross-attention, aimed to effectively integrate information from different views of each breast and improve the coherent transfer of this information between different views. This model was tested on the CBIS-DDSM and VinDr datasets, achieving an AUC of 66.43% for CBIS-DDSM data and 96.08% for VinDr data.

In some studies [49], researchers have used intrinsic features of each image as inputs rather than directly using the images as the main features. In [49], researchers first applied shearlet transforms to the input images and then extracted the desired features from these transformed images. The energy of the shearlet-transformed images was used as a feature, which is similar to the GLCM feature extraction method used in study [30]. This study was conducted on the MIAS dataset, achieving an accuracy of 95.83% in the best case.

One important aspect, especially in medical imaging, is noise removal. The importance of noise removal in medical image processing is significant due to several major reasons. Noise removal improves the quality of medical images, which is crucial for more accurate and efficient diagnosis of diseases. Medical images typically encounter various noises due to imaging conditions and limitations of imaging devices. The presence of noise can interfere with the accurate and timely diagnosis of diseases and may even lead to incorrect medical decisions. Removing noise from medical images helps to reveal more details of internal body structures, increasing the accuracy in image analysis and interpretation. Numerous studies have addressed the importance of noise removal in medical image processing. For instance, the book [50] states that using various filters to reduce noise can improve image quality, which is highly effective in clinical diagnoses. Additionally, research [51] shows that using advanced noise removal methods like "Shrinkage" in medical images can significantly enhance diagnostic accuracy.

In this study, we focus on the processing and classification of images from the CBIS-DDSM dataset. Inspired by recent achievements of transformer-based methods in various fields, we aim to develop an appropriate system for processing these images. For this purpose, we proposed new model DualShearNAT, based on NAT model, a relatively newer transformer-based model. Instead of using self-attention, this model employs neighborhood-attention. This means that in processing each pixel of the image, the information from the surrounding neighborhood of the pixel is also considered in the calculations, allowing the model to benefit from both global processing capabilities and local processing similar to convolutional models. Additionally, we utilize the shearlet transform for feature extraction to support our base model. Due to its unique properties, the shearlet transform can extract information that ordinary models cannot. Inspired by study [49], we also use the energy of shearlet-transformed images, which enhances the system's performance. Furthermore, to improve image quality, we have implemented noise removal inspired by research [51]. Additionally, we use techniques to tackle dataset imbalance that affect the outcomes of networks.

3. Proposed Method

In this section, we propose a new model called DualShearNAT for breast cancer detection using mammography images. This model employs neighborhood attention (NA) instead of self-attention for both global and local pixel processing. We enhance the model with shearlet transform for unique feature extraction capabilities, leveraging energy from shearlet-transformed images to boost performance. Our approach also includes noise removal techniques and also strategies to address dataset imbalance, crucial for improving network outcomes.

The block diagram of the proposed model is depicted in Figure 2. This model is composed of three parallel branches: 1. Spatial Branch, 2. Shearlet Branch, and 3. Energy Branch. Then the combination of these branches is sent to multilayer perceptron (MLP) to classify these combined features. The reason for using this structure is that we aim to combine three features to access a better model that exhibits improved structural features. This will allow MLP layers to create better classifications using these features. The reason for using the neighborhood attention in this paper is due to its unique characteristics. By utilizing the neighborhood attention, the model can leverage the structure around each pixel and produce suitable results with the help of this neighborhood. This is particularly important in mammogram images and their lesions, where the neighborhood of each pixel, both within and at the border of the lesion, is crucial. This type of transformer can examine both the local neighborhood and the long-range features of the images, thereby extracting appropriate features.

We have also considered enhancing our model by incorporating frequency domain and statistical features. Features, especially those related to Shearlet transforms, which are connected to the frequency domain, can yield specific characteristics. This is particularly relevant for mammography images, where radio wave radiation is used in imaging. In this context, we use a 2-level Shearlet transforms to extract the features of these images, resulting in 17 output images. These extracted features are then re-examined by the presented model, which outputs the enhanced features with Shearlet transformations.

In the next step, the Shearlet-transformed output features (shearlet branch) are combined with the features extracted using NAT (Spatial branch) to create a stronger combination of features. This combination again utilizes neighborhood attention layers to highlight important features and examine the output features of the two branches together for better results. Alongside these two branches, a third branch uses the statistical energy features of the Shearlet subbands. We use the energy of each image as an important feature in this model. The energy of each image is the average of the features within each image, representing the strength of each feature, which can play a significant role. We use these three branches and their combination to enhance our model, covering various important features and allowing MLP layers to yield the best possible results. In the following, we will explain each component of our proposed model in detail:

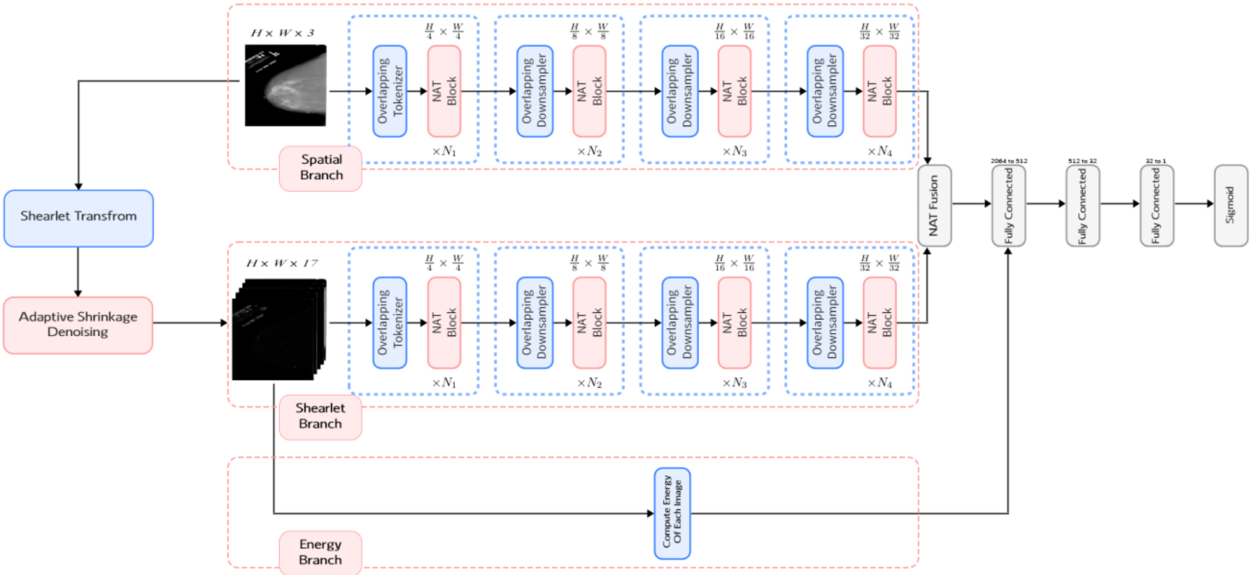


Figure 2. The block diagram of the DualShearNAT Model (Proposed Model)

3.1.Spatial Branch

In the Spatial Branch, we have used the base NAT model. Transformers are used instead of traditional convolutional models for several reasons. First, transformers leverage global features. They excel at capturing long-range dependencies and global features in images. Unlike CNNs, which are inherently local in their operation, transformers can attend to all parts of an image simultaneously, allowing them to understand complex relationships between distant pixels. Second, transformers can effectively scale with increased data and computational resources. They have demonstrated significant performance improvements with increased size, depth, or training on larger datasets [17]. Due to the good characteristics of NAT model such as ability to localize attention to the neighborhood around each token, introduce local inductive biases, maintain translational equivariance, and enable receptive field growth without additional operations, we use this transformer and, in the following, we discuss NAT model:

3.1.1. Neighborhood Attention Transformer Model

In this section, we discuss neighborhood attention (NA), a localized form of self-attention designed for visual data. neighborhood attention reduces computational costs compared to traditional self-attention by incorporating useful inductive biases similar to those in convolutions. Additionally, the Neighborhood Attention Transformer (NAT) uses neighborhood attention instead of self-attention. NAT employs a multi-level hierarchical design akin to Swin [52], utilizing overlapping convolutions to sample feature maps across different levels. This approach, in contrast to non-overlapping patches, enhances model performance by introducing useful inductive biases.

Swin's Window Self Attention (WSA) [52] is recognized as one of the fastest methods to constrain self-attention, aiming to reduce the quadratic attention cost by partitioning inputs and applying self-attention to each partition separately. To extend its effectiveness, WSA is complemented by a shifted variant, SWSA, which adjusts partition boundaries to facilitate interactions beyond separate sections, essential for expanding the receptive field. However, a more direct approach to localizing self-attention is to allow each pixel to attend to its neighboring pixels. This creates a dynamically shifted window around most pixels, effectively expanding the receptive field without manual adjustments. Unlike Swin but similar to convolutions, this dynamic form of localized self-attention can maintain translational equivariance [53].

Given an input sequence in the form of a matrix, $X \in R^{n \times d}$, where the rows are d-dimensional token vectors, X is first transformed into queries and key-value pairs. Then, the neighborhood attention weights for the i-th token, denoted as A with a neighborhood size k, are defined as a matrix-vector product as follows:

$$A_i^k \;=\; \begin{bmatrix} Q_i K_{\rho_1(i)}^T \\ Q_i K_{\rho_2(i)}^T \\ \vdots \\ Q_i K_{\rho_k(i)}^T \end{bmatrix}$$

Equation 1

where ρ represents the j-th nearest neighbor to token i. Similarly, the neighboring values, V, are defined, upon which the attention weights are applied, as a matrix whose rows are the k nearest predicted value neighbors of the i-th token:

$$V_i^k \;=\; [V_{\rho_1(i)}^T \;\; V_{\rho_2(i)}^T \;\; \cdots \;\; V_{\rho_k(i)}^T]^T$$

Equation 2

The final output from the neighborhood attention for the i-th token with neighborhood size k is:

$$NA_k(i) \;=\; softmax\left(\frac{A_i^k}{\sqrt{\mathrm{d}}}\right) V_i^k$$

Equation 3

Which includes the scaling parameter √d and the embedding dimension d. As shown in Figure 3, this operation is applied to each pixel in the feature map. As the neighborhood size k increases, A_i^k approximates the self-attention weights, and V^k approximates V_i, making the neighborhood attention approach self-attention.

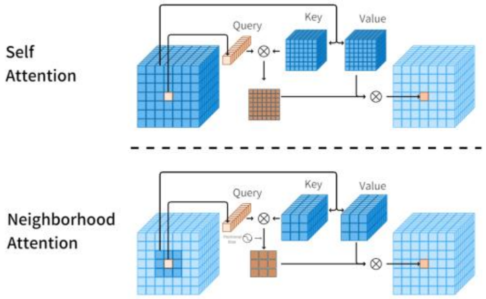


Figure 3. Comparison of the Query-Key-Value Structure in Neighbor Attention (NA) and Self-Attention (SA) for a Pixel [53].

NAT employs a method involving two consecutive 3x3 convolutions with 2x2 strides, resulting in a reduction of the spatial size to 1/4 of the input. This approach is similar to using a patch layer and embedding with 4x4 patches but utilizes overlapping convolutions to introduce useful inductive biases. However, the use of overlapping convolutions may increase computational costs, and using two convolutions adds more parameters. Nonetheless, these challenges are managed through model reconfiguration, resulting in an optimal balance [53].

NAT consists of four levels, each followed by a downsampler (except for the final level). Downsampling reduces the spatial dimensions by half while doubling the number of channels. Unlike Swin's patch merging technique using 2x2 non-overlapping convolutions, NAT uses 3x3 convolutions with 2x2 strides. Given the 4x downsampling factor by the tokenizer, the model produces feature maps with sizes h/4 x w/4, h/8 x w/8, h/16 x w/16, and h/32 x w/32. This adaptation is inspired by established CNN architectures and aligns with hierarchical attention-based methods like PVT [54], ViL [55], and Swin Transformer [52]. Additionally, LayerScale is integrated to enhance stability in larger models.

Considering all the given explanations, we chose the NAT model. This model, while having the general features of transformers, also possesses other unique capabilities. By utilizing neighborhood attention, this model allows each pixel to play a specific role relative to its neighboring pixels and significantly impact the results of its neighbors [53]. This feature, besides the inherent global feature extraction of transformer models, enables this model to also cover local features to a certain extent. Moreover, this model performs better than the Swin model in feature extraction, especially for local features, as evidenced by the results of the NAT paper [53]. This model acts similarly to convolutions in extracting local features to some extent. Like the Swin model, it uses a hierarchical structure. This model has shown better results than its counterparts, Swin and ConvNeXt [56]. For instance, in the classification task on the ImageNet-1K [57] dataset, the top-1 accuracy of this model is 84.3, compared to 83.5 and 83.8 for Swin and ConvNeXt models, respectively [53]. For the input to this model, we used 224*224*3 mammography images. These images, after passing through a preprocessing stage detailed in subsequent sections, are fed into the model. Initially, the model uses two 3x3 convolutional layers for initial embedding. The model then consists of four main blocks with dimension reducers. Each block contains 3, 4, 18, and 5 NAT blocks, respectively. The output of these blocks is an array of length 1024.

3.2.Shearlet Branch

In the Shearlet Branch, we first apply a 2-level shearlet transform to our images to obtain 17 subbands. These subbands are then sent to the denoising module. In this module, each subband is denoised, and the outputs are combined into an image with 17 channels, which is then input to the NAT model. It's important to note that the NAT model used in this branch is the same as the spatial model. However, we have adjusted the input size of the image to accommodate the 17 channels in this branch, compared to the 3 channels in the spatial branch. The model processes these images as it would typical images, aiming to extract their features. The advantage of using these images is that they contain shearlet-transformed images that have been denoised through adaptive shrinkage, thus capturing valuable features which are then passed to the NAT model. First, we will describe the shearlet transform. Following this, we will explain the adaptive shrinkage denoising technique. Since the model used in this branch is identical to the spatial model, we will omit its description to avoid redundancy:

3.2.1. Shearlet Transform

Wavelet theory, developed in the 1980s, struggled with anisotropic features and directional information in higher dimensions [58]. To overcome these issues, Donoho and Kutyniok introduced Shearlets in their 2002 paper, which manage anisotropic features using shear transforms. Subsequent advancements by Labate, Lim, Kutyniok, and Weiss, notably in a 2005 paper, detailed the Shearlet transform's structure and effectiveness in multi-dimensional data representation. The comprehensive 2012 book by Kutyniok and Labate solidified Shearlet theory and its applications as fundamental. The Shearlet transform can be applied in both continuous and discrete domains. Below, the continuous and discrete Shearlet transforms are explained [58].

Continuous Shearlet Transform (CST)

The Continuous Shearlet Transform (CST) is a mathematical tool for analyzing complex data in signal and image processing, offering high accuracy by operating in a continuous domain [58]. It uses shearlets, defined by scale, shear, and translation, to decompose signals and images into different details. This decomposition is achieved through a generating function that is scaled, sheared, and translated to create various shearlets:

$$SH_{\psi}f(a,s,t) = \langle f, \psi(a,s,t) \rangle \quad \text{Equation 4}$$

In this formula, f(x) represents the signal or image, W is the Shearlet transform that is parameterized by a, s, and t. SHψf(a,s,t) are the Shearlet coefficients that represent the correlation between the original signal and the Shearlet-transformed signal at different scales, shifts, and locations. In Figure 4 we can see the overview of shearlets in frequency domain.

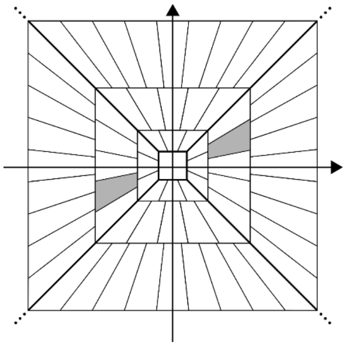


Figure 4. Support of shearlets ψ(a, s, t) in the frequency domain [73].

Discrete Shearlet Transform (DST)

The Discrete Shearlet Transform (DST) decomposes signals or images into components sensitive to scale and orientation using shear matrices, which incorporate directional scaling factors. It can be converted from its continuous form by sampling scale, shear, and translation parameters [58]. Unlike wavelets, DST includes additional directional subbands at each scale, capturing features such as edges and textures effectively [59]. This directional sensitivity makes it useful for edge detection [60], texture classification [61], and image inpainting [62]. The DST is applied in fields like medical imaging [63] and computer vision, enhancing tasks such as image denoising [64] and feature extraction [65]. Its ability to provide sparse representations makes it efficient for data compression [66].

3.2.2. Adaptive Shrinkage Denoising

We proceed within the model after shearlet transform. next, we need to denoise our shearlet images. Medical imaging differs significantly from many other imaging tasks due to the high attention to detail required. For instance, an unusual artifact in an X-ray image could lead to a misdiagnosis of a tumor. Often, the signs of many medical conditions are subtle and hard to detect, making precise processing of such data essential. There has been relatively little published on the effects of image noise reduction on image classification tasks compared to the amount of computer vision work released in recent years. This is especially true in the field of medical imaging, where even less work has been done on this topic [67]. Shearlet coefficients, like wavelet coefficients, exhibit correlation within a small neighborhood. This means that a large coefficient likely has many significant neighboring coefficients. However, unlike wavelets, the shearlet transform offers a unique combination of properties. It uses directional elements at fine scales to capture elongated features and an isotropic father wavelet at coarse scales to capture low-frequency information. This combination of features makes the shearlet transform suitable for noise reduction tasks.

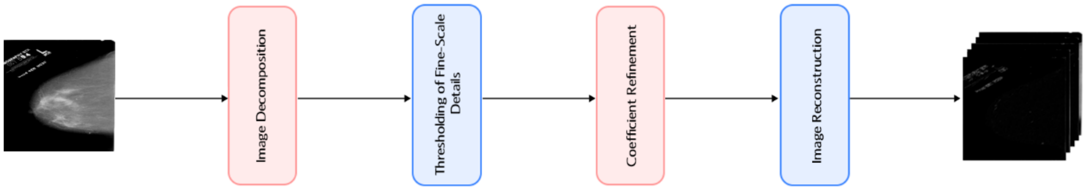


Figure 5. The block diagram of adaptive shrinkage denoising.

There are various types of noise reduction methods. They can generally be divided into two main categories: spatial filtering methods and transform domain filtering methods. Spatial filtering methods include median and Wiener filtering. Transform domain filtering methods include Fast Fourier Transform, VISUShrink [68], and BayesShrink [69]. To exploit denoising, a new local adaptive shrinkage thresholding approach is proposed. This approach utilizes the inherent correlation between neighboring shearlet coefficients. The main idea of this method lies in using the mean of the absolute values of shearlet coefficients within a local neighborhood. This local mean acts as a dynamic threshold for each coefficient. Coefficients above this significant threshold are considered to potentially contain valuable information. Conversely, coefficients that fall below the threshold are likely to be noise and are reduced towards zero [70]. The following section delves deeper into the architecture of the shearlet-based adaptive shrinkage method, which is depicted in Figure 5, and explains its implementation details:

1. **Image Decomposition** We apply the shearlet transform to the image to create coefficients denoted as C (j,l,m,n) . Here, j represents the scale level, l represents the directional information, and (m, n) are the spatial coordinates.

2. **Thresholding of Fine-Scale Details** To denoise the shearlet coefficients, we focus on the coefficients at specific scale levels (e.g., levels 2, 3, ... 5) to capture finer details. A new function, $\tilde{C}(\mathrm{j,l,m,n})$, is created by applying a mean filter to the absolute values of these coefficients. To compute $\tilde{C}(\mathrm{j,l,m,n})$ for each coefficient, a local neighborhood is considered, ensuring a balanced size (e.g., 3x3 or 5x5) around the coefficient. This neighborhood includes pixels above, below, to the left, and to the right of the target coefficient. The mean filter operation is shown by Equation 5.

$$\tilde{C}_{mean}\,(j,l,m,n) \;=\; \underset{(m,n)\,\in\,B}{\mathrm{mean}}\,|C\,(j,l,m,n)| \qquad \text{Equation 5}$$

Where B is the neighborhood window of the shearlet coefficient. In this neighborhood, two values are calculated: the mean value (e_{mean}(j,l)) of the modified coefficients $\tilde{C}(\mathrm{j,l,m,n})$ and the minimum value (M_{mean}(j,l)) . These values are then used to define a local adaptive threshold function based on the mean filter, as shown in Equation 6.

$$T_{mean}\,(j,l,m,n) \;=\; \left(1 \;-\; \frac{\tilde{C}_{mean}\,(j,l,m,n) \;-\; e_{mean}\,(j,l)}{\tilde{C}_{mean}\,(j,l,m,n) \;-\; M_{mean}\,(j,l)}\right) k\ \varepsilon_{j,l}\sigma \qquad \text{Equation 6}$$

$\varepsilon_{j,l}$ represents the mean distribution of white noise energy in the shearlet coefficient at scale j and direction l. σ is defined as the standard deviation of white noise, and k is a constant, often k = 3 or k = 4 for different scales.

3. **Coefficient Refinement** With the newly calculated thresholds for each coefficient, the method refines the coefficients using a hard thresholding approach, as shown by Equation 7.

$$\tilde{C}_k(\mathrm{j,l,m,n}) = \begin{cases} 0 & |C_k(\mathrm{j,l,m,n})| \;<\; T^{new}(j,l,m,n) \\ C_k & |C_k(\mathrm{j,l,m,n})| \;>\; T^{new}(j,l,m,n) \end{cases} \qquad \text{Equation 7}$$

4. **Image Reconstruction** Finally, to obtain the denoised image, the method performs an inverse transform using the refined coefficients ($\tilde{C}(\mathrm{j,l,m,n})$ after thresholding).

As mentioned, we used the shearlet transform into our model to enhance its performance by leveraging its advanced mathematical features. Unlike traditional wavelet transforms, shearlets provide superior directional sensitivity and multi-scale, multi-directional analysis, which is crucial for capturing edges and singularities in complex biomedical images like mammography [71]. Studies have shown that incorporating frequency domain features, such as those offered by shearlets, improves system performance, with Curvelet features outperforming wavelets [72]. Shearlets' ability to handle anisotropic features, their sparse representation, and excellent localization properties in both spatial and frequency domains make them particularly useful for precise edge and texture analysis, image compression, noise reduction, and other image processing tasks [73]. These characteristics led us to use shearlet subbands in our model, as illustrated by the CBIS-DDSM dataset image and subbands in Figure 6.

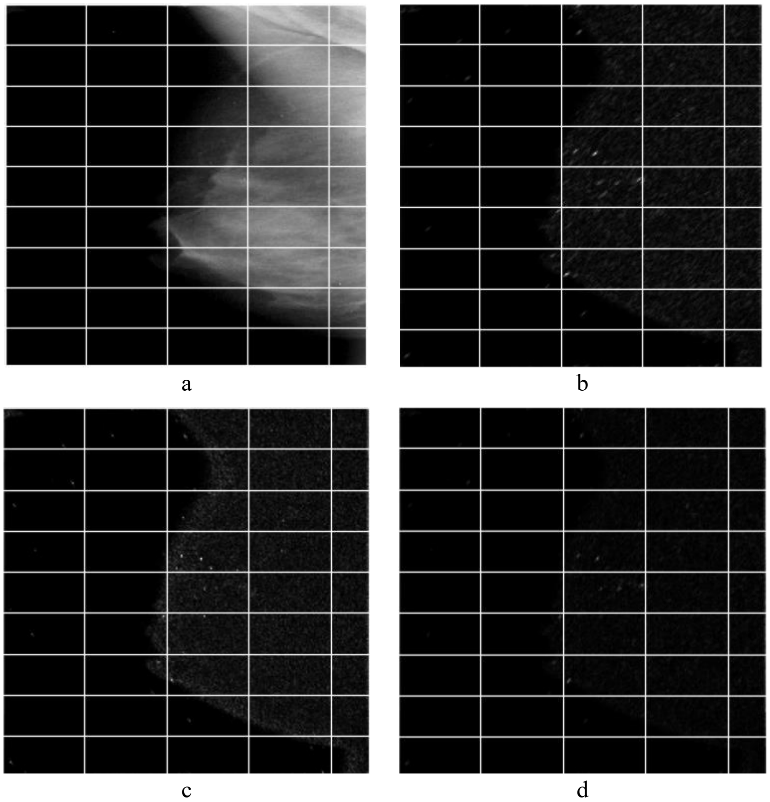


Figure 6. an image from the CBIS-DDSM dataset and several images of shearlet transform. a) CBIS-DDSM regular image b,c,d) the subbands of shearlet transform

3.3.Energy Branch

We have utilized the energy of shearlet images to enhance our model's structure. Using the energy of each subband in the discrete shearlet transform provides several benefits. it simplifies feature representation by summarizing data into scalar values, highlights significant patterns by capturing variations in magnitude, and offers robustness to noise and invariance to transformations like rotations. Energy features are effective in image and signal processing tasks by emphasizing important structural information and facilitating statistical analysis [74]. In the proposed model, the training mammography image is decomposed using a 2-level Discrete Shearlet Transform with different directions. The energy of each directional sub-band is then used as a feature for the corresponding training image. This feature extraction method is applied to all training samples, and the features are stored in a database, which is used as one of the inputs for the classification stage. The energy of each directional sub-band of image I is calculated using Equation 8:

$$Energy_e = \frac{1}{\mathrm{RC}} \sum_{\mathrm{i}=1}^{\mathrm{W}} \sum_{\mathrm{j}=1}^{\mathrm{H}} |I_e\ (\mathrm{i},\mathrm{j})| \qquad \text{Equation 8}$$

where I is the pixel value of the e-th sub-band, and W and H are the width and height of the sub-band, respectively. In this branch, we have only used the 16 high-frequency sub-bands and excluded the LL sub-band, which has minimal noise and maximum base structure. The reason for this is that the LL sub-band, due to its base structure, has relatively higher energy compared to other sub-bands, which can disrupt the regular and appropriately ranged energy data. Furthermore, experimental results have shown that excluding this sub-band improves the output metrics.

3.4.Attention Fusion

As mentioned before, in our proposed model, the attention module is used to combine and highlight the important features from the output of the two branches, spatial and shearlet. Among all the available attention modules, Squeeze-and-Excitation (SE) attention [39] is the most popular. Initially, we used this module for our model, but it did not yield satisfactory results. We also explored the attention fusion feature method [75], which also did not provide suitable outcomes. In these experiments, we concatenated the output features of the two branches and used the new feature vector as the input vector to this attention module. Based on these results, we realized that using the attention mechanism alone is not sufficient to achieve the best result. To achieve the best result, the two output vectors from the spatial and shearlet branches need to be processed to create a new feature vector derived from the processing of these two vectors. By doing this, we effectively reprocess the two output feature vectors together to achieve better features. The NAT layers shown in Figure 7, due to the presence of two modules, NAT attention and MLP, can process the input feature vector to the layer while also paying attention to more important features. Therefore, we decided to use NAT layers to combine and highlight the important features from the output vectors of the two branches. First, we concatenate the two feature vectors from the two branches. Then, we feed them into four NAT layers. This process simultaneously handles both spatial and shearlet data, highlighting the best features. Each of these layers normalizes the input data before entering the Neighborhood Attention module. In this module, the calculations for Neighborhood Attention are performed, and the final result is summed with the initial value. The result is then passed through a network of fully connected layers, where the output, having the same shape as the input of the fully connected layers, is added back to form the output of the NAT layer. Clearly, this integration processes the spatial and shearlet branches together, producing a suitable output for our system. These branches pass through four identical NAT layers and then an avgpool layer, creating our feature vector.

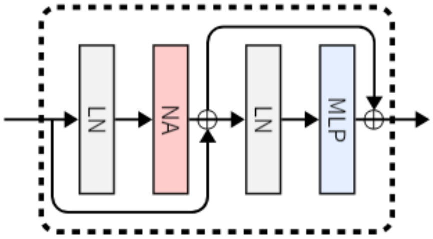


Figure 7. The structure of NAT layers [53].

3.5.Fully Connected Layers

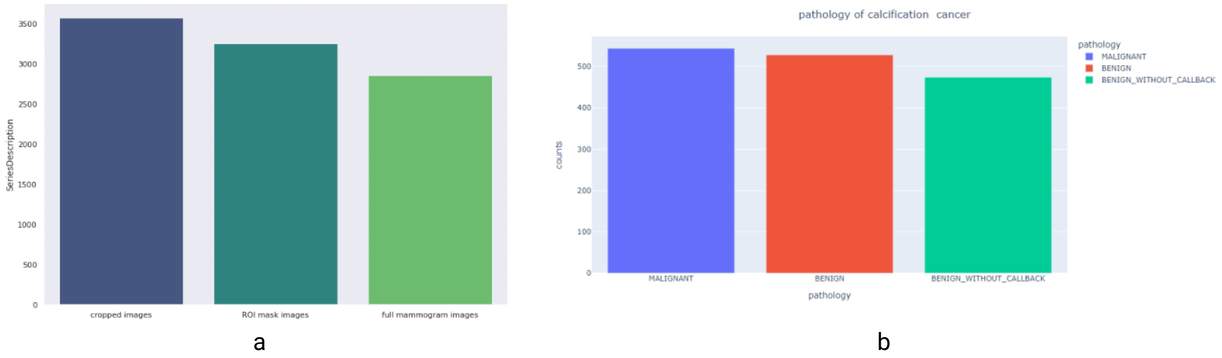
Next, the feature vector output from the attention module is concatenated with the energy feature vector. It is then fed into a network of fully connected layers. Our system employs three fully connected layers with ReLU activation functions. At the beginning of these layers, we use Dropout to minimize overfitting. This network consists of three layers with 512, 32, and 1 nodes, respectively. The final layer uses a Sigmoid activation function to ensure the neural network outputs interpretable probabilities, which is essential for binary classification tasks. This ensures that the output is within a useful range and that the network can effectively learn during training. The number of layers and nodes for each network was determined through trial and error, with the mentioned configurations providing the best results.

4. Experiments and results

This section examines the results, experiments, and related aspects. We will present the various experiments and evaluations.

4.1.Dataset

The dataset used in this study is CBIS-DDSM [23], a subset of 3D breast images sorted from the Digital Database for Screening Mammography (DDSM). CBIS-DDSM is an updated and standardized version of the DDSM database. It includes a subset of DDSM data selected and meticulously sorted by a trained mammographer. The images are compressed and converted to DICOM format. Additionally, new annotations for ROI segmentation and pathology diagnosis for training data are included in this set. The database comprises 6775 studies. In this study, there are 10239 images from 1566 patients, including full images, cropped images, and tumor masks. The database has three classes: 1. Benign 2. Malignant 3. Benign without callback. The term "benign without callback" in CBIS-DDSM refers to non-cancerous (benign) anomalies identified in breast imaging studies that do not require further action or follow-up based on the initial assessment. In medical imaging, "callback" typically indicates the need for additional tests, evaluations, or procedures to further assess suspicious findings. In the CBIS-DDSM dataset, which focuses on breast imaging for breast cancer detection and diagnosis, the "benign without callback" category usually includes cases where radiologists or medical professionals have identified anomalies in the breast images as non-cancerous and, therefore, no further follow-up or investigation is recommended. Figure 8 shows the chart related to the number of each type of image and the number for each class. In this experiment, we used the full mammogram images for analysis. This dataset contains two types of images: mass images and calcification images. After appropriate data preprocessing, we obtained 3103 images, including 1592 mass images and 1511 calcification images. We then converted the dataset classes from three to two categories by considering "benign without callback" as part of the benign class. The dataset was initially divided into two main sets for training and testing, with the training set comprising 1354 benign and 1104 malignant images, and the testing set comprising 385 benign and 260 malignant images, as detailed in Table 1. Since the testing data constitute 20% of the total data, we also split the training data into 80% for training and 20% for validation.



a

b

Figure 8. The specified number of each class and each type of image in the data. Image (a) shows the number of each image in the dataset and image (b) shows the number specified for each class in the dataset.

Table 1. The Number of data points for each class in the training and testing data for the model.

Split	Classes		All Data
	Malignant	Benign	
Train	1104	1354	2458
Test	260	385	645

4.2.Preprocessing

Deep learning methods and models have shown that the more data available, the better the output results [76]. This is especially true for transformer-based models, which are designed to be "data-hungry" and require large amounts of data for proper training [76]. Various studies have demonstrated that transformers, given ample data, outperform powerful and leading convolutional models [76]. In medical imaging, collecting real data is particularly challenging due to the sensitive nature of medical images, and data is often limited. This is even more pronounced in breast cancer imaging. Given these limitations, we decided to use data augmentation to diversify our dataset. We applied rotations, horizontal and vertical flipping, and random small window deletion from images. Additionally, we resized the input images to 224x224x3. We also experimented with CLAHE [77], but it did not produce satisfactory results.

4.3.Used Tools

In this paper, we utilized shared systems from Kaggle. Due to the extensive and lengthy nature of the experiments, we used these systems for an extended period. Typically, we employed two T4 GPUs in parallel. For implementation, we used the PyTorch framework. We also used the NATTEN framework [78] developed for the NAT model to enhance processing speed.

4.4.Evaluation Metrics

We used three evaluation metrics in this paper: the F1 score, the Kappa statistic, and the AUC. We did not use accuracy as a metric due to the class imbalance in the dataset. Using accuracy could misleadingly inflate or deflate the performance measure due to this imbalance. Therefore, we opted for the mentioned metrics. Note that TP stands for True Positive, TN for True Negative, FP for False Positive, and FN for False Negative.

$$\text{F1 Score} = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \qquad \text{Equation 9}$$

$$\text{Kappa} = \frac{2\ x\ (TP\ x\ TN - FN\ x\ FP)}{(TP + FP)\ x\ (FP + TN) + (TP + FN)\ x\ (FN + TN)} \qquad \text{Equation 10}$$

In machine learning, AUC (Area Under the Curve) refers to the area under the ROC curve (Receiver Operating Characteristic curve), which plots the true positive rate against the false positive rate at various threshold settings. To compute AUC, the predicted probabilities are sorted and used as thresholds for classifying samples, then the true positive and false positive rates are calculated for each threshold. The ROC curve is plotted, and the AUC is determined, often using the trapezoidal rule to approximate the integral. This value provides an overall measure of model performance, where an AUC of 1 indicates perfect classification and 0.5 indicates performance equivalent to random guessing.

4.5.Results

We explored various configurations, and the optimal model, DualShearNAT, is illustrated in Figure 2. Table 2 shows the comparison of our proposed method with other researched methods on the CBIS-DDSM dataset.

Table 2. Comparison of the proposed method with some existing methods.

Model	AUC	F1
Jaamour et al. [42]	-	0. 6648
Almeida et al. [80]	0.684	-
Shen et al. [12]	0.75	-
Quintana et al. [15]	0.809	-
Wei et al. [81]	0.7964	-
Petrini et al. [43]	0.8044	-
Sarker et al. [48]	0.6643	-
DualShearNAT	0.845	0.7680

In the study by Jaamour et al. [42], various convolutional models were examined. This study reviewed well-known convolutional neural networks including VGG19, ResNet50, InceptionV3, DenseNet121, and MobileNetV2, using weighted classes, different input sizes, several transfer learning techniques, and multiple datasets. Ultimately, they achieved F1 of 66.48% on the CBIS-DDSM dataset with the MobileNetV2 model. The study by Almeida et al [80]. was a comparative study that evaluated the performance of two XGBoost models against the VGG16 model. Additionally, this study examined whether it is better to use ROI images or full mammogram images. According to the results of this study, using full mammogram images yielded better results. Ultimately, this model achieved an AUC of 0.684. In the study by Quintana et al [15], researchers employed the DenseNet-121 model to optimize performance by experimenting with various patch sizes and resolution settings. Their methodical approach led to a significant achievement, culminating in an AUC score of 80.9%. This result underscores the effectiveness of DenseNet-121 in this context. In the study by Petrini et al [43], a parallel processing approach was utilized to handle the CBIS-DDSM dataset, which includes at least two images per breast: one from the side and one from the top. The study processed both images simultaneously using a pre-trained EfficientNet model, yielding two distinct results. With the default dataset division, the model achieved an AUC of 80.33%. However, when using a modified dataset division, the AUC improved to 84.83%. These findings demonstrate the impact of dataset division on model performance and the effectiveness of parallel processing with EfficientNet. In the study by Sarker et al. [48], researchers utilized a Swin-based structure that incorporated cross-attention mechanisms to enhance the integration of information from different views of each breast. This approach aimed to improve the coherent transfer of information between these views. The model was evaluated on both the CBIS-DDSM and VinDr datasets, achieving an AUC of 66.43% for the CBIS-DDSM data and an impressive 96.08% for the VinDr data. These results highlight the model's variable performance across different datasets, demonstrating its effectiveness in certain contexts while suggesting areas for further improvement in others. In the study by Shen et al. [12], initial training on patches of the DDSM dataset is performed using their convolutional models. This study employs ResNet50 and VGG16. After training the models, they are then trained on full mammogram images from the CBIS-DDSM dataset. Several approaches are used in this study. When using the official combined dataset, they achieved an AUC of 0.75. In the study by Wei et al. [81], a new framework called morphHR is introduced to improve the quality of transfer learning. In this framework, not only can the final layers be managed, but the initial layers of the model can also be modified. This study utilizes the ResNet model.

As explained in previous sections, we have introduced a new model named DualShearNAT, which consists of three branches for feature extraction. The first branch is the Spatial branch, where the full mammogram image with three color channels is input into the NAT model. In the second branch, we use images from the shearlet subbands. First, we apply a shearlet transform to the full mammogram images. Then, we use adaptive shrinkage denoising based on shearlets to remove noise. The obtained subbands are then input as a 17-channel image into another NAT network. The third branch utilizes the energy of the shearlet subband images. Finally, we combine the feature vectors from the spatial and shearlet branches using neighborhood attention. The output data are then fed along with the energy output data into MLP layers.

Additionally, to address the imbalance in the CBIS-DDSM dataset, we have used the focal loss function. A focal loss function addresses class imbalance during training in tasks by applying a modulating term to the cross-entropy loss [79] to focus learning on hard-to-classify samples. It is a dynamically scaled cross-entropy loss where the scaling factor decays to zero as confidence in the correct class increases [79]. Using all these techniques, we achieved an AUC of 0.845, which is significantly better compared to other methods discussed in this section.

4.6. Ablation Study

Here, we conduct an ablation study to evaluate the proposed method considering the following factors:
Comparison of different combinations of the spatial and shearlet feature vectors and examination of various backbones to achieve the most optimal model for the spatial branch

Comparison of Models Based on Feature Combination. In this section, we compare our feature fusion method with other methods used in various studies. The reason for combining features in such methods is that by combining different features, we can access better features that can be used for improved classification. There are various methods for combining features, and in this section, we compare some of these methods. In the concatenation method, we concatenate 2 feature vectors. First, we place the spatial feature vector, followed by the shearlet feature vector. In this combination, we transform two vectors of length 1024 into a single vector of length 2048. In the summation method, we add these 2 vectors element by element, transforming two vectors of length 1024 into a single vector of length 1024. However, these methods are not the best possible ways to combine 2 vectors. Attention-based methods can produce better results because the feature fusion module can make better decisions regarding the selection and combination of the 2 vectors, placing more importance on the more critical features. Therefore, we compare some attention-based methods. In the Attention Feature Fusion (AFF) method [75], we input 2 vectors into this module to be combined using the attention method. In the Squeeze-and-Excitation Networks (SE) [40] method, we first concatenate the 2 vectors and then input this vector into the SE module to emphasize the important features. In the NAT method, we also concatenate the 2 vectors and then input the resulting vector into 4 consecutive NAT blocks. Based on the results in Table 3, our proposed method significantly outperforms other methods.

This experiment demonstrates that feature vectors need to be processed together before being combined to create better features, as these vectors are generated from the spatial branch and the shearlet branch. In our NAT method, we use NAT blocks where input feature vectors are processed together. Additionally, using the NA module, these vectors receive neighborhood attention, prioritizing more critical features and enhancing the combined features.

Table 3. Results of various model experiments with different combinations.

Model	F1	AUC	Kappa
DualShearNAT with concatenation	0.736	0.812	0.443
DualShearNAT with summation	0.731	0.804	0.437
DualShearNAT with AFF	0.738	0.815	0.451
DualShearNAT with SE fusion	0.747	0.832	0.472
DualShearNAT with NAT fusion	0.768	0.845	0.509

Examination of various backbones for spatial branch. In this section, we compare our proposed NAT-based method for spatial branch with several other models. Specifically, we evaluate three models: VIT, Swin, and ConvNext. The VIT and Swin models, which are based on attention mechanisms, are used to compare neighborhood attention with self-attention-based approaches. Furthermore, ConvNext represents one of the most recent and robust convolution-based models, developed to compete with transformer-based models. Additionally, to ensure a consistent evaluation, we use the focal loss function to address the dataset imbalance. At this stage, we trained and tested the mentioned models using focal loss. The results of these experiments are shown in Table 4.

Table 4. Results of different models to process spatial branch.

Model	F1	AUC	Kappa
VIT [17]	0.691	0.734	0.376
Swin [52]	0.685	0.729	0.371
Convnext [56]	0.702	0.748	0.385
NAT [53]	0.716	0.762	0.414

According to the experimental results, the baseline NAT model achieved an AUC of 0.762, which is better than the other baseline models. The reason for this result can be found in the features of the NAT model. NAT, due to its use of neighborhood attention, can leverage local features similar to convolutional models, while also incorporating the global feature extraction capabilities of other transformer methods. In this approach, each pixel plays an important role in processing its neighboring pixels directly, and the receptive field expands without manual adjustments. Additionally, this model can maintain translational equivariance, much like convolutional methods. Notably, the NAT model achieves the best performance when compared to various models for the spatial branch. In our approach, we added the Shearlet and energy branches alongside the spatial branch, and by integrating the Shearlet transform and noise reduction techniques, our model attains an impressive AUC of 0.845.

5. Conclusion

In this paper, a novel method for breast cancer detection using mammography images, has been introduced. To extract more efficient features from mammography images, a novel model called DualShearNAT has been proposed. This model combines the features extracted from mammography images and the features extracted from shearlet domain images. Two parallel NAT models have been used to simultaneously process the original image and the shearlet-transformed images. Subsequently, we processed the feature vectors output from the two models using four additional NAT layers. The rationale for this approach is twofold: 1) Using this module, we aim to process the extracted features from different data sources together, achieving a combined processing of these two types of data for optimal results. 2) This module allows us to highlight more important features from the combined data processing and feature extraction. Following this, we concatenated these features with the energy data of the shearlet images, which contain more features, and fed the resulting vector into an MLP (multilayer perceptron) or fully connected network to perform the classification using this vector. One of the enhancement techniques employed in this research is the adaptive shrinkage threshold technique, applied to the input data to reduce the noise in mammography images using shearlet transforms. By using all these techniques, we achieved an AUC of 0.845, placing our model among the top-performing ones in prior research.

Despite the achievements of this research, there are still areas for improvement in this method. Exploring different shearlet levels, using NSST [82], extracting patches and applying the method to ROIs, and using other techniques similar to shearlet could be investigated to enhance this model. One challenge in using this method was the substantial resources required to run the research code, which, given our limited hardware capabilities, resulted in slow model processing—a noted weakness of this research. To further challenge the model, the proposed method could be tested on other datasets such as Mias [20], Inbreast [21], or even newer and more extensive datasets like VinDr [22]. One of the major issues with this research was the limited number of images available in this dataset, and running this model on datasets with a larger number of samples could potentially yield even better results.

References

1. National Breast Cancer Foundation. Retrieved from https://www.nationalbreastcancer.org/

2. The case for early detection. Nature Reviews Cancer, 3. https://doi.org/10.1038/nrc1041

3. American Cancer Society. Retrieved from https://www.cancer.org/

4. Siegel, R. L., Miller, K. D., Wagle, N. S., & Jemal, A. (2023). Cancer statistics, 2023. CA: A Cancer Journal for Clinicians, 73 (1). https://doi.org/10.3322/caac.21763

5. Mammography. Retrieved from https://www.nibib.nih.gov/science-education/science-topics/mammography

6. Sickles, E., d'Orsi, C., Bassett, L., Appleton, C., Berg, W., Burnside, E., Feig, S., Gavenonis, S., Newell, M., & Trinh, M. (2013). ACR BI-RADS® Atlas Breast Imaging Reporting and Data System. American College of Radiology.

7. Warren, R., & Duffy, W. (1995). Comparison of single reading with double reading of mammograms, and change in effectiveness with experience. *The British Journal of Radiology*, 68(813), 958–962.

8. Agrawal, S., Rangnekar, R., Gala, D., Paul, S., & Kalbande, D. (2018). Detection of breast cancer from mammograms using a hybrid approach of deep learning and linear classification. In *2018 International Conference on Smart City and Emerging Technology (ICSCET)* (pp. 1-6). IEEE. https://doi.org/10.1109/ICSCET.2018.8537250

9. Eremici, I., Borlea, A., Dumitru, C., & Stoian, D. (2024). Factors associated with false positive breast cancer results in the real-time sonoelastography evaluation of solid breast lesions. *Medicina, 60*(1023). https://doi.org/10.3390/medicina60071023

10. Abdelrahman, L., Al Ghamdi, M., Collado-Mesa, F., & Abdel-Mottaleb, M. (2021). Convolutional neural networks for breast cancer detection in mammography: A survey. *Computerized Medical Imaging and Graphics*, 131, 104248. https://doi.org/10.1016/j.compbiomed.2021.104248

11. Desai, M., & Shah, M. (2021). An anatomization on breast cancer detection and diagnosis employing multi-layer perceptron neural network (MLP) and convolutional neural network (CNN). *Clinical eHealth, 4*. https://doi.org/10.1016/j.ceh.2020.11.002

12. Shen, L., Margolies, L. R., Rothstein, J. H., Fluder, E., McBride, R., & Sieh, W. (2019). Deep learning to improve breast cancer detection on screening mammography. *Scientific Reports, 9*, 12495. https://doi.org/10.1038/s41598-019-48995-4

13. Redmon, J., Divvala, S. K., Girshick, R., & Farhadi, A.. (2016, June 27). *You Only Look Once: Unified, Real-Time Object Detection*. https://doi.org/10.1109/CVPR.2016.91

14. Ayana, G., Park, J., & Choe, S. W. (2022). Patchless multi-stage transfer learning for improved mammographic breast mass classification. *Cancers, 14*(5), 1280. https://doi.org/10.3390/cancers14051280

15. Quintana, G. I., Vancamberg, L., Mougeot, M., Desolneux, A., & Muller, S. (2023). Exploiting patch sizes and resolutions for multi-scale deep learning in mammogram image classification. *Bioengineering, 10*(5), 534. https://doi.org/10.3390/bioengineering10050534

16. Hatamizadeh, A., Yang, D., Roth, H., & Xu, D. (2021). UNETR: Transformers for 3D medical image segmentation. *arXiv preprint arXiv:2103.10504*.

17. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021, May 3). An image is worth 16x16 words: Transformers for image recognition at scale.

18. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (Vol. 30).

19. Hussain, D., & Hyeon Gu, Y. (2024). Exploring the impact of noise and image quality on deep learning performance in DXA images. *Diagnostics, 14*(13), 1328. https://doi.org/10.3390/diagnostics14131328

20. Suckling, J., Parker, J., Dance, D., Astley, S., Hutt, I., Boggis, C., Ricketts, I., Stamatakis, E., Cerneaz, N., Kok, S., Taylor, P., Betal, D., &amp; Savage, J. (2015). Mammographic Image Analysis Society (MIAS) database v1.21. Apollo - University of Cambridge Repository. https://doi.org/10.17863/CAM.105113

21. Moreira, I. C., Amaral, I., Domingues, I., Cardoso, A., Cardoso, M. J., & Cardoso, J. S. (2012). INbreast: toward a full-field digital mammographic database. Academic radiology, 19(2), 236–248. https://doi.org/10.1016/j.acra.2011.09.014

22. Nguyen, H.T., Nguyen, H.Q., Pham, H.H. et al. (2023). VinDr-Mammo: A large-scale benchmark dataset for computer-aided diagnosis in full-field digital mammography. Sci Data 10, 277 https://doi.org/10.1038/s41597-023-02100-7

23. Rebecca Sawyer Lee, Francisco Gimenez, Assaf Hoogi , Daniel Rubin. (2016). Curated Breast Imaging Subset of DDSM . The Cancer Imaging Archive. https://doi.org/10.7937/K9/TCIA.2016.7O02S9CY

24. Král, P., & Lenc, L. (2016). LBP features for breast cancer detection. In Proceedings of the International Conference on Image Processing (ICIP) , 2643-2647.

25. Hadid, A. (2008). The local binary pattern approach and its applications to face analysis. In Proceedings of the First Workshops on Image Processing Theory, Tools and Applications (IPTA) , 1-9. https://doi.org/10.1109/IPTA.2008.4743795

26. Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., & Scholkopf, B. (1998). Support vector machines. IEEE Intelligent Systems and Their Applications, 13 (4), 18-28. https://doi.org/10.1109/5254.708428

27. Buciu, I., & Gacsadi, A. (2011). Directional features for automatic tumor classification of mammogram images. Biomedical Signal Processing and Control, 6 , 370-378.

28. Jiang, W., Shen, T., Zhang, J., Hu, Y., & Wang, X.-Y. (2008). Gabor wavelets for image processing. In Proceedings of the 2008 International Conference on Computer and Computational Sciences (CCCM) , 1. https://doi.org/10.1109/CCCM.2008.62

29. Maćkiewicz, A., & Ratajczak, W. (1993). Principal components analysis (PCA). Computers & Geosciences, 19 (3), 303-342. https://doi.org/10.1016/0098-3004(93)90090-R

30. Pratiwi, M., Alexander, J., Harefa, S. Nanda. (2015). Mammograms classification using gray-level co-occurrence matrix and radial basis function neural network. Procedia Computer Science, 59 , 83-91.

31. El Houby, E. M. F., & Yassin, N. I. R. (2021). Malignant and nonmalignant classification of breast lesions in mammograms using convolutional neural networks. Biomedical Signal Processing and Control, 70 . https://doi.org/10.1016/j.bspc.2021.102954

32. Macit, H. S., & Sabanci, K. (2023). Benchmarking of ResNet models for breast cancer diagnosis using mammographic images. International Journal of Applied Methods in Electronics and Computers, 11 (3), 128–133. https://doi.org/10.58190/ijamec.2023.39

33. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) .

34. Malebary, S. J., & Hashmi, A. (2021). Automated breast mass classification system using deep learning and ensemble learning in digital mammogram. IEEE Access, 9 , 55312-55328. https://doi.org/10.1109/ACCESS.2021.3071297

35. Mahmood, T., Li, J., Pei, Y., Akhtar, F., Rehman, M. U., & Wasti, S. H. (2022). Breast lesions classifications of mammographic images using a deep convolutional neural network-based approach. PLoS ONE, 17 (1). https://doi.org/10.1371/journal.pone.0263126

36. Soriano, D., Aguilar, C., Ramirez-Morales, I., Tusa, E., & Rivas, W. (2018). Mammogram classification schemes by using convolutional neural networks. In Botto-Tobar, M., Esparza-Cruz, N., León-Acurio, J., Crespo-Torres, N., & Beltrán-Mora, M. (Eds.), Technology Trends. CITT 2017. Communications in Computer and Information Science, vol 798 . Springer, Cham. https://doi.org/10.1007/978-3-319-72727-1_6

37. Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. Journal of Machine Learning Research, 13 , 281-305. https://doi.org/10.5555/2188385.2188395

38. Thwin, S. M., Malebary, S. J., Abulfaraj, A. W., & Park, H.-S. (2024). Attention-based ensemble network for effective breast cancer classification over benchmarks. Technologies, 12 (2), 16. https://doi.org/10.3390/technologies12020016

39. Hu, J., Shen, L., Albanie, S., Sun, G., & Wu, E. (2018). Squeeze-and-excitation networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42 (8). https://doi.org/10.1109/TPAMI.2019.2913372

40. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) . https://doi.org/10.1109/CVPR.2016.308

41. Oza, P., Sharma, P., & Patel, S. (2023). Deep ensemble transfer learning-based framework for mammographic image classification. Journal of Supercomputing, 79 , 8048–8069. https://doi.org/10.1007/s11227-022-04992-5

42. Jaamour, A., Myles, C., Patel, A., Chen, S.-J., McMillan, L., & Harris-Birtill, D. (2023). A divide and conquer approach to maximise deep learning mammography classification accuracies.

43. Petrini, D. G. P., Shimizu, C., Roela, R. A., Valente, G. V., Folgueira, M. A. A. K., & Kim, H. Y. (2022). Breast cancer diagnosis in two-view mammography using end-to-end trained EfficientNet-based convolutional network.

44. Borah, N., Varma, P. S. P., Datta, A., Kumar, A., Baruah, U., & Ghosal, P. (2022). Performance analysis of breast cancer classification from mammogram images using vision transformer. In Proceedings of the 2022 IEEE Calcutta Conference (CALCON) , 238-243. https://doi.org/10.1109/CALCON56258.2022.10060315

45. Panambur, A. B., Yu, H., Bhat, S., Madhu, P., Bayer, S., & Maier, A. (2024). Attention-guided erasing. In Maier, A., Deserno, T. M., Handels, H., Maier-Hein, K., Palm, C., & Tolxdorff, T. (Eds.), Bildverarbeitung für die Medizin 2024. Informatik aktuell . Springer Vieweg, Wiesbaden. https://doi.org/10.1007/978-3-658-44037-4_8

46. Al-Tam, R. M., Al-Hejri, A. M., Narangale, S. M., Samee, N. A., Mahmoud, N. F., Al-masni, M. A., & Al-antari, M. A. (2022). A hybrid workflow of residual convolutional transformer encoder for breast cancer classification using digital x-ray mammograms. Biomedicines, 10 (11), 2971. https://doi.org/10.3390/biomedicines10112971

47. Chen, X., Zhang, K., Abdoli, N., Gilley, P. W., Wang, X., Liu, H., Zheng, B., & Qiu, Y. (2022). Transformers improve breast cancer diagnosis from unregistered multi-view mammograms. Diagnostics, 12 (7), 1549. https://doi.org/10.3390/diagnostics12071549

48. Sarker, S., Sarker, P., Bebis, G., & Tavakkoli, A. (2024). MV-Swin-T: Mammogram classification with multi-view Swin Transformer.

49. Ali, J. A., & Janet, J. P. (2013). Discrete shearlet transform-based classification of microcalcification in digital mammograms.

50. Gonzalez, R. C., & Woods, R. E. (2008). Digital image processing. Prentice Hall.

51. Donoho, D. L. (1995). De-noising by soft-thresholding. IEEE Transactions on Information Theory, 41 (3), 613-627.

52. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021, March 25). Swin transformer: Hierarchical vision transformer using shifted windows.

53. Hassani, A., Walton, S., Li, J., Li, S., & Shi, H. (2022). Neighborhood attention transformer. arXiv preprint arXiv:2204.07143 . https://doi.org/10.48550/arXiv.2204.07143

54. Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., Liang, D., Lu, T., Luo, P., & Shao, L. (2021, February 24). Pyramid vision transformer: A versatile backbone for dense prediction without convolutions.

55. Zhang, P., Dai, X., Yang, J., Xiao, B., Yuan, L., Zhang, L., & Gao, J. (2021). Multi-scale vision longformer: A new vision transformer for high-resolution image encoding. In IEEE/CVF International Conference on Computer Vision (ICCV) .

56. Liu, Z., Mao, H., Wu, C., Feichtenhofer, C., Darrell, T., & Xie, S. (2022, January 10). A ConvNet for the 2020s. https://doi.org/10.1109/CVPR52688.2022.01167

57. Deng, J., Dong, W., Socher, R., Li, L. -J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA (pp. 248-255).

58. Vafaie, S., & Salajegheh, E.. (2023). A Comparative Study of Shearlet, Wavelet, Laplacian Pyramid, Curvelet, and Contourlet Transform to Defect Detection. *Journal of Soft Computing in Civil Engineering*, *7*. https://doi.org/10.22115/scce.2023.356475.1505

59. Lim, W.-Q. (2010). The discrete shearlet transform: A new directional transform and compactly supported shearlet frames. IEEE Transactions on Image Processing, 19 (5), 1166-1180. https://doi.org/10.1109/TIP.2010.2041410

60. Duval-Poo, M. A., Odone, F., & De Vito, E. (2015). Edges and corners with shearlets. IEEE Transactions on Image Processing, 24 (11), 3768-3780. https://doi.org/10.1109/TIP.2015.2451175

61. Meshkini, K., & Ghassemian, H. (2017). Texture classification using shearlet transform and GLCM. In 2017 Iranian Conference on Electrical Engineering (ICEE) (pp. 1845-1850). IEEE. https://doi.org/10.1109/IranianCEE.2017.7985354

62. Guo, K., Labate, D., & Rodriguez Ayllon, J. P. (2020). Image inpainting using sparse multiscale representations: Image recovery performance guarantees. Applied and Computational Harmonic Analysis, 49 (2). https://doi.org/10.1016/J.ACHA.2020.05.001

63. Zhou, S., Shi, J., Zhu, J., Cai, Y., & Wang, R. (2013). Shearlet-based texture feature extraction for classification of breast tumor in ultrasound image. Biomedical Signal Processing and Control, 8 (6), 688-696.

64. Yang, H. Y., Wang, X. Y., Niu, P. P., & Liu, Y. C. (2014). Image denoising using nonsubsampled shearlet transform and twin support vector machines. Neural Networks, 57 , 152-165.

65. Ziani, C., & Sadiq, A. (2020). SH-CNN: Shearlet convolutional neural network for gender classification. Advances in Science, Technology and Engineering Systems Journal, 5 (6), 158.

66. Thayammal, S., & Selvathi, D. (2012). Image compression using multidirectional anisotropic transform: Shearlet transform. In 2012 International Conference on Devices, Circuits and Systems (ICDCS) (pp. 254-258). IEEE. https://doi.org/10.1109/ICDCSyst.2012.6188739

67. Michael, P. F., & Yoon, H.-J. (2020). Survey of image denoising methods for medical image classification.

68. Donoho, D. L., & Johnstone, I. M. (1994). Ideal spatial adaptation via wavelet shrinkage. Biometrika, 81 (3), 425-455.

69. Simoncelli, E. P., & Adelson, E. H. (1996). Noise removal via Bayesian wavelet coring. In Third International Conference on Image Processing (ICIP) (Vol. 1, pp. 379-382). IEEE Signal Processing Society.

70. Chen, X., Deng, C., & Wang, S. (2010). Shearlet-based adaptive shrinkage threshold for image denoising. In 2010 International Conference on E-Business and E-Government (ICEE) (pp. 1616-1619). IEEE. https://doi.org/10.1109/ICEE.2010.409

71. Olaide, O., & Ezugwu, A. (2022). A novel wavelet decomposition and transformation convolutional neural network with data augmentation for breast cancer detection using digital mammogram. Scientific Reports, 12 . https://doi.org/10.1038/s41598-022-09905-3

72. Kanagaraj, K., & Priya, G. G. L. (2018). Curvelet transform based feature extraction and selection for multimedia event classification. https://doi.org/10.1016/J.JKSUCI.2018.11.006

73. Kumar, D., Pandey, R. C., & Mishra, A. K. (2024). A review of image features extraction techniques and their applications in image forensic. Multimedia Tools and Applications . https://doi.org/10.1007/s11042-023-17950-x

74. Kanchana, M., & Varalakshmi, P.. (2012). *Texture Classification Using Discrete Shearlet Transform*. *2*(6). https://doi.org/10.15373/22778179/JUNE2013/61

75. Dai, Y., Gieseke, F., Oehmcke, S., Wu, Y., & Barnard, K. (2021, January 1). Attentional feature fusion. https://doi.org/10.1109/WACV48630.2021.00360

76. Deininger, L., Stimpel, B., Yüce, A., Abbasi-Sureshjani, S., Schönenberger, S., Ocampo, P. S., Korski, K., & Gaire, F. (2022). A comparative study between vision transformers and CNNs in digital pathology. arXiv . https://doi.org/10.48550/arXiv.2206.00389

77. Zuiderveld, K. (1994). Contrast limited adaptive histogram equalization. In P. Heckbert (Ed.), Graphics Gems IV . Academic Press.

78. Hassani, A., Hwu, W.-M., & Shi, H. (2024, March 7). Faster neighborhood attention: Reducing the O(n^2) cost of self attention at the threadblock level.

79. Huang, C., Li, Y., Loy, C. C., & Tang, X. (2016). Learning deep representation for imbalanced classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 5375–5384).

80. Almeida, R., Chen, D., Filho, A., & Brandão, W. (2021). Machine learning algorithms for breast cancer detection in mammography images: A comparative study. In Proceedings of the 23rd International Conference on Enterprise Information Systems (pp. 660–667).

81. Wei, T., Aviles-Rivero, A. I., Wang, S., Huang, Y., Gilbert, F. J., Schönlieb, C.-B., & Chen, C. W. (2021). Beyond fine-tuning: Classifying high-resolution mammograms using function-preserving transformations. arXiv . https://arxiv.org/abs/2101.07945

82. Da Cunha, A. L., Zhou, J., & Do, M. N. (2006). The nonsubsampled contourlet transform: Theory, design, and applications. IEEE Transactions on Image Processing, 15 (10), 3089–3101. https://doi.org/10.1109/TIP.2006.877507