BİTLİS EREN ÜNİVERSİTESİ
FEN BİLİMLERİ DERGİSİ

ISSN: 2147-3129 / e-ISSN: 2147-3188

Article Type	: Research Article	**Year**	: 2025
Received	: November 12, 2024	**Volume**	: 14
Revised	: February 25, 2025	**Issue**	: 1
Accepted	: March 6, 2025	**Pages**	: 287-313
DOI	: 10.17798/bitlisfen.1583948		

EVALUATING VISION TRANSFORMER MODELS FOR BREAST CANCER DETECTION IN MAMMOGRAPHIC IMAGING

Uğur DEMIROĞLU [1], **Bilal ŞENOL** [2 *]

[1] Kahramanmaraş İstiklal University, Software Engineering Department, Kahramanmaraş, Türkiye

[2] Aksaray University, Software Engineering Department, Aksaray, Türkiye

**** Corresponding Author:*** *ugurdemiroglu@istiklal.edu.tr*

ABSTRACT

Breast cancer is a leading cause of mortality among women, with early detection being crucial for effective treatment. Mammographic analysis, particularly the identification and classification of breast masses, plays a crucial role in early diagnosis. Recent advancements in deep learning, particularly Vision Transformers (ViTs), have shown significant potential in image classification tasks across various domains, including medical imaging. This study evaluates the performance of different Vision Transformer (ViT) models—specifically, base-16, small-16, and tiny-16—on a dataset of breast mammography images with masses. We perform a comparative analysis of these ViT models to determine their effectiveness in classifying mammographic images. By leveraging the self-attention mechanism of ViTs, our approach addresses the challenges posed by complex mammographic textures and low contrast in medical imaging. The experimental results provide insights into the strengths and limitations of each ViT model configuration, contributing to an informed selection of architectures for breast mass classification tasks in mammography. This research underscores the potential of ViTs in enhancing diagnostic accuracy and serves as a benchmark for future exploration of transformer-based architectures in the field of medical image classification.

Keywords: Breast mammography with masses, Image classification, Vision transformers, base-16, small-16, tiny-16.

1 INTRODUCTION

Cancer is a leading cause of death worldwide; one in two people diagnosed with cancer will require treatment, and early detection is the best method of preventing the progression of the disease to a later stage. More than one million cases of breast cancer are diagnosed each year, and despite increased survival rates, it remains the leading cause of death among women [1]. The small size of malignant masses has been shown to correlate with treatment success, making early diagnosis treatments such as mammography critical for improved long-term survival and quality of life [2]. Historically, most women would not become aware of their cancer until later stages of tumorigenesis due to the delayed onset of symptoms or would actively avoid detection as societal attitudes were against early mastectomy [3]. Current cancer detection guidelines have begun to promote public health education on their necessity; with increased awareness of need comes the resources to support potential candidates. A significant portion of individuals fail to access treatment due to systemic inability to pay for or seek care, a symptom of false patient belief or clinician default that mammography results in frequent benign biopsy and consequently low patient satisfaction [4]. Despite these informational hurdles, mammography remains our gold standard in breast screening, potentially changing long-term prognoses with easy attainability today. The accessibility of mammography has risked malignant tumor diagnosis, allowing those under poverty to live with progressive disease until care is both useful and affordable. The message of our work is not to catalog a list of diagnoses lost to accessibility, but instead to reinforce the idea that education on available services can potentially lead to lifestyle changes that drastically improve public health. Recommending yearly checks, even with breast self-exams, can increase the 5-year survival rate by 94% for early detection of tumor outgrowth and subsequent apoptosis [5]. Our increasing knowledge base on heterogenic carriers and specific risks, coupled with decreased test invasiveness, could ensure increased preventive lifestyle changes. Early cancer detection guidelines aimed at potentially affected groups of the general public could encourage preventive lifestyle changes in those on the brink, who cannot or choose not to earn access to screening [6].

Breast cancer is a major cause of mortality worldwide. To survive, affected individuals must receive timely treatment. Early diagnosis is also associated with reduced treatment toxicity and healthcare costs. Because breast-focused physical examination alone often fails to detect small lesions early on, imaging technologies have been utilized in cancer screening. Among these techniques, mammography has the strongest supporting evidence [7]. Mammography was first adopted for convenient use in healthcare screening in the 1960s. Both the sensitivity and specificity of the imaging machines have gradually improved. Rates of interval cancers and those detected beyond the screen-detected tumor have also marginally decreased. Widespread mammographic screening has thus been adopted in many countries in some form [8]. Because mammography is performed in a private area and may have a variety of outcomes, it might be intimidating. It is essential to have a basic idea of what mammography entails and how it should be performed before attending your appointed days [9]. Traditional mammography machines help discover breast cancer as they employ a functioning X-ray system that allows them to detect abnormalities in the breast, such as tumors, before the patient or doctor notices them. Understanding what takes place when you receive a mammogram can help you make an informed decision about your breast health with the information you have. Inform your doctor if you have breast implants or have been diagnosed with breast cancer [10].

Mammography is an imaging technique focused on breast composition to screen for breast cancer and is widely considered the gold standard in the investigation for detecting early breast cancer, in addition to its important role in diagnostic evaluation [11]. Breast tomosynthesis is essentially an advanced type of mammography that creates three-dimensional images of the breast from a two-dimensional radiograph image. The purpose of mammography is to provide detailed images of the breast by passing a very low dose of radiation through the tissue. Mammography can detect tumors that are not easily felt. It can also identify some non-cancerous abnormalities, which surgeons may review to know if biopsies are required in the future [12]. A screening mammogram is part of regular healthcare. This test is designed to detect early signs of breast cancer in women who do not display clinical symptoms or signs of breast disease. Diagnostic mammography, on the other hand, is used to investigate tissue changes that were detected as a result of a screening mammography or not, or following clinical and/or self-exam detection [13]. Recently, computer-aided techniques developed to classify mammogram images gained a significant place in machine learning world. There can be found numerous methods proposed in this regard. This paper implements the Vision Transformers (ViTs) for this issue.

Deep learning has made a breakthrough in many fields, especially in computer vision, where convolutional networks have played a major role. Recent years have shown a shift from traditional convolutional networks to transformer-based approaches, mostly in language processing, as they have outperformed benchmark datasets [14]. ViTs have shown the ability to capture spatial information effectively, replacing convolutional networks. Vision Transformers have shown the ability to capture spatial information effectively, replacing convolutional networks. The extensive analysis of Vision Transformer models and their workings will help us improve the performance of such transformer-based models further [15]. Traditional

Convolutional Neural Networks (CNNs) have shown impressive results in a variety of tasks, from image classification to object segmentation. CNNs are popular in image classification tasks because they are translationally invariant and have a compact representation of the input image at each layer. For instance, classical networks for image classification may capture intricate hierarchical relationships, but they are entirely dependent on convolutional and max-pool layers, which limit the size of the receptive field. Large images are difficult to process, leading to an increased number of layers and parameters, which may result in computational inefficiency. Large stride values lose a lot of local features [16]. The transformer architecture is a multi-head self-attention mechanism with various layers. Interest in the transformer has increased with respect to computer vision applications' best pre-trained model. However, the transformer architecture is completely image-agnostic, which means it can stitch data of any kind [17]. More information about the architecture is provided in the further sections.

Mammogram image classification is one of the predominant approaches to detect breast cancer. The classification is either performed within different categories of tumors or between different types of tumors. In our study, we have differentiated the mammogram images of a general mammogram dataset into three classes: benign, malignant, and normal. In the modern era, any model requires several types of enhancements to classify a complex real-life dataset with high accuracy and minimal time complexity. Therefore, one of the state-of-the-art models suffers from this drawback. Thus, in our study, we have performed a comparative analysis between various models.

ViTs have recently gained attention in medical image analysis, offering an alternative to traditional CNNs by capturing long-range dependencies through self-attention mechanisms. Studies have demonstrated that ViTs can achieve performance comparable to or even surpassing that of CNNs in breast cancer detection tasks, as evidenced by research applying ViTs to classify breast ultrasound images with promising results [18]. However, CNNs remain prevalent in medical imaging due to their efficiency and strong inductive biases, which are advantageous for learning spatial hierarchies in complex medical datasets [19]. To leverage the strengths of both architectures, hybrid models that combine CNN-based feature extraction with transformer-based self-attention have been proposed. These hybrid approaches have been applied in various studies, such as one that integrated a convolutional backbone with transformer layers to enhance feature representation in histopathological images [20]. Another study proposed a token-mixer hybrid architecture, demonstrating improved diagnostic accuracy in breast cancer classification by effectively balancing local and global feature extraction [21]. Additionally, recent investigations have explored novel training strategies and data augmentation techniques that further boost the performance of hybrid models in medical image classification tasks [22]. Overall, the integration of transformer-based architectures with traditional CNNs not only enhances diagnostic performance but also provides a flexible framework that can adapt to different imaging modalities and clinical requirements, marking a significant step forward in the evolution of automated breast cancer detection systems.

Recent advancements in ViT models have significantly enhanced breast cancer detection in medical imaging. For instance, a study introduced a ViT-based transfer learning method for breast mass classification, achieving an impressive area under the curve (AUC) of 1.0 on both ultrasound and mammogram datasets, thereby outperforming traditional CNN-based approaches [23]. Another innovative approach, the TokenMixer hybrid architecture, combines CNNs and ViTs to improve feature extraction and classification accuracy in histopathological image analysis, effectively balancing local and global feature representations [24]. Additionally, the NHS has launched the world's largest trial of AI for breast cancer diagnosis, aiming to expedite detection by using AI to analyze a significant portion of mammograms, potentially reducing the workload on radiologists and decreasing patient wait times [25]. These developments underscore the potential of ViT-based models and AI integration in enhancing the accuracy and efficiency of breast cancer diagnostics.

The article contributes significantly to the literature by providing a comprehensive evaluation of ViT models for breast cancer detection in mammographic imaging. Unlike traditional CNNs, which rely on local receptive fields, the study highlights the effectiveness of ViTs in capturing long-range dependencies and complex patterns in mammographic textures through self-attention mechanisms. By assessing three different ViT configurations—Base-16, Small-16, and Tiny-16—the research offers a comparative analysis of model performance, considering factors such as accuracy, computational efficiency, and training time. The findings indicate that while the Small-16 model achieves the highest accuracy, the Tiny-16 model provides a computationally efficient alternative with moderate performance. This study underscores the importance of selecting appropriate model architectures based on computational resources and diagnostic accuracy requirements. Furthermore, by leveraging a publicly available mammographic dataset and implementing a standardized preprocessing pipeline—including resizing, normalization, and Contrast Limited Adaptive Histogram Equalization (CLAHE)—the study ensures reproducibility and robustness in medical image classification. The research also contributes to the broader understanding of transformer-based architectures in medical imaging, positioning ViTs as a viable alternative to CNNs for automated breast cancer diagnosis. This work serves as a benchmark for future studies, encouraging further exploration of ViTs and hybrid deep learning approaches to enhance diagnostic accuracy in medical applications.

This paper implements a comprehensive analysis of the classification structure of ViTs. The three sub-models in the ViTs architecture which are the base-16, small-16 and tiny-16 models have been separately considered for classifying mammographs. Thus, a detailed analysis of ViTs and its sub-models are discussed. In the paper next section presents the dataset and the third section gives information about the ViTs. The case study is shown in the fourth section and the last section includes the conclusions.

2 THE DATASET OF MAMMOGRAPHY

The *Breast Mammography Images with Masses* dataset, available through the Digital Object Identifier (DOI) 10.17632/ywsbh3ndr8.2, is an essential asset for professionals and researchers focused on medical imaging, breast cancer detection, and computer-aided diagnostic systems [26]. This comprehensive dataset comprises mammographic images that contain masses, which are critical indicators in assessing the likelihood of breast cancer. The images provide high-resolution visual information that supports the differentiation between benign and malignant masses, enhancing diagnostic accuracy. Each image in the dataset is carefully labeled and categorized, supporting diverse research applications such as mass detection, segmentation, and classification. This thorough annotation allows researchers to utilize the dataset effectively in machine learning contexts, where high-quality labeled data is essential for training algorithms designed to identify early-stage breast cancer. Such annotated data proves particularly valuable for developing and testing deep learning models in diagnostic radiology, where advancements in automated mass detection and classification can make a substantial impact on early diagnosis and patient outcomes. Figure 1 contains sample images from the dataset.

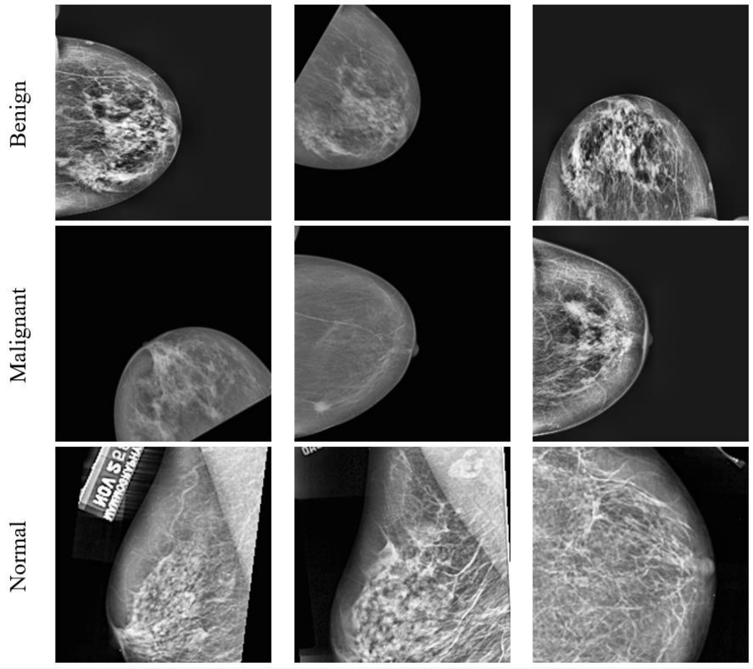

Figure 1. Sample images from the dataset.

The dataset is freely accessible for public use, making it a critical resource for both academic and industry researchers committed to advancing breast cancer detection and diagnosis technologies. Its availability allows a broad community of researchers to contribute to the development of new methodologies and models, ultimately aiding the early detection and treatment of breast cancer.

The dataset is a structured collection designed for cancer detection and classification, containing approximately 625MB of data across 26,602 images. These images are divided into three main categories—benign, malignant, and normal—organized in subfolders corresponding to each class. Specifically, the dataset includes 10,866 images labeled as benign, 13,710 as malignant, and 2,026 as normal. Each image is 8 bits deep, primarily in PNG format, with a minimum resolution of 227x227 pixels, providing adequate detail for analysis. The PNG format used for these images supports clear labeling of each sample as benign, malignant, or normal, which is essential for training machine learning models. In addition, the dataset is publicly licensed, making it widely accessible and frequently utilized in medicine, oncology research, and computer vision applications. This open access allows researchers and developers uninterrupted downloading and use, supporting diverse initiatives in medical imaging. For deep learning model development, the dataset underwent further enhancements. After being resized to standardized dimensions, the images were enhanced using CLAHE, a technique that improves image contrast by adjusting local histogram intensities. This preprocessing step enhances image quality, making it suitable for machine learning applications where subtle differences in tissue appearance are critical for model accuracy. The processed dataset thus serves as a valuable foundation for developing, testing, and refining deep learning models focused on breast cancer detection and classification. While this dataset provides high-resolution images with well-annotated benign, malignant, and normal cases, we acknowledge the importance of assessing its applicability across different populations. The dataset primarily consists of images collected under specific clinical conditions, which may not fully represent the diversity of real-world patient populations, including variations in age, ethnicity, breast density, and imaging protocols. Additionally, potential sources of error in the dataset could arise from factors such as label inaccuracies, imaging artifacts, or biases introduced during data collection and annotation. For example, mammograms from different devices and institutions may exhibit variations in contrast and noise, which could impact model generalization. Moreover, the presence of class imbalances—such as fewer normal cases compared to benign and malignant ones—could affect classification performance. Addressing these limitations would require additional validation on diverse, multi-institutional datasets and collaboration with clinicians to assess model reliability in real-world scenarios. Future work could also incorporate domain adaptation techniques and bias mitigation strategies to improve the robustness and fairness of ViT-based models across different populations.

It would be useful to give brief information about some existing studies using the dataset. The mammography dataset comprising INbreast, MIAS, and DDSM has been extensively utilized in various studies to enhance breast cancer detection and diagnosis through advanced machine learning techniques. For instance, a study by Al-Antari et al. developed a fully integrated computer-aided diagnosis (CAD) system employing deep learning models, achieving an overall classification accuracy of 95.64% on the INbreast dataset [27]. Similarly, Li et al. proposed a method combining deep learning with an extreme learning machine, resulting in accuracies of 97.19% on DDSM, 98.13% on MIAS, and 98.26% on INbreast datasets [28]. Another notable approach by Falconí et al. utilized transfer learning on NasNet Mobile and fine-tuned VGG models to classify mammogram images according to the BI-RADS scale, achieving an accuracy of 90.9% on the INbreast dataset [29]. These studies underscore the efficacy of integrating deep learning methodologies with traditional machine learning models to improve the accuracy and reliability of breast cancer diagnostics using the INbreast, MIAS, and DDSM datasets.

3 THE VISION TRANSFORMERS

Vision transformers (ViTs) are a revolutionary development in computer vision that have the potential to replace traditional convolutional neural networks (CNNs) as the backbone of various vision tasks. CNNs process visual data in blocks, and at each layer, more abstract representations that capture spatial hierarchies are generated [30]. The modern architectures employ average pooling in the last layers to produce task-specific outputs. For example, if an image is being classified, the final average pooling layer is replaced by a classification head, and if the task at hand is object detection, the final average pooling layer is omitted altogether. Overall, the functions of CNNs are very different from global attention-based vision transformers. Thus, while the architectural details may differ, the overall task outputs are still closely related to each other for CNNs [31]. ViTs, on the other hand, decompose the input images into fixed-sized patches that are fed into conventional transformer blocks, which is a self-attention-based deep learning architecture. This decomposition allows the end-to-end training of large transformers acting on very large image datasets by ensuring a linear scaling in complexity with respect to the size of the images independent of dataset size [32]. All of the self-attention mechanisms in the transformer allow the model to perform efficient deduplication of work in processing pixel interactions because each operation is not applied between every possible pair of pixels. Rather, operations are applied across groups of patches, and information between the groups is incorporated sparsely from some operations across particular patches in each group. This allows for a more controlled and modular learning process that operates at the level of entire patches while leveraging information from different parts of the visual input [33].

The self-attention mechanism is the cornerstone of vision transformers, allowing them to weigh different local parts of the image differently without losing any global information. The weighted mean of this local information is then calculated to obtain the image representation. The self-attention mechanism calculates attention scores that describe the similarity of the query feature against the key feature for all positions in the input space. [34] The attention score is calculated by taking the softmax of a scaled dot-product score, defined as the matrix multiplication (after scaling element-wise) between the query and key [35]. These attention scores specify the distribution of relative importance of signals from different positions. The output of the final self-attention layer is then calculated as a weighted sum of the value features, where each value is weighted by the normalized attention score [36]. The self-attention mechanism introduces the representation of each position to be influenced by the context surrounding the position. In a more global scope, self-attention leads to potential relationships between every position, and in turn makes it difficult for any position to contain the same amount of information [37].

Unlike conventional computer vision project names that consist of the phrase “smaller,” these project names instead describe the scale of the vision and vision model, similar to a hardware platform descriptor. We utilize this descriptor to emphasize that there are three variations of the model that cater to different hardware platforms while maintaining a consistent vision size. The vision part of the model is linear by complexity [38]. The “Base-16” refers to a model with a standard complexity structure that conventionally achieves a balance between size and speed. The “Small-16” and “Tiny-16” variations project the complexities and outputs to 64 and 32 respectively, making them cheaper and more accessible. Keeping the vision part models constant between Base-16, Small-16, and Tiny-16 allows us to test the different models under similar settings [39].

General workflow of ViTs is shows in figure 2 [40]. The Vision Transformer (ViT) workflow draws inspiration from the transformer architecture, which has shown great success in natural language processing. By applying transformers to image processing, Vision Transformers follow a specific sequence of steps to analyze image data and generate highly precise outcomes in computer vision tasks. Let's take a closer look at the detailed breakdown of the Vision Transformer workflow.

Image Preprocessing

Vision Transformers necessitate fixed input dimensions, which means that all images must be resized to a consistent size, such as 224x224 pixels for ViT. Following resizing, the image is split into a grid of set-size patches, with each patch flattened into a 1D vector. For example, a 224x224 image can be divided into 16x16 patches, resulting in a 14x14 grid, with each patch containing 256 pixels.

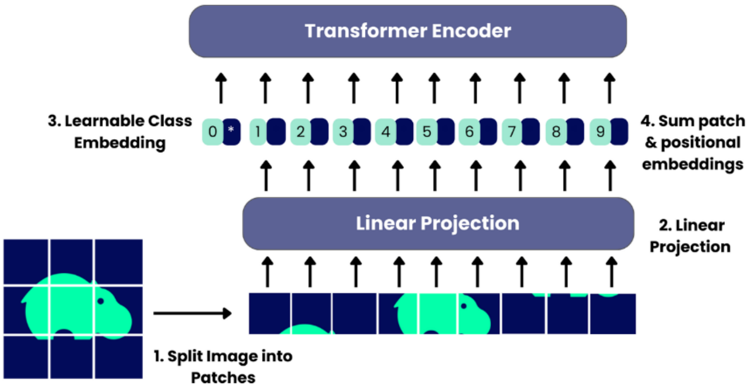
Transformer Encoder
3. Learnable Class Embedding
0
*
1
2
3
4
5
6
7
8
9
4. Sum patch & positional embeddings
Linear Projection
2. Linear Projection
1. Split Image into Patches

Figure 2. General workflow of ViTs.

These flattened patches are then passed through a linear projection to map them to a higher-dimensional space, such as 768 dimensions, ensuring that the model has a consistent feature size irrespective of the original patch size. The resulting vectors, one for each patch, are known as patch embeddings [41].

Positional Encoding

In contrast to CNNs, transformers do not naturally grasp spatial information, which is why positional encoding is included in each patch embedding. This encoding is crucial for helping the transformer comprehend the spatial connections between patches. Positional encodings commonly consist of learned or sinusoidal values that contribute specific position-related details to each patch embedding, thereby preserving the image's spatial arrangement within the sequence. [42, 43].

Input Embedding Construction

Following the incorporation of patch embedding and positional encoding, there is an output of a series of vectors, each of which represents a patch with position-aware details. A "class token" is also included at the beginning of the sequence, serving a similar purpose to the [CLS] token in NLP transformers. This class token is intended to consolidate information from all patches for use in classification tasks [44].

Transformer Encoder

The sequence of embeddings undergoes several layers of transformer encoders. Each layer includes: Multi-Head Self-Attention (MHSA), where each embedding interacts with others to capture global relationships; Layer Normalization to stabilize and speed up training; a Feed-Forward Network (FFN) with a ReLU activation that processes each embedding independently; and Residual Connections to help gradients flow through the model. This sequence is iterated across all transformer encoder layers, gradually learning complex relationships across patches [45].

Classification Head (or Task-Specific Head)

Upon completion of the encoder layers, the ultimate form of the class token serves as the representation of the image. In classification activities, this representation is processed through an MLP head, generating class probabilities for each category through softmax. This head is adaptable for other tasks like segmentation or object detection [46].

Training and Optimization

Vision Transformers are usually trained for classification tasks using cross-entropy loss, adjusting parameters through backpropagation and gradient descent. Pretraining on extensive datasets such as ImageNet can improve the ViT model's ability to perform well on related tasks. Fine-tuning the model on a specific dataset, like medical images, can also improve its performance by making it more adaptable to domain-specific characteristics [47].

Inference

In the process of inference, the image goes through identical preprocessing procedures such as resizing, patch extraction, patch embedding, and positional encoding. The representation of the ultimate class token is employed for classification, enabling the model to anticipate the category of fresh, unobserved images [48].

Vision Transformers offer a more organic approach to capturing overall connections across an entire image as opposed to CNNs, which are limited by local receptive fields. They show efficient scalability with larger datasets, making them suitable for extensive image datasets. In contrast to CNNs, Vision Transformers do not impose a rigid hierarchical feature structure, enabling more adaptable feature learning [49].

4 THE CLASSIFICATION STUDY

The Breast Mammography dataset was methodically divided to optimize model training, validation, and testing phases, with 80% of the images dedicated to training, 10% set aside for validation, and the remaining 10% reserved for testing. This split was chosen to provide a balanced approach that maximizes training data while ensuring ample samples for unbiased validation and evaluation. To prepare the images for deep learning model input, each image was resized to a uniform dimension of 384x384 pixels with three color channels (RGB). This resizing ensures that all images share a consistent structure, which is essential for convolutional neural networks that rely on uniform input shapes for accurate learning. Additionally, the choice to process the images as color (RGB) images, rather than grayscale, preserves critical color details that could aid in distinguishing between benign, malignant, and normal tissue types. Normalization was applied to each image, adjusting pixel values to a standardized range, typically between 0 and 1, or to a distribution centered around zero. This step is critical as it minimizes variations across the dataset, enabling the network to focus on important visual features rather than being affected by differing brightness or contrast levels. Such preprocessing ensures that the model can learn effectively from the images without bias introduced by inconsistent pixel intensities. This preprocessing pipeline—including resizing, color preservation, and normalization—was applied identically to images used in both training and testing. This approach guarantees that the model encounters images of identical quality and format during training and evaluation, reducing any risk of performance discrepancies due to preprocessing differences. Overall, this careful preparation of the Breast Mammography dataset supports robust and reliable model training, validation, and testing, fostering a more accurate classification performance across breast tissue image categories. The training parameters are selected as follows. **MiniBatchSize** =12, **MaxEpochs** =5, **IterationsPerEpoch** =443, **ObservationsTrain**=1773 and **Iterations**=8865. This study utilized the base, small, and tiny ViT models with their default parameters. No particular hyperparameter modification was conducted to improve performance, as our emphasis was on examining the influence of the default models on the dataset. The ViT models used in this study were trained using a standardized set of hyperparameters to optimize performance while maintaining computational efficiency. The models utilized a patch size of 16×16 pixels, with input images resized to 384×384 pixels. Training was conducted using the Adam optimizer with an initial learning rate of 1e-4 and a weight decay of 0.01. Each model was trained for five epochs with a batch size of 12, processing 443 iterations per epoch, totaling 8,865 iterations. The loss function employed was cross-entropy loss, and images were normalized to a [0,1] range to enhance model stability. The activation function used was Gaussian Error Linear Unit (GELU), and dropout was set at 0.1 to prevent overfitting. The transformer architecture varied across models, with the Base model featuring 12 multi-head self-attention heads, 12 transformer encoder layers, a hidden dimension of 768, and a feed-forward network dimension of 3072. The Small model had 6 attention heads, 8 encoder layers, a hidden dimension of 384, and a feed-forward network dimension of 1536, while the Tiny model was the most compact, with 3 attention heads, 4 encoder layers, a hidden dimension of 192, and a feed-forward network dimension of 768. Positional encoding was learnable across all models. These hyperparameters were carefully selected to balance accuracy and efficiency, with the Small model achieving the highest classification performance, while the Tiny model offered a computationally efficient alternative with moderate accuracy.

In training the network, the *Adam* (Adaptive Moment Estimation) algorithm was applied as an optimization solver for deep learning. This optimizer was selected for its adaptive learning rate capability, which efficiently handles dynamic learning rates and accelerates convergence during training. The training process was executed on a GPU using parallel computing, with 16 parallel workers operating concurrently to maximize processing efficiency and reduce training time. Key training parameters are as follows:

Initial Learning Rate: Set at 1e-4 to provide a stable starting point that adjusts adaptively during training, facilitating consistent model updates.

Shuffle: Data shuffling was configured to occur at every epoch, ensuring that the training data is randomly reordered with each cycle to enhance generalization and minimize overfitting.

Execution Environment: Configured for parallel processing, utilizing the GPU's multi-threading capabilities for optimized computation.

This configuration was designed to support efficient, stable, and resource-effective training, as detailed in the computational setup, and provided the necessary foundation for robust model convergence.

The computational setup, detailed in Table 1, illustrates the hardware specifications and the configuration that enabled efficient and stable training. The optimized environment supported a faster convergence rate, enhancing model performance while ensuring consistent resource utilization throughout the learning process.

Table 1. Hardware specifications of the computer used in this study.

Processor	**12th Gen Intel(R) CoreTM i9-12900F 2.40 GHz**
Cores, Processors	16, 24
Installed RAM	64.0 GB (63.7 GB usable)
GPU	NVIDIA RTX A4000
DirectX version	12 (FL 12.1)
GPU Memory	47.9 GB (16.0 GB Dedicated, 31.9 GB Shared)

The Vision Transformers (ViT) models employed in this study consist of the following configurations:

1. **Base-16-ImageNet-384**: This base-sized model contains 86.8 million parameters, with a patch size of 16 pixels, and is fine-tuned on the ImageNet 2012 dataset. It processes images at a resolution of 384x384 pixels, making it suitable for capturing complex features in larger images.

2. **Small-16-ImageNet-384**: This smaller model includes 22.1 million parameters, also with a patch size of 16, and is fine-tuned on the ImageNet 2012 dataset at the same resolution of 384x384 pixels. It provides a balance between model size and computational efficiency, enabling effective feature extraction with lower resource requirements.

3. **Tiny-16-ImageNet-384**: The smallest of the three models, this configuration has 5.7 million parameters and a patch size of 16. It is similarly fine-tuned on the ImageNet 2012 dataset with an image resolution of 384x384 pixels. This model is optimized for scenarios with limited computational resources while still leveraging the benefits of the ViT architecture.

Each of these ViT models, fine-tuned with ImageNet 2012 data, offers unique trade-offs in terms of parameter count and processing capacity, making them adaptable to various resource constraints and performance needs in image classification tasks.

In the study, the dataset was divided into three subsets: 80% for training, 10% for validation, and 10% for testing. These groups were separated before training began, ensuring that they consisted of independent images.

The training and testing performance graphs, along with the computed values for the base, small, and tiny ViT models, are presented below. Figure 3 shows the training visualities for the base model.

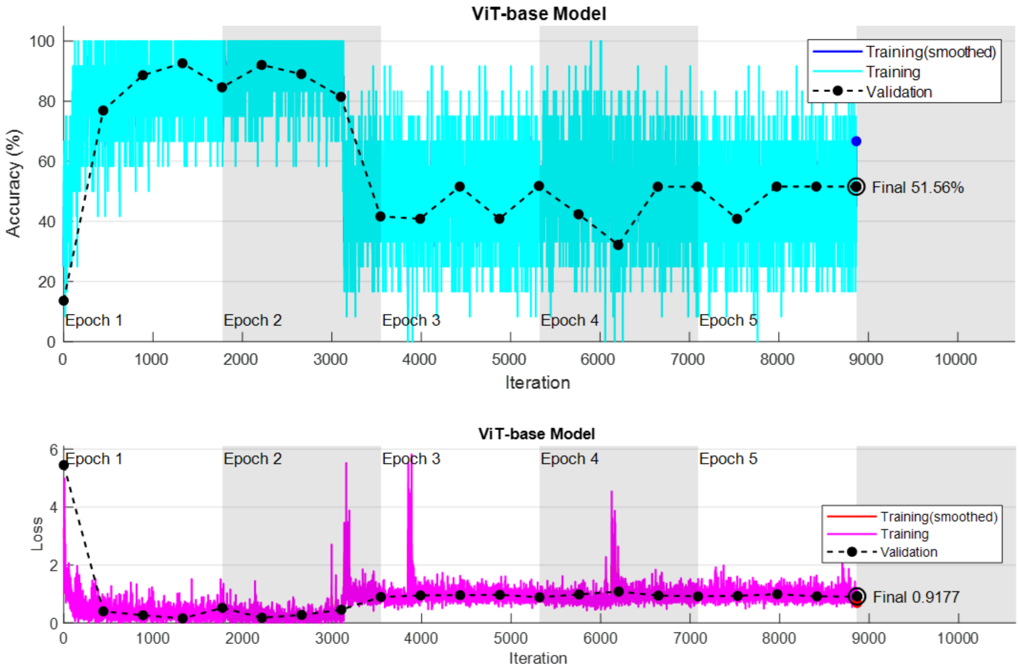

Figure 3. Training process of the base model.

The test performance parameters of the ViT base model—Accuracy, Error, Recall, Specificity, Precision, F1-Score, Geometric Mean Precision Recall (G-Measure PR), Geometric Mean Sensitivity Specificity (G-Measure SS), and Matthews Correlation Coefficient (MCC)—are presented in the table below.

Table 2. Performance metrics for the base model.

Accuracy	Error	Recall	Specificity	Precision	F1-Score	G-Measure PR	G-Measure SS	MCC
83.9534	16.0466	87.4203	89.5156	88.4118	87.4151	87.6649	88.0407	71.9342

The training visualities for the small model are illustrated in Figure 4. The test performance metrics of the model are listed in Table 3.

Table 3. Performance metrics for the small model.

Accuracy	Error	Recall	Specificity	Precision	F1-Score	G-Measure PR	G-Measure SS	MCC
87.9369	12.0631	91.1346	92.4234	91.1982	91.1647	91.1656	91.7628	78.5135

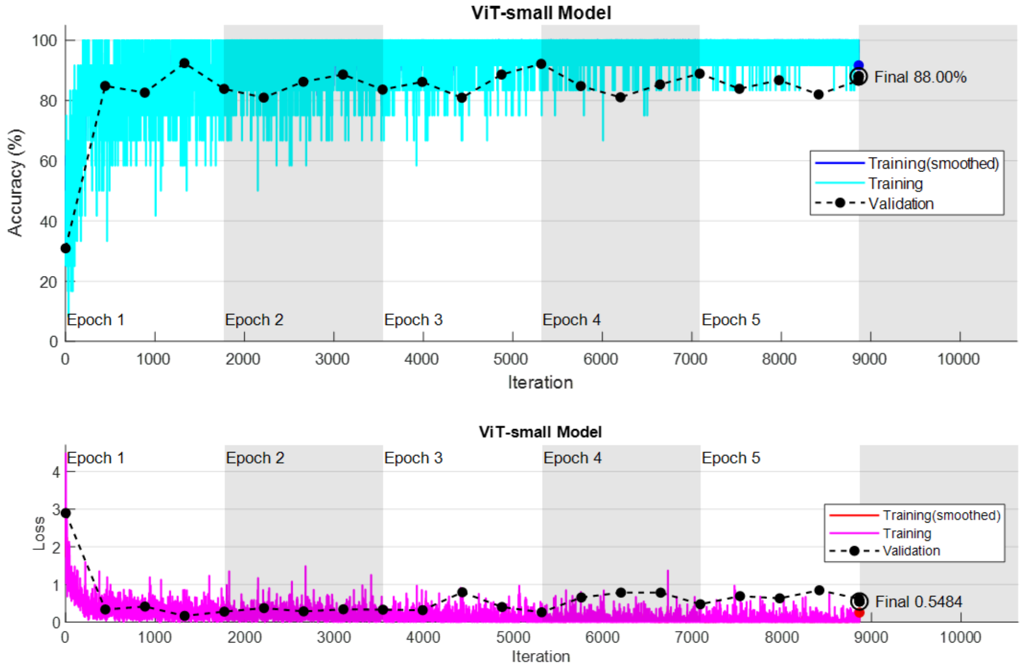

Figure 4. Training process of the small model.

Finally, the training of the tiny model is visualized in Figure 5. The test performance parameters of the model can be seen in Table 4.

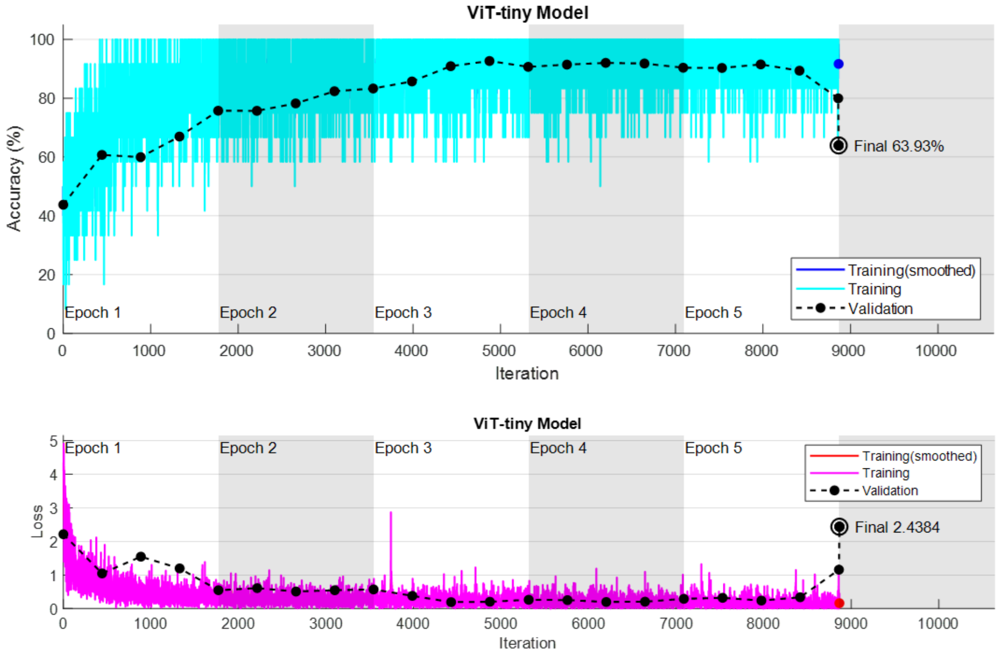

Figure 5. Training process of the tiny model.

Table 4. Performance metrics for the tiny model.

Accuracy	Error	Recall	Specificity	Precision	F1-Score	G-Measure PR	G-Measure SS	MCC
81.5859	18.4141	86.3841	88.3812	86.4489	86.3963	86.4064	87.3136	67.1630

For all three ViT model types, training was conducted using the same parameters on the 80% training dataset, while validation was performed independently using the 10% validation dataset. Testing was carried out after training was completed using a separate and independent 10% test data.

As a result of the training process, the test accuracies for each model type are as follows:

- **Base model**: Accuracy = 0.8395
- **Small model**: Accuracy = 0.8794
- **Tiny model**: Accuracy = 0.8159

These results indicate that the accuracy rates vary according to the model size, with the small model achieving the highest accuracy, while the tiny model achieved the lowest.

The training durations for each model are as follows:

- **Base Model**: Training completed in 32 hours 37 minutes 40 seconds.
- **Small Model**: Training completed in 3 hours 12 minutes 39 seconds.
- **Tiny Model**: Training completed in 1 hour 47 minutes 1 second.

These training times demonstrate a significant reduction in duration as the model size decreases. However, there is also a noticeable trade-off in accuracy, with the smaller models taking less time but showing some differences in accuracy. This highlights the impact of model size on both accuracy and training time. Confusion matrixes showing the training accuracies obtained for three models are given in figure 6.

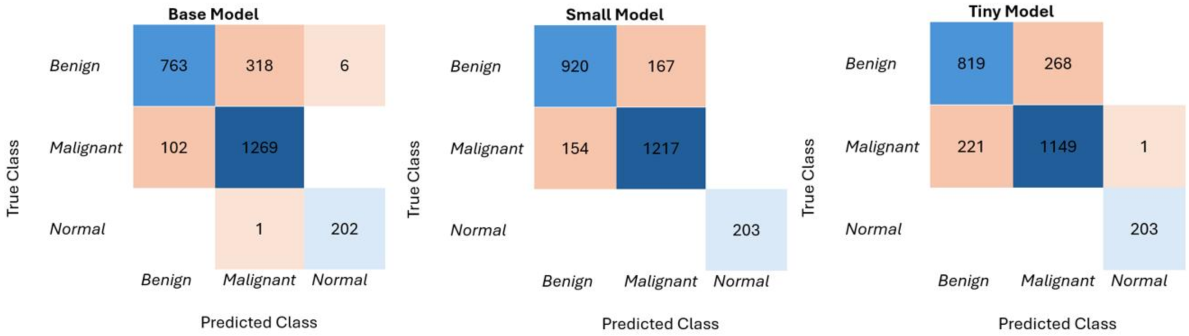

Figure 6. Confusion matrixes obtained for Base, Small and Tiny Models.

Figure 7 illustrates the Area Under the Curve (AUC) graphics for the three models. In the figure, the classes 1, 2 and 3 represents Bening Malignant and Normal classes respectively.

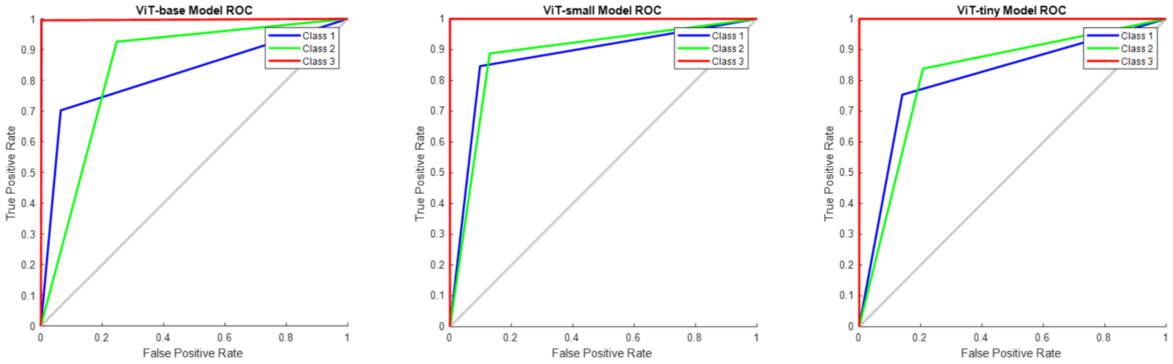

Figure 7. AUC graphics obtained for Base, Small and Tiny Models.

For the base model, the performance values were calculated as, Benign: 0.8186, Malignant: 0.8392, and Normal: 0.9963. For the small model, performance values are found as Benign: 0.8743, Malignant: 0.8791 and Normal: 1.0 and for the tiny model, the performance is found as Benign: 0.8065, Malignant: 0.8152 and Normal: 0.9998.

Figure 8 shows the progress of accuracy and loss during the ViTs Base model. As shown in Figure 4, the validation performance of the training process for the base model of the ViT network was 51.56%. This indicates that, during the validation phase, the model was able to correctly classify 51.56% of the samples, reflecting its ability to generalize from the training data to unseen data. While the performance might suggest room for improvement, it provides valuable insight into the network's capabilities, and further optimizations or fine-tuning could lead to enhanced accuracy in future iterations.

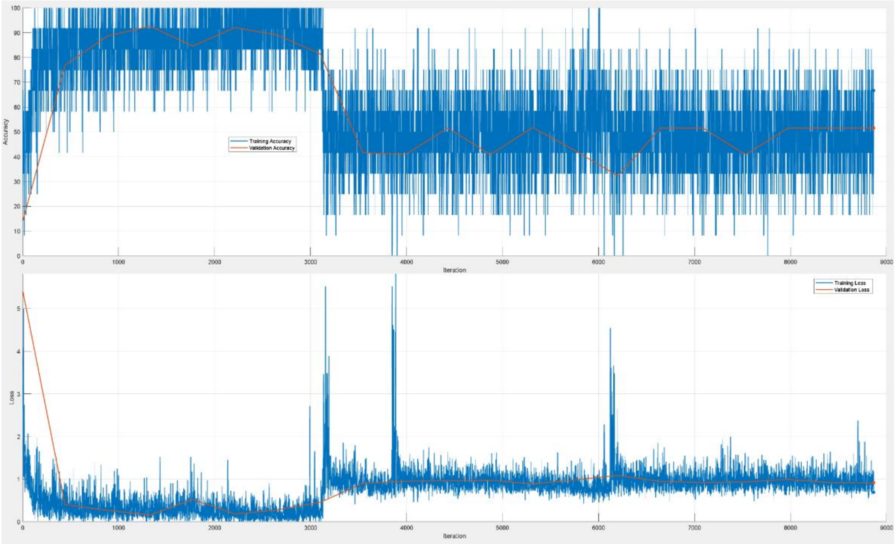

Figure 8. Accuracy and Loss values obtained for the base model.

Similarly, figure 9 is for the progress of accuracy and loss obtained for the small model of ViTs. According to the Small model, the validation performance of the training process was 88.00%. This high validation accuracy suggests that the model has effectively learned the features necessary for distinguishing between the classes in the dataset. The improved performance of the Small model compared to the base model demonstrates the advantages of using a more compact architecture, which may offer better generalization and efficiency in training. Further analysis could be conducted to understand the specific factors contributing to this improvement, such as the choice of hyperparameters or the model's ability to capture intricate patterns in the data.

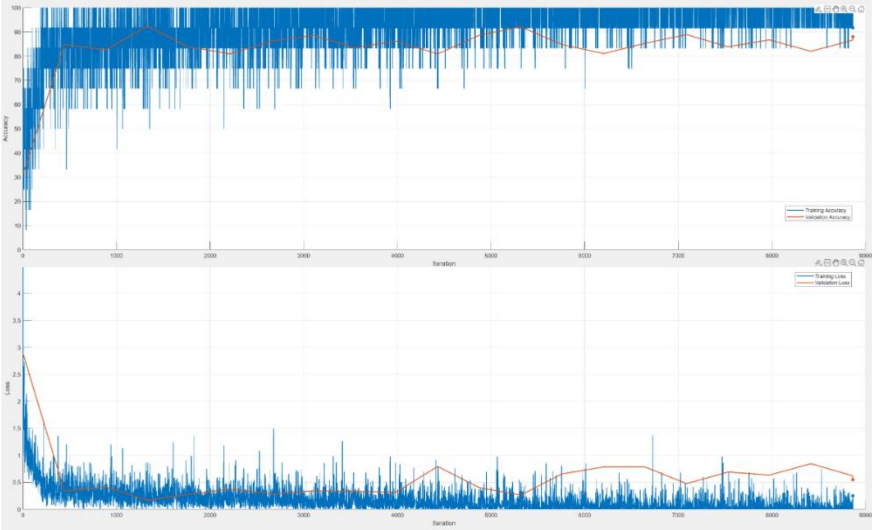

Figure 9. Accuracy and Loss values obtained for the small model.

Finally, figure 10 presents the progress of accuracy and loss values obtained for the tiny model.

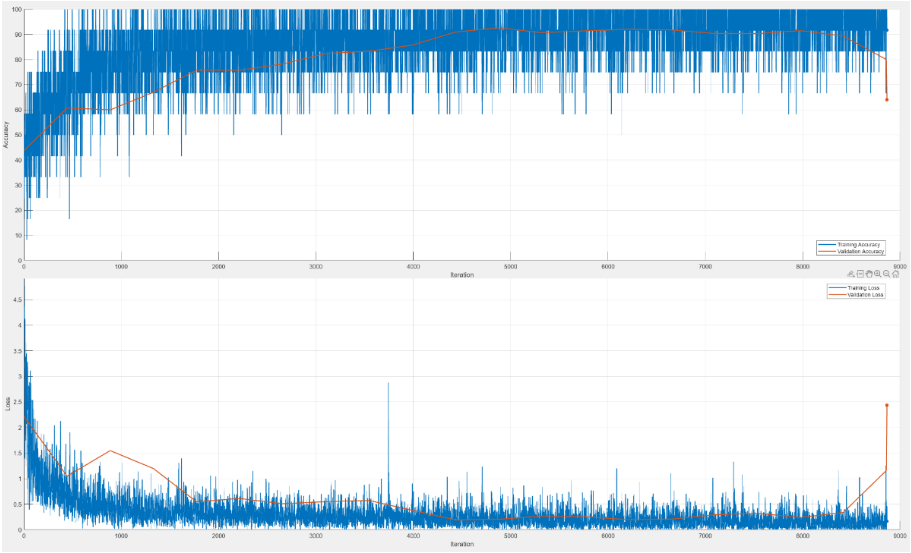

Figure 10. Accuracy and Loss values obtained for the tiny model.

According to the Tiny model, the validation success of the training process was 63.93%. This result indicates the model's performance in classifying images it had not seen during training, both during the intermediate validation tests and after the final training completion. These validation tests used unseen images, which are crucial for assessing the model's generalization ability. While the Tiny model shows a moderate level of validation accuracy, it suggests there may be potential for improvement, possibly through further tuning or more sophisticated techniques for feature extraction and model optimization. The test success for the Tiny model is also detailed in the confusion matrix in figure 3, which provides a deeper insight into the misclassifications and the overall model performance.

For all three models, table 5 provides a comprehensive overview of the training iteration progress, including the duration of each iteration, training and test performance, and calculation data from both the beginning and the end of the training process:

Table 5. Training iteration progress.

		Base			**Small**			**Tiny**		
Iteration	**Epoch**	**Time Elapsed**	**Training Accuracy**	**Validation Accuracy**	**Time Elapsed**	**Training Accuracy**	**Validation Accuracy**	**Time Elapsed**	**Training Accuracy**	**Validation Accuracy**
0	0	00:01:28		13.652	00:00:41		30.951	00:00:27		43.738
1	1	00:01:29	25		00:00:41	25		00:00:27	50	
50	1	00:08:32	50		00:01:42	75		00:01:00	41.667	
100	1	00:17:28	66.667		00:02:40	91.667		00:01:36	33.333	
150	1	00:26:23	50		00:03:39	75		00:02:11	58.333	
200	1	00:34:57	83.333		00:04:37	58.333		00:02:43	41.667	
250	1	00:43:23	91.667		00:05:36	83.333		00:03:16	50	
300	1	00:51:47	58.333		00:06:35	75		00:03:50	58.333	
350	1	01:00:12	91.667		00:07:32	100		00:04:28	66.667	
400	1	01:08:39	91.667		00:08:30	75		00:05:07	58.333	
443	1	01:17:09	83.333	76.909	00:09:50	91.667	84.806	00:06:08	91.667	60.7
…										
8417	5	30:15:06	41.667	51.561	03:01:06	100	81.986	01:40:30	91.667	89.357
8450	5	30:23:41	75		03:01:51	100		01:40:54	91.667	
8500	5	30:36:30	50		03:02:58	100		01:41:29	91.667	
8550	5	30:49:25	58.333		03:04:05	100		01:42:05	100	
8600	5	31:02:18	66.667		03:05:11	100		01:42:41	91.667	
8650	5	31:15:14	66.667		03:06:18	100		01:43:16	100	
8700	5	31:28:08	33.333		03:07:25	100		01:43:52	100	
8750	5	31:41:03	50		03:08:32	100		01:44:28	100	
8800	5	31:53:59	50		03:09:39	91.667		01:45:02	91.667	
8850	5	32:06:55	25		03:10:46	100		01:45:39	91.667	
8860	5	32:23:00	50	51.561	03:11:35	100	86.574	01:46:10	100	79.955
8865	5	32:37:17	66.667	51.561	03:12:18	91.667	88.003	01:46:38	91.667	63.934

The training performances of the network models used, including their success rates, error rates, parameter sizes, and training times, are summarized in table 6. This table provides an overview of the performance of each model, helping to compare their relative effectiveness in terms of classification accuracy, model complexity (parameter size), and the time taken to complete the training process. The models demonstrate different strengths, with the Small model achieving the highest accuracy, while the Tiny model offers a more compact architecture with moderate performance. The Base model, although achieving lower accuracy, can still serve as a useful baseline for comparison against more optimized configurations.

Table 6. Performance metrics of the training and test process.

Model	Accuracy Train (%)	Accuracy Validation (%)	Accuracy Test (%)	Training Loss	Validation Loss	Parameters	Training Time
ViT-Base	66.667%	51.561%	83.9534%	0.69153	0.91771	86.8 million	32:37:40
ViT-Small	91.667%	88.003%	87.9369%	0.25361	0.54841	22.1 million	03:12:39
ViT-Tiny	91.667%	63.934%	81.5859%	0.16297	2.4384	5.7 million	01:47:01

When interpreting Table 6, it is evident that the Small model of the Vision Transformer (ViT) networks achieves the highest accuracy, with a performance difference of 3.99% compared to the Base model, despite having approximately 4 times fewer parameters. This highlights the effectiveness of the Small model in terms of both performance and computational efficiency.

Although the Tiny model has about 15 times fewer parameters than the Base model and approximately 4 times fewer parameters than the Small model, its performance remains close to that of the other models. Notably, the Tiny model achieves this level of success in significantly less time, which makes it a viable option for scenarios where computational speed is critical.

While the Base model ranks second in terms of accuracy, it is worth considering for certain applications where a balance between performance and computational time is required. Therefore, the choice of model depends heavily on the specific needs of the research or application, whether that is achieving the highest accuracy, minimizing computational time, or balancing both factors.

The ViT base model consists of 86.8 million parameters. Given the complexity of such a large model, processing images involves an extensive number of computations, leading to a significant demand for computational resources at maximum capacity. As a result, obtaining outcomes takes considerably longer compared to the small and tiny models. Although the ViT base model's validation performance during training appears significantly lower than that of the other two models, its test performance was found to be very close to theirs. Ultimately, when analyzing test performance, the results indicate that the accuracy rates across all three models are relatively close and fall within an acceptable range.

5 RESULTS AND DISCUSSION

Breast cancer remains one of the leading causes of mortality among women worldwide, with early detection playing a crucial role in improving treatment outcomes. One of the most important diagnostic tools in this context is mammography, which allows for the detection of breast masses, a key indicator of potential malignancy. In recent years, advancements in deep learning have demonstrated promising results in improving the accuracy and efficiency of medical image analysis, particularly through the use of Vision Transformers (ViTs). ViTs have gained significant attention for their ability to model long-range dependencies in images and leverage self-attention mechanisms, making them ideal candidates for complex medical imaging tasks. This study evaluates the performance of three different Vision Transformer models—base-16, small-16, and tiny-16—on a dataset of breast mammography images containing masses. These models were specifically chosen to assess how different configurations of ViTs, with varying sizes and parameter counts, perform on the task of classifying mammographic images into categories such as benign, malignant, and normal. The ViTs' self-attention mechanism helps address challenges posed by the inherent complexity of mammographic textures and the low contrast that is often seen in medical imaging, which can make traditional image classification techniques less effective. Through comparative analysis, this study highlights the strengths and limitations of each ViT model. The findings provide valuable insights into the performance of these models in terms of accuracy, training time, and computational efficiency. By evaluating their effectiveness in breast mass classification tasks, this research aims to contribute to the broader understanding of how transformer-based architectures can enhance diagnostic accuracy in medical imaging. The results serve as a benchmark for future research, paving the way for further exploration of ViTs and other transformer-based models in medical image classification and diagnosis. This study underscores the potential of Vision Transformers in advancing the field of medical image analysis, particularly for early breast cancer detection, and supports their future application in clinical settings.

This paper provides a thorough examination of the classification architecture of Vision Transformers (ViTs) in the context of breast mammography image classification. Specifically, it investigates three sub-models within the ViTs framework: the base-16, small-16, and tiny-16 models. Each of these sub-models has been individually evaluated for their performance in classifying mammographic images, which are crucial for early breast cancer detection. The study aims to provide an in-depth understanding of how these different ViT configurations, with varying model sizes and complexities, contribute to the classification task. A comprehensive analysis is conducted to highlight the strengths, weaknesses, and performance characteristics of each sub-model in the ViT architecture. The base-16 model, with its larger parameter size, is assessed for its ability to capture complex patterns in mammographic images, while the small-16 and tiny-16 models, with fewer parameters, are evaluated for their efficiency and speed, particularly in clinical settings where computational resources may be limited. By discussing the specific advantages and limitations of each model, this paper offers valuable insights into the trade-offs between accuracy, computational cost, and training time. This analysis also explores the role of the self-attention mechanism inherent in ViTs, which enables the models to effectively focus on relevant features within the mammographic images, addressing challenges such as low contrast and intricate textures commonly found in medical imaging. The study emphasizes how the selection of the appropriate ViT sub-model can be tailored to different research or clinical needs, depending on the available computational resources and the required diagnostic accuracy.

The study primarily focuses on a theoretical and computational comparison of ViT models for breast cancer detection in mammographic imaging. By evaluating the Base-16, Small-16, and Tiny-16 ViT configurations on a standardized dataset, we provide a benchmark analysis that highlights their strengths and limitations in terms of accuracy, training time, and computational efficiency. However, we acknowledge that clinical validation and real-world testing with physicians would enhance the applicability of our findings. Future research directions could involve collaboration with radiologists to assess model performance in real clinical settings, integrating physician feedback to improve interpretability and usability. Additionally, testing the models on diverse, real-world datasets with varying imaging conditions and patient demographics could further validate their robustness and reliability. Such efforts would bridge the gap between theoretical performance and practical deployment, making these models more applicable for clinical decision-making.

The primary objective of this research is to evaluate the outputs of ViT models on the same dataset in order to draw conclusions about their strengths, weaknesses, benefits, and disadvantages. This will be accomplished by using the default parameters of the models. If we had begun out with the intention of surpassing the existing performance of the dataset in the literature by utilizing a single ViT model, our primary focus would have been on further improving the dataset through the application of image processing techniques.

Overall, this paper serves as a detailed guide for understanding the application of ViTs in medical image classification, particularly for breast cancer detection, and it lays the groundwork for further research into optimizing and refining transformer-based models for use in medical diagnostics.

Conflict of Interest Statement

There is no conflict of interest between the authors.

Statement of Research and Publication Ethics

The study is complied with research and publication ethics.

Artificial Intelligence (AI) Contribution Statement

This manuscript was composed, revised, analyzed, and prepared without the aid of any artificial intelligence techniques. All content, encompassing text, data analysis, and figures, was exclusively produced by the authors.

Contributions of the Authors

Uğur DEMİROĞLU formulated the theoretical framework, conducted the analytical computations, and executed the numerical simulations. Bilal ŞENOL took the lead in writing the manuscript. Both authors Uğur DEMİROĞLU and Bilal ŞENOL contributed to the final version of the work. Bilal ŞENOL supervised the project.

REFERENCES

[1] M. Arnold *et al.*, "Current and future burden of breast cancer: Global statistics for 2020 and 2040," *The Breast*, vol. 66, pp. 15-23, 2022.

[2] C. I. Lee and J. G. Elmore, "Beyond survival: a closer look at lead-time bias and disease-free intervals in mammography screening," *JNCI: Journal of the National Cancer Institute*, vol. 116, no. 3, pp. 343-344, 2024.

[3] L. N. Fuzzell *et al.*, "Cervical cancer screening in the United States: Challenges and potential solutions for underscreened groups," *Preventive Medicine*, vol. 144, p. 106400, 2021.

[4] G. Savarese *et al.*, "Global burden of heart failure: a comprehensive and updated review of epidemiology," *Cardiovascular Research*, vol. 118, no. 17, pp. 3272-3287, 2022.

[5] S. Sriussadaporn *et al.*, "Ultrasonography increases sensitivity of mammography for diagnosis of multifocal, multicentric breast cancer using 356 whole breast histopathology as a gold standard," *Surgical Practice*, vol. 26, no. 3, pp. 181-186, 2022.

[6] N. Pashayan *et al.*, "Personalized early detection and prevention of breast cancer: ENVISION consensus statement," *Nature Reviews Clinical Oncology*, vol. 17, no. 11, pp. 687-705, 2020.

[7] L. Nicosia *et al.*, "History of mammography: analysis of breast imaging diagnostic achievements over the last century," *Healthcare*, vol. 11, no. 11, p. 1596, 2023.

[8] C. Poggi, "The Evolution of the Radiographer's Educational Path: EBP and Communication Skills in the Mammography Room," in *Breast Imaging Techniques for Radiographers*, Springer Nature Switzerland, 2024, pp. 259-276.

[9] H. O. Kolade-Yunusa and U. D. Itanyi, "Outcome of mammography examination in asymptomatic women," *Annals of African Medicine*, vol. 20, no. 1, pp. 52-58, 2021.

[10] L. Abdelrahman *et al.*, "Convolutional neural networks for breast cancer detection in mammography: A survey," *Computers in Biology and Medicine*, vol. 131, p. 104248, 2021.

[11] D. Barba *et al.*, "Breast cancer, screening and diagnostic tools: All you need to know," *Critical Reviews in Oncology/Hematology*, vol. 157, p. 103174, 2021.

[12] W. Y. Sung *et al.*, "Experiences of women who refuse recall for further investigation of abnormal screening mammography: A qualitative study," *International Journal of Environmental Research and Public Health*, vol. 19, no. 3, p. 1041, 2022.

[13] H. J. Han *et al.*, "Characteristics of breast cancers detected by screening mammography in Taiwan: a single institute's experience," *BMC Women's Health*, vol. 23, no. 1, p. 330, 2023.

[14] A. Aleissaee *et al.*, "Transformers in remote sensing: A survey," *Remote Sensing*, vol. 15, no. 7, p. 1860, 2023.

[15] Y. Liu *et al.*, "A survey of visual transformers," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[16] A. Khan *et al.*, "A survey of the recent architectures of deep convolutional neural networks," *Artificial Intelligence Review*, vol. 53, pp. 5455-5516, 2020.

[17] A. Agarwal and N. Ratha, "Deep Learning in Computer Vision Progress and Threats," in *Applications of Artificial Intelligence, Big Data and Internet of Things in Sustainable Development*, vol. 23, 2022.

[18] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *arXiv:2010.11929*, 2020.

[19] Y. Yuan *et al.*, "Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet," in *Proc. ECCV*, 2021, pp. 558–576.

[20] A. Steiner *et al.*, "How to train your vit? data, augmentation, and regularization in vision transformers," *arXiv preprint arXiv:2106.10270*, 2021.

[21] X. Wu *et al.*, "CTransCNN: Combining transformer and CNN in multilabel medical image classification," *Knowledge-Based Systems*, vol. 281, p. 111030, 2023.

[22] M. Hayat *et al.*, "Hybrid Deep Learning EfficientNetV2 and Vision Transformer (EffNetV2-ViT) Model for Breast Cancer Histopathological Image Classification," *IEEE Access*, 2024.

[23] G. Ayana and S. W. Choe, "Vision transformers-based transfer learning for breast mass classification from multiple diagnostic modalities," *Journal of Electrical Engineering & Technology*, vol. 19, no. 5, pp. 3391-3410, 2024.

[24] M. L. Abimouloud *et al.*, "Advancing breast cancer diagnosis: token vision transformers for faster and accurate classification of histopathology images," *Visual Computing for Industry, Biomedicine, and Art*, vol. 8, no. 1, p. 1, 2025.

[25] "NHS to launch world's biggest trial of AI breast cancer diagnosis," *The Guardian*, Feb. 4, 2025. [Online]. Available: https://www.theguardian.com/society/2025/feb/04/nhs-to-launch-worlds-biggest-trial-of-ai-breast-cancer-diagnosis.

[26] Kaggle, "Mammography Dataset from INbreast, MIAS and DDSM," accessed Nov. 4, 2024. [Online]. Available: https://www.kaggle.com/datasets/emiliovenegas1/mammography-dataset-from-inbreast-mias-and-ddsm.

[27] M. A. Al-Antari *et al.*, "Evaluation of deep learning detection and classification towards computer-aided diagnosis of breast lesions in digital X-ray mammograms," *Computer Methods and Programs in Biomedicine*, vol. 196, p. 105584, 2020.

[28] X. Li *et al.*, "Multiparametric magnetic resonance imaging for predicting pathological response after the first cycle of neoadjuvant chemotherapy in breast cancer," *Investigative Radiology*, vol. 50, no. 4, pp. 195-204, 2015.

[29] L. G. Falconí *et al.*, "Transfer learning in breast mammogram abnormalities classification with mobilenet and nasnet," in *2019 International Conference on Systems, Signals and Image Processing (IWSSIP)*, 2019, pp. 109-114.

[30] W. Hu *et al.*, "A state-of-the-art survey of artificial neural networks for whole-slide image analysis: from popular convolutional neural networks to potential visual transformers," *Computers in Biology and Medicine*, vol. 161, p. 107034, 2023.

[31] S. K. Hamed *et al.*, "Enhanced Feature Representation for Multimodal Fake News Detection Using Localized Fine-Tuning of Improved BERT and VGG-19 Models," *Arabian Journal for Science and Engineering*, pp. 1-17, 2024.

[32] Y. Fang *et al.*, "You only look at one sequence: Rethinking transformer in vision through object detection," *Advances in Neural Information Processing Systems*, vol. 34, pp. 26183-26197, 2021.

[33] Z. Zeng *et al.*, "You only sample (almost) once: Linear cost self-attention via Bernoulli sampling," in *International Conference on Machine Learning*, PMLR, 2021, pp. 12321-12332.

[34] A. Rehman, "Transformers in Computer Vision: Recent Advances and Applications," *International Journal of Advanced Engineering Technologies and Innovations*, vol. 1, no. 1, 2022.

[35] M. Hassanin *et al.*, "Visual attention methods in deep learning: An in-depth survey," *Information Fusion*, vol. 108, p. 102417, 2024.

[36] M. Li *et al.*, "SACNN: Self-attention convolutional neural network for low-dose CT denoising with self-supervised perceptual loss network," *IEEE Transactions on Medical Imaging*, vol. 39, no. 7, pp. 2289-2301, 2020.

[37] M. H. Guo *et al.*, "Beyond self-attention: External attention using two linear layers for visual tasks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 5, pp. 5436-5447, 2022.

[38] R. Divya and J. Prabhakar, "Vision Transformer-based Model for Human Action Recognition in Still Images," *Journal of Computational Analysis and Applications*, vol. 33, no. 08, pp. 522-531, 2024.

[39] E. Şahin *et al.*, "Multi-objective optimization of ViT architecture for efficient brain tumor classification," *Biomedical Signal Processing and Control*, vol. 91, p. 105938, 2024.

[40] Marqo, "Introduction to Vision Transformers," accessed Nov. 4, 2024. [Online]. Available: https://www.marqo.ai/course/introduction-to-vision-transformers.

[41] A. Sriwastawa and J. A. Arul Jothi, "Vision transformer and its variants for image classification in digital breast cancer histopathology: A comparative study," *Multimedia Tools and Applications*, vol. 83, no. 13, pp. 39731-39753, 2024.

[42] V. Jain *et al.*, "Transformers are adaptable task planners," in *Conference on Robot Learning*, PMLR, 2023, pp. 1011-1037.

[43] S. Wang *et al.*, "When transformer meets robotic grasping: Exploits context for efficient grasp detection," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 8170-8177, 2022.

[44] L. Xu *et al.*, "Mctformer+: Multi-class token transformer for weakly supervised semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[45] A. Deihim *et al.*, "STTRE: A Spatio-Temporal Transformer with Relative Embeddings for multivariate time series forecasting," *Neural Networks*, vol. 168, pp. 549-559, 2023.

[46] W. Wang *et al.*, "Semi-supervised vision transformer with adaptive token sampling for breast cancer classification," *Frontiers in Pharmacology*, vol. 13, p. 929755, 2022.

[47] H. E. Kim *et al.*, "Transfer learning for medical image classification: A literature review," *BMC Medical Imaging*, vol. 22, no. 1, p. 69, 2022.

[48] L. Xu *et al.*, "Multi-class token transformer for weakly supervised semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4310-4319.

[49] O. S. Khedr *et al.*, "The classification of bladder cancer based on Vision Transformers (ViT)," *Scientific Reports*, vol. 13, no. 1, p. 20639, 2023.