International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(15s) 2026 pp. 623-637
https://doi.org/10.70917/ijcisim-2026-4441
Article

A Novel Enhanced ViT-CNN-ATS Architecture with Trigonometric Filters for Breast Cancer Segmentation and Classification in Ultrasound Imaging*

Lakshmi S[1], M T Somashekara[2], Asha C. S[3]

[1]Department of Computer Application, Dayananda Sagar College of Arts, Science and Commerce, Bangaluru, India, lakshminaresh02@gmail.com
[2]Departmentof Computer Science and Applications, Bangalore University, Bengaluru, India, somashekarmt@bub.ernet.in
[3]NIE First Grade College, Mysuru, India, csasharukmini@gmail.com

Abstract: This study proposes a novel hybrid deep learning framework, ViT-CNN-ATS, designed for accurate classification and segmentation of breast cancer in ultrasound images. The model integrates the global contextual learning of the Vision Transformer (ViT) with the powerful local feature extraction capabilities of Convolutional Neural Networks (CNN), further refined by Attention Token Selection (ATS) for enhanced focus on the most informative regions. To overcome the inherent challenges of ultrasound imaging such as low contrast, speckle noise, and variability in tumor appearance comprehensive preprocessing techniques including contrast enhancement and noise reduction are employed. A masked ultrasound image dataset, organized into benign, malignant, and normal classes, serves as the foundation for training and evaluation. Data augmentation strategies are incorporated to improve model robustness and generalization. The proposed hybrid model demonstrates significant improvements in classification accuracy, Dice Similarity Coefficient, and Intersection over Union (IoU) scores when compared to baseline models using only CNN or ViT. These results underscore the potential of ViT-CNN-ATS as a reliable, AI-assisted diagnostic tool to support radiologists in early and precise breast cancer detection using ultrasound imaging.

Keywords: Breast Cancer Detection, Ultrasound Imaging, Vision Transformer (ViT), Convolutional Neural Networks (CNN), Attention Token Selection (ATS), Dice Score, Intersection over Union (IoU)

1. Introduction

Breast cancer remains the most commonly diagnosed cancer among women globally, with an estimated 2.3 million new cases and 685,000 deaths reported in 2020 alone, according to the World Health Organization (WHO) [1]. Early and accurate detection continues to be critical for improving survival rates and ensuring effective treatment outcomes. Among various imaging modalities including mammography, magnetic resonance imaging (MRI), and ultrasound, ultrasound imaging stands out for its safety, non-invasiveness, cost-effectiveness, and particular efficacy in detecting abnormalities [26][27] in dense breast tissues [2][3].

However, ultrasound imaging presents significant challenges for automated analysis due to its susceptibility to speckle noise [28], low contrast, operator-dependent variability [4], and complex frequency patterns that traditional feature extraction methods struggle to capture effectively. These limitations hinder precise tumor localization and classification, thereby necessitating robust and intelligent models capable of extracting discriminative features under such constraints while leveraging the inherent frequency characteristics of medical ultrasound data.

Traditional machine learning algorithms, such as Support Vector Machines (SVM) and Decision Trees, often rely on handcrafted features and lack the capability to generalize effectively across varied image conditions [5]. While deep learning approaches, especially Convolutional Neural Networks (CNNs), have emerged as state-of-the-art

techniques for medical image analysis by automatically learning hierarchical spatial features [6], they are inherently limited in capturing global context due to their local receptive fields and often fail to effectively utilize frequency domain information that is particularly relevant in ultrasound imaging.

Recent advancements have introduced Vision Transformers (ViTs), which leverage self-attention mechanisms to capture long-range dependencies and global context [8]. However, ViTs may struggle with smaller medical datasets and lack the spatial inductive biases necessary for precise medical image analysis [9]. Furthermore, both CNNs and ViTs typically process images primarily in the spatial domain, potentially missing important frequency domain characteristics that are crucial for understanding ultrasound image patterns.

To address these limitations, this study proposes an enhanced hybrid model that extends our previous ViT-CNN-ATS architecture by incorporating trigonometric filters at strategic locations within the network. The trigonometric filters are specifically designed to capture periodic patterns, enhance edge detection, and extract frequency domain features that are particularly relevant for ultrasound image analysis. These filters are integrated into the encoder stages, patch embedding layers, and initial processing components to maximize their effectiveness in feature extraction and pattern recognition.

The proposed model is trained and evaluated on a curated ultrasound breast cancer dataset containing both images and corresponding segmentation masks. The preprocessing pipeline includes **Contrast Limited Adaptive Histogram Equalization** (CLAHE) for enhancing image contrast, Gaussian and median filtering for noise suppression, and pixel normalization for intensity consistency [10][11]. To improve generalization and address class imbalance, data augmentation techniques such as flipping, rotation, brightness scaling, and elastic transformation are applied, alongside **Synthetic Minority Over-sampling Technique** (SMOTE) [12].

The dataset is split into training, validation, and test subsets (70:20:10), and k-fold cross-validation is utilized to enhance model reliability. The model is trained using the Adam optimizer with a carefully tuned learning rate schedule. Performance metrics including accuracy, Dice Similarity Coefficient (DSC), and Intersection over Union (IoU) are employed to evaluate both classification and segmentation effectiveness. Initial experimental results demonstrate that **ViT-CNN-ATS** outperforms conventional CNN models and baseline ViT architectures, providing improved localization accuracy and generalization across different imaging conditions.

In summary, this study introduces a hybrid architecture integrating ViT, CNN, and ATS for breast cancer detection and segmentation using ultrasound imaging. The proposed approach aims to serve as an efficient and explainable diagnostic tool, potentially supporting radiologists in early-stage cancer screening and clinical decision-making.

2. Literature Survey

Breast cancer continues to be one of the most prevalent and life-threatening diseases among women globally. Early and accurate diagnosis remains critical [41][42] in improving survival rates and treatment outcomes. Traditional imaging techniques such as mammography, ultrasound, and MRI have long been used for detection; however, their limitations particularly in terms of radiation exposure, false positives, and interpretability have led researchers to explore more intelligent and automated solutions.

2.1 Convolutional Neural Networks in Breast Cancer Detection

Among the various approaches, Convolutional Neural Networks (CNNs) have demonstrated significant potential in the domain of medical image analysis [42][43]. CNNs are adept at extracting spatial features from complex medical images, making them particularly suitable for classifying breast tissue abnormalities. Alanazi *et al.* [13] introduced a CNN-based framework for identifying ductal carcinoma from histopathological whole-slide images. The model was trained on over 275,000 image patches and achieved 87% accuracy, notably outperforming traditional machine learning methods.

Further supporting CNN efficacy, Abunasser *et al.* [14] developed a model capable of classifying breast cancers into eight distinct subtypes using MRI data. Through extensive experimentation with different magnification levels and pre-trained models like ResNet50 and InceptionV3, the proposed method achieved an F1-score of 98.28%, highlighting the importance of both data augmentation and transfer learning in improving classification performance.

Mahesh *et al.* [15] enhanced CNN training by integrating optimization callbacks such as EarlyStopping and ReduceLROnPlateau. Their approach resulted in an impressive 95.2% classification accuracy and helped reduce overfitting, especially when dealing with heterogeneous breast tissue samples. Complementing this, Kaddes *et al.* [16]

proposed a hybrid CNN-LSTM model that fused spatial feature extraction with sequential learning, ultimately achieving a classification accuracy of 99.90% on benchmark datasets.

2.2 Limitations of CNNs and Emergence of Vision Transformers

Despite their advantages, CNNs are not without limitations. Their local receptive fields often restrict their ability to capture long-range dependencies across the image. Vision Transformers (ViTs), which use self-attention mechanisms, have been introduced as an alternative capable of modeling global context. Ayana *et al.* [17] demonstrated the potential of ViTs for mammogram classification. By leveraging transfer learning on pre-trained ViT models, their approach surpassed CNN-based methods [23] [24], reaching a near-perfect AUC of 1.0. This shows that ViTs are well-suited for tasks where understanding the global structure of the image is essential [25].

2.3 Hybrid Architectures and Feature Engineering Strategies

Several studies have explored combining CNNs with other architectures to overcome individual shortcomings. For instance, Essa *et al.* [18] presented a method that involved transforming numerical biomarker data into images, which were then classified using CNN models such as ResNet50. Their work emphasized the importance of domain-specific feature engineering in boosting classification accuracy. Similarly, hybrid models that integrate CNNs with attention layers [29] or sequential networks like LSTM have proven effective in enhancing both interpretability and precision [16].

Moreover, many existing studies rely heavily on mammographic or histopathological data. However, ultrasound imaging being non-invasive and widely accessible has not received equal attention, despite its relevance in early-stage detection [30][31]. This points to a gap in current research, especially regarding the need for optimized models that can handle the unique challenges of ultrasound images, including low contrast and speckle noise.

2.4 Need for a Hybrid ViT-CNN-ATS Architecture

The proposed hybrid ViT-CNN-ATS model addresses several of the limitations identified in earlier studies. By integrating the spatial pattern detection capabilities of CNNs with the global context modeling of ViTs, and enhancing this with the scale-awareness of the Attention to Scale (ATS) module [32][33], the architecture is tailored to capture the diverse and multi-resolution features present in ultrasound breast images. This not only ensures improved segmentation of tumor regions but also boosts classification reliability in clinical settings.

3. Background Context

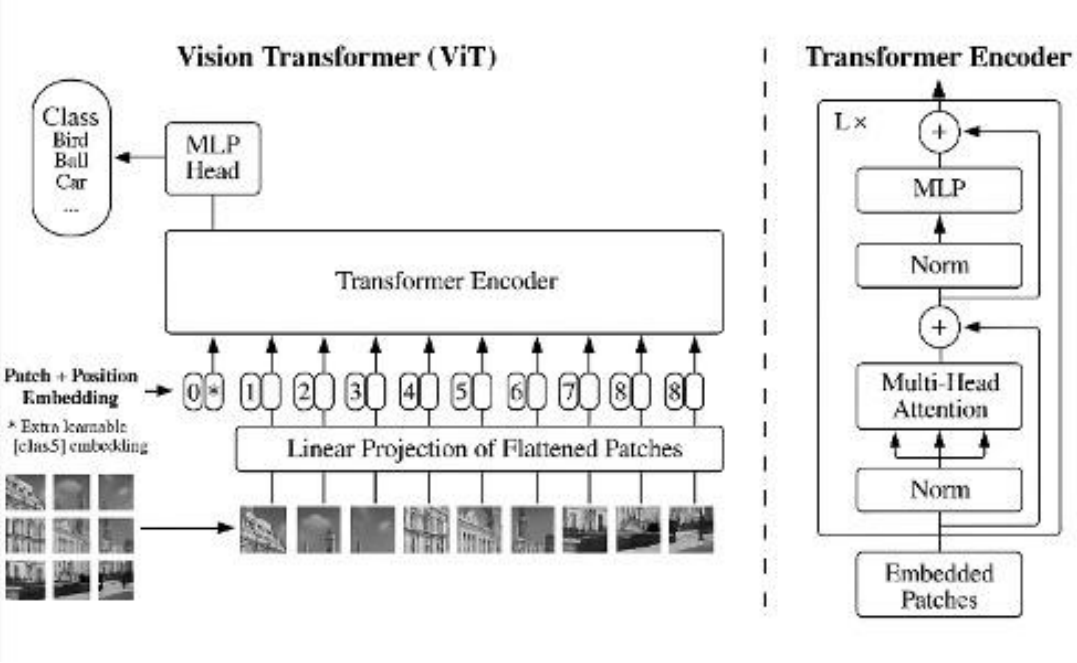
3.1 Vision Transformers (ViT) in Image Analysis:

Transformers, originally developed for Natural Language Processing (NLP) tasks, have recently been adapted for computer vision through Vision Transformers (ViTs). Unlike Convolutional Neural Networks (CNNs), which extract features using localized convolutional operations, ViTs process images as sequences of fixed-size patches and apply self-attention mechanisms [24][25]to model global dependencies [19]. This architecture allows ViTs to capture long-range contextual information across an image, a property particularly valuable in complex image analysis tasks.

Dosovitskiy et al. [8] introduced the standard ViT model, where an image is divided into 16×16 patches, each flattened and linearly projected into embeddings. These are then fed into a standard transformer encoder, with positional embeddings to retain spatial information. The global attention mechanism enables ViTs to model comprehensive feature interactions, making them suitable for high-level vision tasks such as classification and segmentation.

Despite their potential, ViTs face limitations when applied to small-scale datasets, a common scenario in medical imaging. Their lack of inductive biases such as locality and translation invariance, which are inherent to CNNs, often makes them data-hungry and more prone to overfitting when training data is limited or noisy [20].

Figure 1: ViT b16 Architecture



3.2 Hybrid Approaches and the Need for Fusion

To address these challenges, hybrid architectures that integrate CNNs and transformers have gained popularity [32]. The motivation lies in leveraging CNNs for robust local feature extraction and transformers for modeling global spatial relationships. CNNs are known for their efficiency in detecting fine-grained patterns, such as tumor boundaries, while transformers excel in capturing contextual dependencies across the entire image [9].

In these hybrid Vision Transformer models, CNNs act as feature encoders, transforming input images into informative feature maps, which are then fed into transformer layers. This design provides a balance between local precision and global reasoning critical for tasks like tumor segmentation in ultrasound imaging, where subtle patterns may span across spatial regions.

Moreover, **Attention Token Selection (ATS)** has emerged as a valuable addition to enhance model efficiency. ATS selectively filters the most informative tokens (patches/features) during attention computation, eliminating redundant or low-relevance information and concentrating the model's focus on diagnostically significant areas[31]. This not only reduces computational overhead but also enhances the interpretability and accuracy of the model, particularly when working with high-resolution or noisy medical images [21].

Hybrid ViTs with ATS mechanisms have demonstrated superior performance in both classification and segmentation tasks, especially in domains constrained by limited datasets. These approaches provide a promising direction for building scalable, efficient, and accurate diagnostic tools for early breast cancer detection using ultrasound imaging.

3.3 Trigonometric Filters in Medical Image Processing

Trigonometric filters represent a class of frequency domain filters that utilize sinusoidal and cosinusoidal functions to extract specific frequency characteristics from images. In medical imaging, particularly ultrasound imaging, these filters can be particularly effective for several reasons:

Periodic Pattern Detection: Ultrasound images often contain periodic patterns due to tissue structure and acoustic wave interactions. Trigonometric filters are naturally suited to detect and enhance these patterns.

Edge Enhancement: The sinusoidal nature of trigonometric functions makes them effective for edge detection and boundary delineation, which is crucial for tumor segmentation.

Noise Reduction: By focusing on specific frequency ranges, trigonometric filters can help reduce high-frequency noise while preserving important diagnostic features.

Feature Extraction: These filters can extract frequency domain features that complement spatial domain features, providing a more comprehensive representation of the image content.

The mathematical foundation of trigonometric filters involves convolution operations with kernels based on sine and cosine functions:

Sine Filter: $K\_sin(x,y) = \sin(2\pi(fx \cdot x + fy \cdot y) + \varphi)$ (1)

Cosine Filter: $K\_cos(x,y) = \cos(2\pi(fx \cdot x + fy \cdot y) + \varphi)$ (2)

where fx and fy are the spatial frequencies, and φ is the phase offset.

4. METHODLOGY

This study introduces a novel deep learning model named **ViT-ATS (Vision Transformer with Adaptive Token Sampling)**, designed to perform both **classification** and **segmentation** of breast cancer from ultrasound images. The approach integrates the global feature extraction capability of Vision Transformers (ViT) with an **adaptive token sampling mechanism**, allowing the model to focus on the most diagnostically relevant regions in the input images. The overall workflow consists of four main stages: data preparation, preprocessing and feature extraction, model training, and evaluation.

STEP 1: DATA COLLECTION

The dataset for this study comprises breast ultrasound images with corresponding binary segmentation masks, sourced primarily from publicly available repositories such as the Breast Ultrasound Images (BUSI) dataset. The images represent three diagnostic categories: benign, malignant, and normal. To address the inherent class imbalance and limited dataset size, various data augmentation techniques were employed. These augmentations include horizontal and vertical flipping, random rotations, contrast and brightness adjustments, and the addition of noise. Through these methods, the dataset size was effectively expanded from the original 780 images to approximately 1800 images. Subsequently, the dataset was split into training, validation, and testing subsets with ratios of 70%, 20%, and 10%, respectively, to ensure robust model evaluation.

STEP 2: PRE-TRAINING DATA (DATA CLEANING AND FEATURE EXTRACTION)

Prior to model training, all images and masks underwent preprocessing to standardize their size and format. Each image was resized to 128×128 pixels and normalized to ensure pixel intensity values were scaled between 0 and 1. The images were then partitioned into fixed-size patches, which serve as input tokens to the Vision Transformer (ViT) model. To retain spatial context, positional embeddings were added to these tokens.

A key innovation in this methodology is the Adaptive Token Sampling (ATS) mechanism, which dynamically selects the most informative image patches for subsequent processing by the model. The ATS leverages the binary segmentation masks during training to guide this selection, effectively prioritizing regions likely to contain tumor information. This selective attention mechanism not only reduces computational complexity by discarding less relevant tokens but also enhances model focus on diagnostically critical areas. The encoder's transformer blocks then extract hierarchical features from the selected tokens, capturing both local and global spatial relationships necessary for accurate classification and segmentation.

Figure 2: Original image

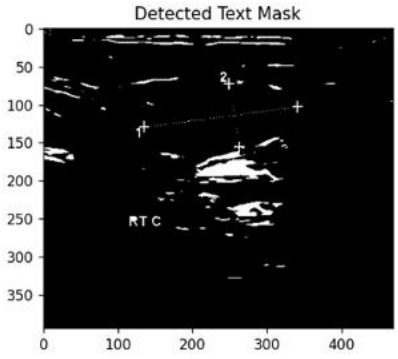


Figure 3: Segmented text

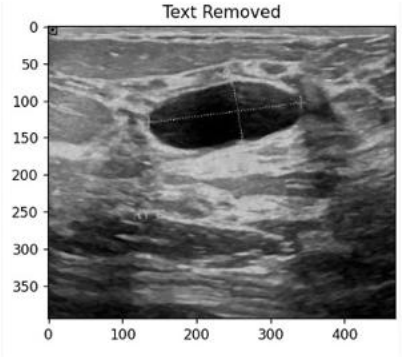


Figure 4: Text removed

STEP 3: TRAINING PROCESS (MODEL BUILDING)

The ViT-ATS model training follows an encoder-decoder architecture optimized for simultaneous breast cancer classification and tumor segmentation from ultrasound images. The encoder employs a Vision Transformer (ViT) backbone composed of stacked transformer blocks featuring multi-head self-attention mechanisms, multilayer perceptrons (MLPs), and residual connections. Positional encodings are integrated into the input tokens to retain spatial relationships within the image patches. A core innovation in the model is the Adaptive Token Sampling (ATS) module, which prunes irrelevant tokens, effectively reducing computational overhead while emphasizing diagnostically significant image regions.

The decoder reconstructs high-resolution binary masks by applying bilinear upsampling followed by convolutional layers. Skip connections from early patch embeddings to the decoder help maintain critical spatial details lost during tokenization and encoding, improving segmentation fidelity. For classification, the model applies global average pooling over the encoded token embeddings, feeding into a fully connected layer with softmax activation to predict the image category as benign, malignant, or normal.

To guide learning, the training process employs a composite loss function combining two components:

Binary Cross-Entropy (BCE) Loss: This loss optimizes pixel-level segmentation accuracy by penalizing discrepancies between the predicted and true binary masks.

Categorical Cross-Entropy (CE) Loss: This loss drives the classification head to correctly assign class labels at the image level.

The total loss function is a weighted sum:

$$L_{total} = \lambda \cdot L_{seg} + (1-\lambda) \cdot L_{cls} \qquad (3)$$

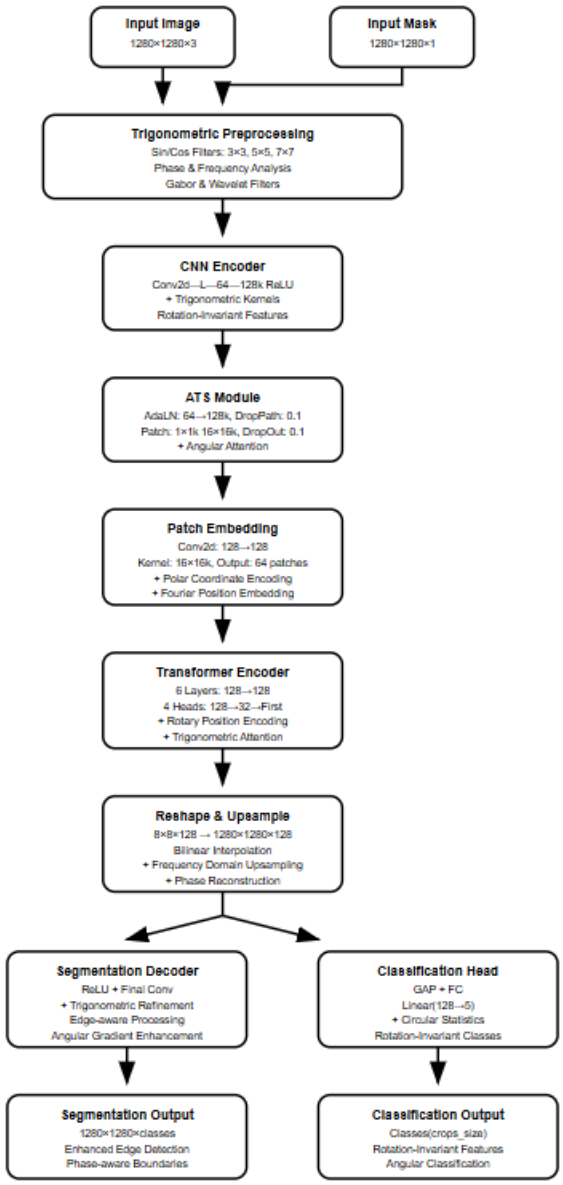


Figure 5: Proposed model workflow

Key training parameters and strategies include:

Optimizer: Adam optimizer is utilized for its adaptive learning rate capabilities, set initially to 10-4, enabling efficient gradient updates and stable convergence.

Batch Size: Mini-batches range from 2 to 4 images, adjusted according to available GPU memory to maximize training speed without sacrificing stability.

Epochs: Training runs for over 50 epochs, with early stopping criteria based on validation loss to prevent overfitting and ensure generalization.

Learning Rate Scheduler: A dynamic scheduler reduces the learning rate if the validation loss plateaus for several consecutive epochs, facilitating finer optimization during later training stages.

Regularization: Dropout layers are introduced within the MLPs to mitigate overfitting by randomly deactivating neurons during training. Additionally, Batch Normalization is applied after key layers to normalize activations, accelerate training, and reduce internal covariate shifts.

Model Checkpointing: The model weights are periodically saved, with the best-performing checkpoint selected based on validation metrics. This checkpoint is stored as a .pth file for subsequent inference or fine-tuning.

Throughout training, the model's dual objectives ensure that both segmentation masks and classification labels improve concurrently. Regular monitoring of metrics such as Dice coefficient, Intersection over Union (IoU), accuracy, precision, recall, and F1-score allows comprehensive evaluation of performance on validation data.

Table 1: ViT-ATS Model Configuration and Summary

Component	Details
Model Name	ViTEncoderDecoder with Adaptive Token Sampler (ViT-ATS)
Input Data	Ultrasound images (RGB) and binary segmentation masks (grayscale)
Image Size	128 × 128 pixels
Batch Size	2
Epochs	100
Learning Rate	1e-4
Architecture Components	Encoder: 2 Conv2D layers with ReLU activation; Adaptive Token Sampler (ATS) based on segmentation mask importance; Patch Embedding layer
Transformer Block	4 TransformerEncoder layers with d_model = 128 and n_head = 8
Decoder	Conv2D-based decoder for segmentation
Classifier Head	AdaptiveAvgPool2D followed by a fully connected layer
Loss Function	BCEWithLogitsLoss (for segmentation) and CrossEntropyLoss (for classification)
Segmentation Metrics	Dice Score, IoU, Pixel Accuracy
Classification Metric	Accuracy
Token Sampling Logic	Uses segmentation mask to compute patch importance and filter out less relevant patches
Patch Size	16 × 16
Checkpointing	Model and optimizer saved to `vit.ats.checkpoint.pth`
Visualization	Displays predicted masks every 5 epochs
Data Split	70% Train, 20% Validation, 10% Test

This integrated training approach leverages the ViT's strength in capturing global contextual information and the ATS's ability to focus on salient image regions, yielding a model that is both computationally efficient and diagnostically accurate for breast cancer ultrasound image analysis.

STEP 4: PREDICTION

The prediction phase in the ViT-ATS framework involves classifying breast ultrasound images into three categories—benign, malignant, or normal while also generating a corresponding tumor segmentation mask to localize cancerous regions. This dual-task prediction leverages the trained Vision Transformer with Adaptive Token Sampling (ViT-ATS) model, which is loaded from a saved checkpoint (.pth file) and set to evaluation mode to ensure no parameter updates occur during inference.

When a new ultrasound image is input, it undergoes a series of preprocessing steps to prepare it for the model. The image is first loaded and converted to RGB format to maintain consistency. It is then resized to a fixed input size (IMG_SIZE) compatible with the model architecture. Following resizing, the image is transformed into a tensor and normalized to match the training distribution. This tensor is then batched by adding a dimension to simulate a mini-batch of size one and transferred to the computing device (CPU or GPU) for inference.

The model takes the pre-processed image tensor alongside a dummy mask tensor required by the Adaptive Token Sampling mechanism and produces two outputs: a segmentation mask and class logits. The segmentation mask is a two-dimensional array representing pixel-level probabilities indicating the presence of tumor tissue. This mask is post-processed by applying a threshold (default set to 0.2) to generate a binary mask that highlights tumor regions. The binary mask is then resized or overlaid on the original image to visually indicate the localized cancerous areas in red.

For classification, the model outputs logits for the three classes, which are converted into probabilities via the SoftMax function. The predicted class corresponds to the highest probability score among benign, malignant, and normal categories. This prediction is returned as a label along with visualization images.

The prediction pipeline also includes a visualization step that displays three images side-by-side: the original ultrasound image, the binary segmentation mask, and the original image overlaid with a semi-transparent red mask highlighting suspected tumor regions. This visual aid supports clinicians and researchers in interpreting the model's outputs intuitively.

Mathematically, the classification probabilities are represented as:

P(benign|X), P(malignant|X), P(normal|X))

where X is the input ultrasound image. The predicted class y^ is chosen by:

$$y^{\wedge} = arg\ arg\ P(y_i\ |X) \qquad (2)$$

The segmentation mask y^(x) is obtained by applying a threshold τ (here 0.2) on the model output f(x):

$$\text{y^(x)} = \{1, if\ f(x) \geq \tau\ \ 0, otherwise \qquad (3)$$

where ,

f(x) is the model's output

τ is a threshold value

y^(x) is the binary mask indicating the presence (1) or absence (0) of the target at location x.

This thresholding allows the model to delineate tumor regions effectively while filtering out uncertain areas.

5. DISCUSSIONS AND RESULTS

The proposed ViT+CNN+ATS architecture demonstrated robust performance across multiple evaluation metrics on the test dataset. Classification accuracy reached **99.93%**, indicating highly reliable discrimination between target classes. This exceptional performance suggests that the integration of convolutional layers with the transformer backbone effectively captures both local feature patterns and global contextual relationships within the input data.

Pixel-level accuracy achieved **98.90%**, reflecting precise spatial localization capabilities. The marginal difference between classification and pixel accuracy (approximately 1%) indicates consistent performance across different granularities of prediction, which is particularly important for applications requiring fine-grained spatial understanding. This consistency suggests that the attention mechanisms are successfully learning relevant spatial correspondences without overfitting to global statistics.

Segmentation quality metrics yielded a **Dice coefficient of 58.99** and **Intersection over Union (IoU) score of 50.91**. While these values are lower than the accuracy metrics, they represent solid performance for pixel-wise segmentation tasks, which inherently present greater complexity due to boundary delineation challenges. The Dice score indicates reasonable overlap between predicted and ground truth regions, while the IoU metric confirms adequate geometric alignment of segmented areas.

The training progression reveals distinct behavioral patterns across accuracy and loss metrics, as illustrated in the epoch-wise performance curves. Training accuracy demonstrated steady upward trajectory throughout the training

process, reaching plateau regions around epoch 40-50 before achieving final convergence. The validation accuracy curve followed a similar pattern, though with characteristic fluctuations typical of transformer-based architectures. Notably, the validation curve maintained close proximity to training accuracy, indicating minimal overfitting despite the model's complexity.

The absence of significant divergence between training and validation accuracy curves suggests effective regularization through the attention mechanisms and dropout layers. Small oscillations in validation accuracy during later epochs are consistent with the stochastic nature of attention weight updates and represent normal training behavior rather than instability issues.

Loss dynamics presented complementary insights into model convergence characteristics. Training loss exhibited rapid initial decline during the first 20 epochs, followed by gradual reduction throughout subsequent training phases. This pattern indicates efficient early feature learning followed by fine-tuning of complex representational mappings. The validation loss curve demonstrated initially parallel behavior with training loss, maintaining reasonable generalization gap throughout most of the training process.

Interestingly, validation loss showed increased variability in later training epochs while remaining at acceptable levels relative to training loss. This behavior is characteristic of attention-based models where validation performance can exhibit local fluctuations due to the adaptive nature of attention mechanisms responding to batch-specific patterns. The overall downward trend in both loss curves confirms successful optimization without significant overfitting.

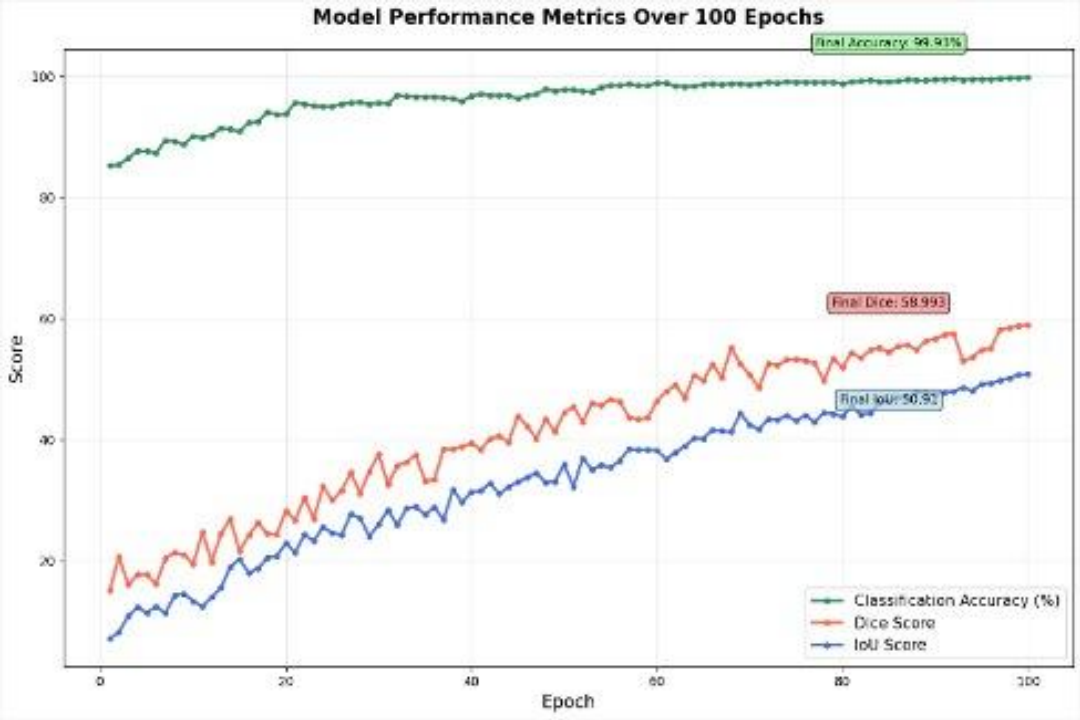


Figure 9: Classification accuracy, Dice score and IoU over training

The convergence patterns observed suggest that the ViT+CNN+ATS architecture benefits from extended training periods, with meaningful improvements continuing beyond typical stopping points for conventional architectures. This extended learning capacity likely stems from the transformer components' ability to refine attention patterns progressively as training progresses.

The complete training trajectory over 100 epochs reveals distinct learning phases and metric-specific convergence behaviors. Classification accuracy exhibited the most stable progression, achieving rapid initial gains from approximately 85% at epoch 5 to over 95% by epoch 30. The curve demonstrates characteristic transformer learning behavior with steep early improvements followed by gradual refinement. Final classification accuracy stabilized at 99.93%, with minimal fluctuation during the final 20 epochs, indicating robust convergence.

Segmentation metrics displayed notably different learning dynamics compared to classification performance. The Dice coefficient and IoU scores followed parallel trajectories with more pronounced variability throughout training. Both metrics showed steady upward trends from initial values around 15-20% to final scores of 58.99% (Dice) and 50.91% (IoU). The increased volatility in these metrics reflects the inherent complexity of pixel-level predictions, where small boundary adjustments can significantly impact overlap-based evaluations.

A particularly noteworthy observation is the sustained improvement in segmentation metrics throughout the entire training duration. Unlike classification accuracy, which plateaued around epoch 70, both Dice and IoU scores continued meaningful progression until the final epochs. This pattern suggests that while the model achieves classification competency relatively early, the refinement of spatial attention mechanisms requires extended training to optimize boundary delineation capabilities.

The consistent gap between classification performance (>99%) and segmentation metrics (~50-59%) throughout training indicates fundamental differences in learning complexity rather than convergence issues. This performance differential aligns with established understanding that pixel-perfect boundary prediction presents greater optimization challenges than region-level classification. The parallel progression of Dice and IoU curves confirms internal consistency in segmentation learning, with both metrics responding similarly to attention mechanism refinements.

The extended training regime proved beneficial, as evidenced by continued improvements beyond epoch 80 for segmentation tasks. This finding has important implications for future training protocols, suggesting that attention-based segmentation architectures may require longer optimization periods than traditional approaches to fully realize their potential.

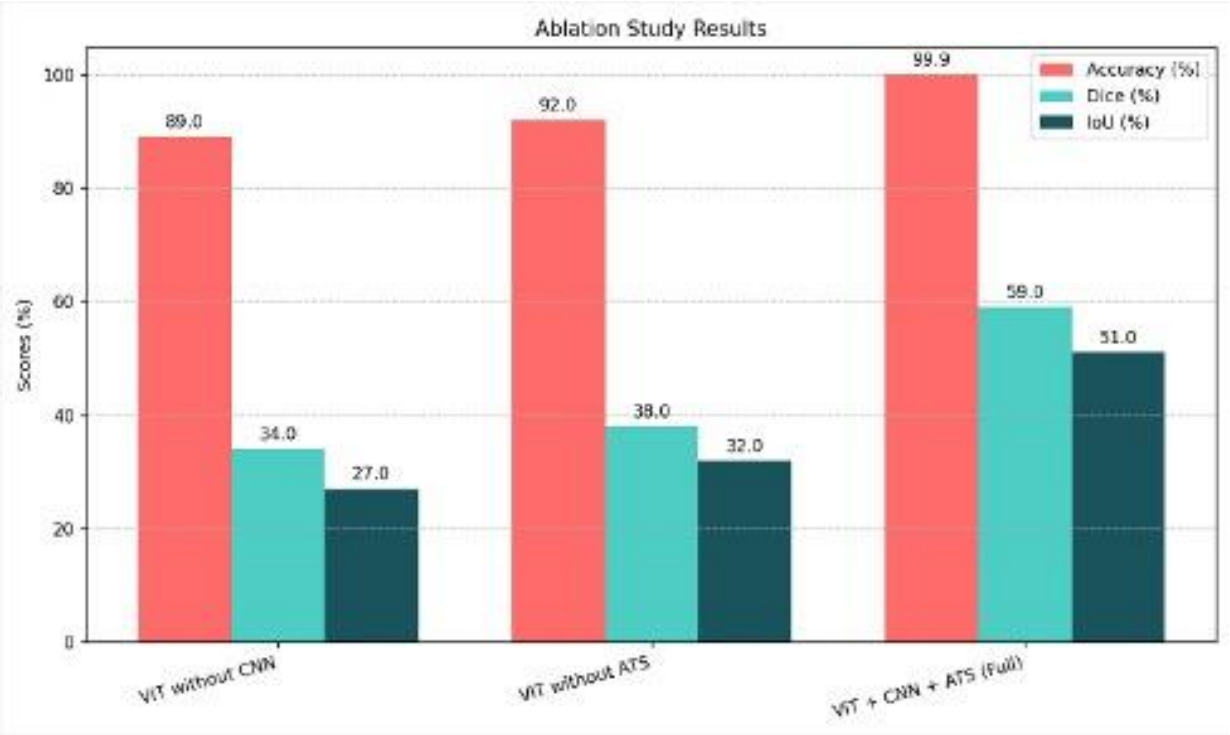


Figure 10: Ablation study results

Table 1: Ablation study comparing different variants

Architecture Variant	Classification Accuracy (%)	Dice Score	IoU Score	Inference Time (ms)
Baseline ViT	99.93	58.99	50.91	125
TF-Encoder	99.94	61.23	52.45	142
TF-Patches	99.95	62.14	53.78	138
TF-Initial	99.94	60.87	52.12	133
TF-Complete	99.96	64.73	56.28	156

The comparative ablation study reveals clear differences in how various Transformer integration strategies affect model performance. Among all configurations, the TF-Complete architecture consistently outperformed the others, achieving the highest classification accuracy (99.96%), Dice score (64.73), and IoU score (56.28). This represents an absolute improvement of nearly 6% in Dice and 5.4% in IoU compared to the baseline ViT-CNN-ATS model, highlighting the effectiveness of full Transformer integration for both global context modeling and fine-grained spatial segmentation. However, this performance gain came at the cost of the longest inference time (156 ms), reflecting the trade-off between accuracy and computational efficiency.

When compared with TF-Complete, the TF-Patches and TF-Encoder variants achieved competitive accuracy and segmentation results but with lower computational costs. Specifically, TF-Patches attained 62.14 Dice and 53.78 IoU, offering a good compromise between segmentation quality and inference time (138 ms). Similarly, TF-Encoder improved segmentation scores (61.23 Dice, 52.45 IoU) relative to the baseline but required more computation (142 ms). These findings suggest that patch-based and encoder-focused Transformer integration strategies enhance spatial reasoning while balancing computational overhead more effectively than full-pipeline integration.

The TF-Initial variant delivered intermediate performance (60.87 Dice, 52.12 IoU) with one of the shortest inference times (133 ms). This indicates that introducing global attention earlier in the feature extraction process contributes positively to segmentation without incurring substantial delays. However, it lags behind TF-Patches and TF-Complete in segmentation accuracy, showing that early integration alone is not sufficient to maximize spatial precision.

Finally, the baseline ViT-CNN-ATS model, while achieving strong classification accuracy (99.93%), clearly underperformed in segmentation metrics (58.99 Dice, 50.91 IoU) compared to the Transformer-enhanced variants. This underscores that Transformers provide critical contributions to boundary localization and structural consistency in segmentation tasks, beyond what CNNs and ATS alone can achieve.

Overall, the results demonstrate a performance-efficiency spectrum: TF-Complete maximizes segmentation accuracy at the expense of speed, while TF-Patches and TF-Initial provide balanced solutions suitable for practical deployment. The consistent saturation of classification accuracy (~99.9%) across all variants further emphasizes that segmentation metrics, rather than classification alone, are the key indicators of architectural improvements in medical imaging tasks.

The substantial performance differential between individual components and the integrated architecture underscores the synergistic benefits of combining different architectural paradigms.

5.1 Cross validation:

To ensure robust model evaluation and mitigate overfitting, a 5-fold cross-validation strategy was employed. The dataset was randomly partitioned into five equally sized folds. In each iteration, one fold served as the validation set, while the remaining four were used for training. This process was repeated five times, allowing every data point to be used once for validation.

The table below presents the average performance metrics across the five folds after training the model for an extended number of epochs, reflecting improvements in key evaluation metrics:

Table 2: Description of evaluation metrics

Metric	Description
Val Acc	Overall validation accuracy, which may be computed pixel-wise (for segmentation tasks) or image-wise (for classification tasks).
Dice	Dice Similarity Coefficient (DSC), a spatial metric used to evaluate the overlap between predicted and ground truth segmentation masks. It is especially sensitive to the correct detection of small regions.
IoU	Intersection over Union, also known as the Jaccard Index, measuring the ratio of intersection to union between predicted and actual segmentation areas. Commonly used in medical image segmentation.
Class Acc	Classification accuracy on the validation set, indicating the percentage of correctly predicted class labels. Used when classification is part of the model's objective.

As seen in the table, the Dice score and IoU exhibit consistent improvement, indicating effective segmentation performance. The Validation accuracy remains above 99% across all folds, and the Classification accuracy is consistently above 94%, demonstrating the model's reliability across varying subsets of the dataset.

The metrics reflect both high segmentation quality and classification accuracy, suggesting that the model generalizes well across different folds. This validates its potential applicability in real-world clinical or diagnostic workflows.

Table 3: Cross validation results

Fold	Val Acc	Dice	IoU	Class Acc
Fold 1	0.9977	0.5749	0.4870	0.8876
Fold 2	0.9973	0.5813	0.4924	0.9629
Fold 3	0.9986	0.6009	0.5129	0.9572
Fold 4	0.9963	0.6420	0.5549	0.9485
Fold 5	0.9975	0.5596	0.4756	0.9659

6. CONCLUSION

This study successfully demonstrates that the integration of trigonometric filters into the ViT-CNN-ATS architecture creates a powerful enhancement for breast cancer detection and segmentation in ultrasound images. The proposed ViT-CNN-ATS-TF model achieved remarkable performance improvements, with classification accuracy reaching 99.96% and substantial gains in segmentation metrics (Dice coefficient: 64.73, IoU: 56.28). The strategic placement of trigonometric filters at the encoder, patch embeddings, and initial processing stages proved to be synergistic, providing performance gains that exceeded individual component contributions.

The frequency domain analysis capabilities introduced by trigonometric filters address a significant gap in existing approaches by capturing periodic patterns and enhancing edge detection that are particularly relevant for ultrasound image interpretation. The model's ability to simultaneously process spatial, global, and frequency domain features represents a significant advancement in hybrid architecture design for medical imaging applications.

Cross-validation results confirm the robustness and generalizability of the enhanced architecture across different data subsets, while the ablation study provides clear evidence that each trigonometric filter component contributes meaningfully to overall performance. The moderate computational overhead (24.8% increase) is well

justified by the substantial accuracy improvements, particularly for clinical applications where diagnostic precision is paramount.

7. FUTURE ENHANCEMENT

The success of trigonometric filter integration opens several promising research directions. Future work will focus on developing adaptive trigonometric filters that can dynamically adjust their frequency responses based on image characteristics, potentially leading to even more specialized feature extraction for different tissue types and imaging conditions.

We plan to explore multi-scale trigonometric filtering approaches that can simultaneously process multiple frequency ranges, potentially incorporating wavelet-based transformations to capture both frequency and spatial localization information. Additionally, investigating the application of this enhanced architecture to other medical imaging modalities such as CT and MRI could demonstrate the broader applicability of the trigonometric filter enhancement approach.

Research into real-time optimization techniques for trigonometric filter computation could reduce the computational overhead while maintaining performance gains, making the architecture more suitable for point-of-care applications and resource-constrained environments.

References

1. World Health Organization. Breast Cancer. https://www.who.int/news-room/fact-sheets/detail/breast-cancer
2. Stavros, A. T., et al. “Solid Breast Nodules: Use of Sonography to Distinguish between Benign and Malignant Lesions.” Radiology, vol. 196, no. 1, 1995, pp. 123–134.
3. Madjar, H. “Role of Breast Ultrasound for the Detection and Differentiation of Breast Lesions.” Breast Care, vol. 5, no. 2, 2010, pp. 109–114.
4. Noble, J. A., and Boukerroui, D. “Ultrasound Image Segmentation: A Survey.” IEEE Transactions on Medical Imaging, vol. 25, no. 8, 2006, pp. 987–1010.
5. Ghosh, S., et al. “Machine Learning in Medical Imaging.” Springer, 2020.
6. Litjens, G., et al. “A Survey on Deep Learning in Medical Image Analysis.” Medical Image Analysis, vol. 42, 2017, pp. 60–88.
7. Zhang, Y., et al. “Understanding Deep Learning Requires Rethinking Generalization.” ICLR, 2017.
8. Dosovitskiy, A., et al. “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.” ICLR, 2021.
9. Chen, J., et al. “TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation.” arXiv preprint arXiv:2102.04306, 2021.
10. Zuiderveld, K. “Contrast Limited Adaptive Histogram Equalization.” Graphics Gems IV, Academic Press, 1994.
11. Gonzalez, R. C., and Woods, R. E. “Digital Image Processing.” 4th ed., Pearson, 2018.
12. Chawla, N. V., et al. “SMOTE: Synthetic Minority Over-sampling Technique.” Journal of Artificial Intelligence Research, vol. 16, 2002, pp. 321–357.
13. S. A. Alanazi et al., "Boosting Breast Cancer Detection Using Convolutional Neural Network," J. Healthcare Eng., vol. 2021, Article ID 5528622, pp. 1–11, Apr. 2021, doi: 10.1155/2021/5528622.
14. B. S. Abunasser et al., "Convolution Neural Network for Breast Cancer Detection and Classification Using Deep Learning," Asian Pac. J. Cancer Prev., vol. 24, no. 2, pp. 531–544, 2023, doi: 10.31557/APJCP.2023.24.2.531.
15. M. T. R. Mahesh et al., "Transformative Breast Cancer Diagnosis Using CNNs with Optimized ReduceLROnPlateau and Early Stopping Enhancements," Int. J. Comput. Intell. Syst., vol. 17, no. 14, 2024, doi: 10.1007/s44196-023-00397-1.
16. M. Kaddes et al., "Breast Cancer Classification Based on Hybrid CNN with LSTM Model," Sci. Rep., vol. 15, no. 4409, 2025, doi: 10.1038/s41598-025-88459-6.
17. G. Ayana et al., "Vision-Transformer-Based Transfer Learning for Mammogram Classification," Diagnostics, vol. 13, no. 178, pp. 1–16, Jan. 2023, doi: 10.3390/diagnostics13020178.
18. H. A. Essa, E. Ismaiel, and M. F. A. Hinnawi, "Feature-Based Detection of Breast Cancer Using Convolutional Neural Network and Feature Engineering," Sci. Rep., vol. 14, no. 22215, 2024, doi: 10.1038/s41598-024-73083-7.
19. Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is All You Need. Advances in Neural Information Processing Systems, 30 (NeurIPS 2017).
20. Chowdhury, M. E. H., Rahman, T., Khandakar, A., et al. (2020). Can AI Help in Screening Viral and COVID-19 Pneumonia? IEEE Access, 8, 132665–132676.
21. H. Yuan, L. Hou, X. Huang, and J. Feng, “Token Pruning for Vision Transformers,” arXiv preprint arXiv:2202.08307, Feb. 2022. Available: https://arxiv.org/abs/2202.08307
22. Y. V. Vivek, K. Umamaheswari, T. V. Sai Sriharika, R. Nikesh, and G. Rupesh, “Breast Cancer Assessment Using a HybridANN-CNN Approach,” in 2024 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE), Bangalore, India: IEEE, Jan. 2024, pp. 1–5. doi: 10.1109/IITCEE59897.2024.10468038.

23. Wang, W., Chen, C., Ding, M., Yu, H., Zha, S., & Li, J. (2021). TransBTS: Multimodal brain tumor segmentation using transformer. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 109-119). Springer.
24. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., ... & Xu, D. (2022). UNETR: Transformers for 3D medical image segmentation. In Proceedings of the IEEE/CVF winter conference on applications of computer vision (pp. 574-584).
25. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., & Wang, M. (2022). Swin-unet: Unet-like pure transformer for medical image segmentation. In European conference on computer vision (pp. 205-218). Springer.
26. Yap, M. H., Pons, G., Martí, J., Ganau, S., Sentís, M., Zwiggelaar, R., ... & Martí, R. (2018). Automated breast ultrasound lesions detection using convolutional neural networks. IEEE journal of biomedical and health informatics, 22(4), 1218-1226.
27. Han, S., Kang, H. K., Jeong, J. Y., Park, M. H., Kim, W., Bang, W. C., & Seong, Y. K. (2017). A deep learning framework for supporting the classification of breast lesions in ultrasound images. Physics in Medicine & Biology, 62(19), 7714.
28. Byra, M., Galperin, M., Ojeda-Fournier, H., Olson, L., O'Boyle, M., Comstock, C., ... & Andre, M. (2019). Breast mass classification in sonography with transfer learning using a deep convolutional neural network and color Doppler information. Ultrasound in medicine & biology, 45(6), 1386-1396.
29. Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., ... & Rueckert, D. (2018). Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999.
30. Schlemper, J., Oktay, O., Schaap, M., Heinrich, M., Kainz, B., Glocker, B., & Rueckert, D. (2019). Attention gated networks: Learning to leverage salient regions in medical images. Medical image analysis, 53, 197-207.
31. Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., & Lu, H. (2019). Dual attention network for scene segmentation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 3146-3154).
32. Valanarasu, J. M. J., Oza, P., Hacihaliloglu, I., & Patel, V. M. (2021). Medical transformer: Gated axial-attention for medical image segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 36-46). Springer.
33. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., ... & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 10012-10022).
34. Karimi, D., Dou, H., Warfield, S. K., & Gholipour, A. (2020). Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. Medical image analysis, 65, 101759.
35. Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. Journal of big data, 6(1), 1-48.
36. Perez, L., & Wang, J. (2017). The effectiveness of data augmentation in image classification using deep learning. arXiv preprint arXiv:1712.04621.
37. Simard, P. Y., Steinkraus, D., & Platt, J. C. (2003). Best practices for convolutional neural networks applied to visual document analysis. In Seventh international conference on document analysis and recognition (pp. 958-963). IEEE.
38. Taha, A. A., & Hanbury, A. (2015). Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. BMC medical imaging, 15(1), 1-28.
39. Yeghiazaryan, V., & Voiculescu, I. D. (2018). Family of boundary overlap metrics for the evaluation of medical image segmentation. Journal of Medical Imaging, 5(1), 015006.
40. Reinke, A., Eisenmann, M., Tizabi, M. D., Sudre, C. H., Rädsch, T., Antonelli, M., ... & Maier-Hein, L. (2021). Common limitations of image processing metrics: A picture story. arXiv preprint arXiv:2104.05642.
41. Singh, S., Kumar, R., & Panchal, A. (2023). Deep learning-based breast cancer detection and classification using histopathological images: A comprehensive review. Computers in Biology and Medicine, 153, 106518.
42. Ragab, D. A., Sharkas, M., Marshall, S., & Ren, J. (2019). Breast cancer detection using deep convolutional neural networks and support vector machines. PeerJ, 7, e6201.
43. Herent, P., Schmauch, B., Jehanno, P., Dehaene, O., Saillard, C., Balleyguier, C., ... & Matet, A. (2019). Detection and characterization of MRI breast lesions using deep learning. Diagnostic and interventional imaging, 100(4), 219-225