Journal of Information & Optimization Sciences
ISSN 0252-2667 (Print), ISSN 2169-0103 (Online)
Vol. 47 (2026), No. 5-A, pp. 1613–1621
DOI : 10.47974/JIOS-2251

Comparative evaluation of transformer and CNN architectures for breast cancer detection across histopathology, mammography, and ultrasound

Kalyani Ghuge *
D. Saravanan †
School of Computing Science & Engineering
VIT Bhopal University
Kothrikalan
Sehore 466114
Madhya Pradesh
India

Abstract

Automated analysis of images in various imaging modalities are beneficial in breast cancer screening and diagnosis. This paper compares four deep learning models Vision Transformer (ViT), EfficientNetV2 with CBAM, ConvMixer, and MobileNetV4 on three publicly available datasets of breast imaging (histopathology, mammography, ultrasound). We measure accuracy, recall, F1-score, computational cost, and suitability to modality. To underline the focus on the reproducibility and modality-specific analysis, the methodology underlines concise descriptions of every model in terms of mathematical intuition, architecture illustration, algorithmic steps. We have found that attention-augmented CNNs (CBAM-EfficientNetV2) perform well in terms of accuracy and robustness across modalities and MobileNetV4 has a high efficiency-accuracy trade-off to be deployed on resource-constrained devices.

Subject Classification: *03G10, 11T71.*

Keywords: *Breast cancer detection, Vision transformer, CBAM-EfficientNetV2, ConvMixer, MobileNetV4, Histopathology, Mammography, Ultrasound.*

I. Introduction

Breast cancer is one of the causes of cancer deaths in the world. The timely and correct diagnosis with medical imaging is a significant factor in patient outcomes. Various imaging techniques reveal different visual information:

* *E-mail:* **Kalyani.ghuge@gmail.com (Corresponding Author)**
† *E-mail:* **saravanan.d@vitbhopal.ac.in**

© Taru Publications

histopathology reveals cellular-level texture and morphology, mammography reveals macro-structures and calcifications and ultrasound reveals structural boundaries and echotexture at the speckle noise [1]. Automated classification solutions should thus be able to conform to modality-specific cues but should be resistant to noise and inter-patient differences [2], [3].

Over the past few years, there has been a plethora of deep architectures that have been used in breast imaging. Medical images now routinely use convolutional neural networks (CNNs) but there is a purely attention-based alternative, based on transformer-based models, including the Vision Transformer (ViT) [4]. Simultaneously, promising backbones such as EfficientNetV2 [4], attention mechanisms such as CBAM [5], patch-based CNNs such as ConvMixer [6] and mobile-friendly models such as MobileNetV4 [7] have appealing accuracy-to-cost trade-offs. In the majority of the available studies, however, the focus is on one of the modalities and the comparison of these various backbone families under a unified experimental setup is seldom done [8].

One of the pipelines that we are designing and on which we test four popular architectures ViT, CBAM-EfficientNetV2, ConvMixer, and MobileNetV4 on three popular breast imaging datasets is BreakHis histopathology images [9], CBIS-DDSM-based mammography [10], and the BUS ultrasound dataset [11].

II. Mathematical Models and Frameworks

A. *Vision Transformer (ViT)*

ViT applies a typical Transformer en- coder to a sequence of non-overlapping patches to an image to model long-range global dependencies [12]. This empowers ViT with learning texture relations and morphology:

MHSA:

$$Q = zWQ, K = zWK, V = zWV$$

$$MHSA(z) = Concat(h1, \ldots, hm)WO$$

Where each head:

Feed-forward block:

$$MLP(z) = W2\,\sigma(W1z + b1) + b2$$

Final classification uses a class token z_{CLS}:

$$p = \sigma\big(Wclsz(L)\big)$$

Further Architecture Description ViT architecture presented in Figure 1 consists of: 1) Patch embedding block 2) L Transformer encoder layers 3) Classification head. ViT does not have any locality bias, as opposed to CNNs; it uses global attention to learn relationships. This renders it very effective in identifying malignant texture clusters that are large in spatial extent [13].

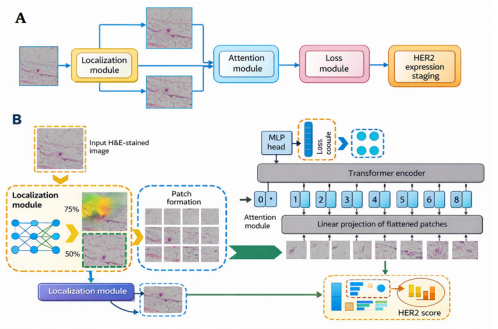


Figure 1
Vision Transformer (ViT) architecture

Algorithm 1: Enhanced ViT Training Algorithm

1: Initialize parameters

$$\theta = \{E, Epos, WQ, WK, WV, WO, Wcls\}$$

2: for epoch = 1 to E do
3: for each batch (x, y) do
4: Extract image patches xp from input x
5: Compute patch embeddings z0 using linear projection + positional encoding
6: for l = 1 to L do
8: $zl = MHSA(zl-1) + zl-1$
9: $zl = MLP(zl) + zl$
10: end for
11: Extract CLS token representation zCLS(L)
13: Compute Binary Cross Entropy (BCE) loss between prediction and label y
14: Update parameters θ using AdamW optimizer
15: end for
16: end for

B. *CBAM-EfficientNetV2*

EfficientNetV2 uses fused convolutions and progressive learning to achieve better speeds in training [3]. The addition of CBAM also adds dynamic channel and spatial attention that allows the model to selectively improve diagnostically significant areas like mass edges or micro-calcifications as shown in Figure 2 [2].

Channel Attention:

$$Mc(F) = \sigma\left(MLP(AvgPool(F)) + MLP(MaxPool(F))\right)$$

This generates per-channel weights which focus on informative feature maps.

Spatial Attention:

$$Ms(F) = \sigma f^{7\times7}([AvgPool(F); MaxPoolc(F)])$$

Combined Feature Refinement:

$$F' = Mc(F) \odot F, F"" = Ms(F') \odot F'$$

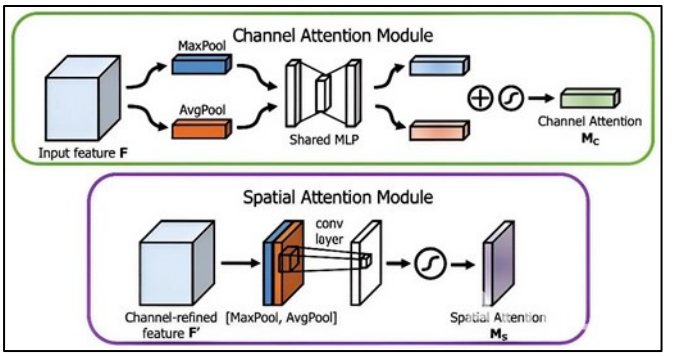


Figure 2
CBAM integrated into EfficientNetV2

C. *ConvMixer*

ConvMixer is an implementation of a violation embedding idea of ViT with pure convolutional mixing layers, generating an efficient substitute of CNNs and Transformers [4]. It switches depth wise (spatial mixing) and pointwise (channel mixing) convolutions as shown in Figure 3.

Patch Embedding:

$$z0 = ConvP,P(x)$$

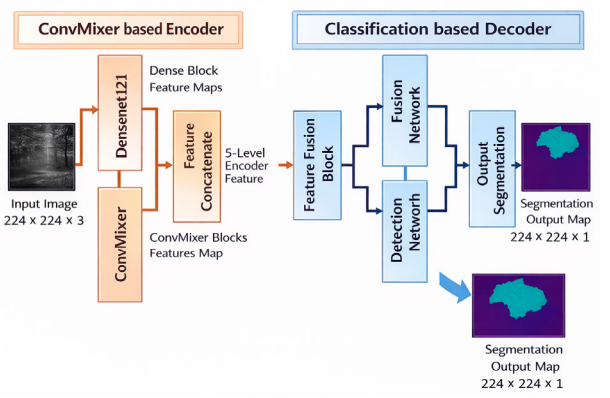


Figure 3
ConvMixer block-level architecture

ConvMixer Block:

$$u = z + DWConvk \times k(z), u = BN\big(GELU(u)\big)$$
$$z' = BN\Big(GELU\big(Conv1 \times 1(u)\big)\Big)$$

D. *MobileNetV4*

MobileNetV4 proposes Universal Inverted Bottlenecks (UIB), unified operator search, and effective scaling principles in mobile implementation [5]. It is also very accurate and lightweight such that it can be used in real time.

UIB Block: Given expansion factor t:

$$h = Conv1 \times 1(x \cdot t)$$
$$h' = DWConvk \times k(h) y = Conv1 \times 1(h')$$

If input and output shapes match:

$$y = y + x\ (residual)$$

III. Result and Discussion

A. *Dataset Used*

We have three publicly available datasets that are concerned with breast-cancer imaging:

- BreakHis (Breast Histopathology Images) High-resolution histopathology image patches of benign and malignant classes [6].
- CBIS-DDSM (Mammography)- Curated mammography cases, having updated segmentations, and pathology labels [7].
- Breast Ultrasound Dataset (BUS) Breast ultrasound images of benign/malignant/normal tissue lesions [8].

Table 1 provides an indisputable comparison of ViT Transformer, CBAM-EfficientNetV2, ConvMixer, and MobileNetV4 across mammography, histopathology, and ultrasound modalities in terms of cross datasets performance. CBAM-EfficientNetV2 has the highest performance and most importantly, in histopathology, the model has 95.20% accuracy, 94.80% recall and 94.90% F1-score, which is the best feature attention and localization ability. ViT Transformer has shown good performance with a high accuracy of 92.80 in mammography implying good global contextual learning.

MobileNetV4 demonstrates an intermediate performance as it shows a more promising result in ultrasound (90.80% accuracy), which is an indication of its effectiveness in feature extraction with low-level features. Conversely, ConvMixer has the least output of all datasets with an accuracy ranging between 80 and 83 percent which implies that ConvMixer has a low generalization aptitude. All in all, the attention-enhanced CNNs are better than the pure convolutional and mixers-based models, whereas the transformer-based learning guarantees consistent and strong cross-modal classification results with respect

to the breast imaging cases. The CBAM-EfficientNetV2 Figure 4 outperforms others overall, and ConvMixer is the worst in each of the modalities [14]-[16].

Table 1

Cross-Dataset Performance Analysis of ViT, CBAM-EfficientNetV2, ConvMixer, and MobileNetV4

Model	Dataset	Accuracy	Recall	F1-Score
ViT Transformer	Mammography	92.80%	91.50%	92.10%
	Histopathology	90.30%	89.20%	89.70%
	Ultrasound	89.50%	88.30%	88.90%
CBAM-EfficientNetV2	Histopathology	95.20%	94.80%	94.90%
	Mammography	93.80%	93.10%	93.40%
	Ultrasound	91.60%	90.80%	90.20%
ConvMixer	Histopathology	80%	79%	77%
	Mammography	82%	81%	80%
	Ultrasound	83%	84%	79.90%
MobileNetV4	Histopathology	86%	85%	82%
	Mammography	82%	81%	81%
	Ultrasound	90.80%	88.70%	87%

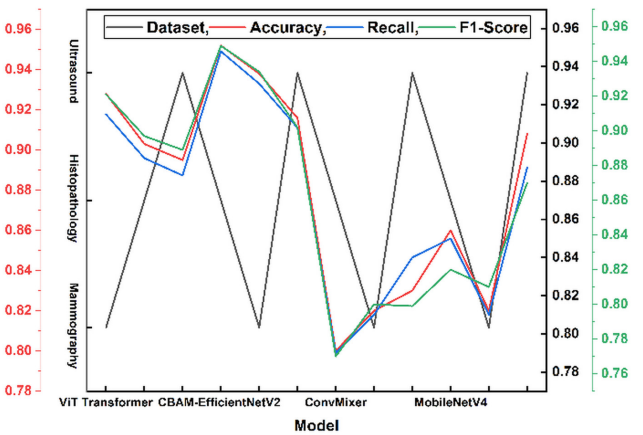


Figure 4

Multi-Modal Performance Comparison of Deep Learning Models Across Breast Imaging Datasets

Table 2 is a performance comparative analysis of ViT, CBAM-EfficientNetV2, ConvMixer, and MobileNetV4 on histopathology, ultrasound, and mammography datasets in terms of modality-specific performance. CBAM-EfficientNetV2 is the most effective model, especially in histopathology (95.20% accuracy, 94.90% F1-score), which proves the usefulness of attention mechanisms in analysing fine-grained tissues.

Table 2

Cross-Modal Classification Performance Analysis Using Accuracy, Recall, and F1-Score

Dataset	Model	Accuracy	Recall	F1-Score
Histopathology	ViT	90.30%	89.20%	89.70%
	CBAM-EffNetV2	95.20%	94.80%	94.90%
	ConvMixer	80%	79%	77%
	MobileNetV4	86%	85%	82%
Ultrasound	ViT	89.50%	88.30%	88.90%
	CBAM-EffNetV2	91.60%	90.80%	90.20%
	ConvMixer	83%	84%	79.90%
	MobileNetV4	90.80%	88.70%	87%
Mammography	ViT	92.80%	91.50%	92.10%
	CBAM-EffNetV2	93.80%	93.10%	93.40%
	ConvMixer	82%	81%	80%
	MobileNetV4	82%	81%	81%

ViT demonstrates good and consistent performance in all modalities, particularly in mammography (92.80% accuracy) which is a sign of effective global feature learning. MobileNetV4 offers competitive outcomes in ultrasound imaging (90.80% accuracy), which demonstrates its effectiveness in working with data with a high level of texture. ConvMixer documents relatively smaller values on datasets, which implies reduced versatility. On the whole, the attention-enhanced architectures are better than the traditional convolutional designs in cross-modal classification of breast images. The overall measures of CBAM-EfficientNetV2 are the highest, ViT measures are stable, and ConvMixer measures are relatively low as indicated in Figure 5.

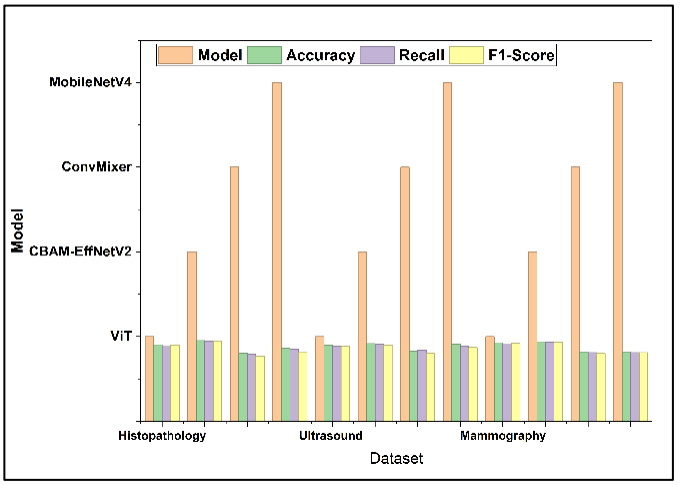


Figure 5

Dataset-Wise Bar Chart Comparison of Model Performance Metrics

IV. Conclusion

We made a systematic comparison of ViT, CBAM-EfficientNetV2, ConvMixer, and MobileNetV4 on three datasets of breast imaging. CBAM-EfficientNetV2 has the highest accuracy and robustness trade-off. MobileNetV4 offers a viable option in the low-resource environment. ViT is better on histopathology with higher compu- cational costs. ConvMixer is a small-scale CNN that is patch-based. Future research involves multi-task learning, domain adaptation and prospective clinical assessment.

References

[1] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.* (ICLR) (2021).

[2] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis*. (ECCV), pp. 3–19 (2018).

[3] M. Tan and Q. V. Le, "EfficientNetV2: Smaller models and faster training," in *Proc. Int. Conf. Mach. Learn*. (ICML), pp. 10096–10106 (2021).

[4] A. Trockman and J. Z. Kolter, "Patches are all you need?," *Trans. Mach. Learn. Res*. (2022).

[5] D. Qin, C. Leichner, M. Delakis, M. Fornoni, S. Luo, F. Yang, W. Wang, C. Banbury, C. Ye, B. Akin, V. Aggarwal, T. Zhu, D. Moro, and A. Howard, "MobileNetV4 – Universal models for the mobile ecosystem," in *Proc. Eur. Conf. Comput. Vis.* (ECCV) (2024).

[6] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, "A dataset for breast cancer histopathological image classification," *IEEE Trans. Biomed. Eng.,* vol. 63, no. 7, pp. 1455–1462 (2016).

[7] R. S. Lee, F. Gimenez, A. Hoogi, K. K. Miyake, M. Gorovoy, and D. L. Rubin, "A curated mammography dataset for use in computer-aided detection and diagnosis research," *Sci. Data,* vol. 4, pp. 170177 (2017).

[8] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy, "Dataset of breast ultrasound images," *Data Brief,* vol. 28, pp. 104863 (2020).

[9] T. Araújo, G. Aresta, E. Castro, J. Rouco, P. Aguiar, C. Eloy, A. Polónia, and A. Campilho, "Classification of breast cancer histology images using convolutional neural networks," *PLOS ONE*, vol. 12, no. 6, pp. e0177544 (2017).

[10] M. Z. Alom, C. Yakopcic, M. S. Nasrin, T. M. Taha, and V. K. Asari, "Breast cancer classification from histopathological images with inception recurrent residual convolutional neural network," *J. Digit. Imaging*, vol. 32, no. 4, pp. 605–617 (2019), doi: 10.1007/s10278-019-00182-7.

[11] Y. Jiménez-Gaona, M. J. Rodríguez-Álvarez, and V. Lakshminarayanan, "Deep-learning-based computer-aided systems for breast cancer imaging: A critical review," *Appl. Sci.*, vol. 10, no. 22, pp. 8298 (2020).

[12] S.-Y. Kim, Y. Choi, E.-K. Kim, B.-K. Han, J. H. Yoon, J. S. Choi, and J. M. Chang, "Deep learning-based computer-aided diagnosis in screening breast ultrasound to reduce false-positive diagnoses," *Sci. Rep.*, vol. 11, no. 1, pp. 395 (2021).

[13] A. Fiaz, M. A. Khan, and M. Sharif, "A deep fusion-based vision transformer for breast cancer classification," *Healthcare Technol. Lett.* (2024).

[14] D. Dhabliya, A. Lavhale, S. Thakur, R. M. Gomathi, A. Kumar, and S. De, "Enhancing cryptographic security through zero-knowledge proofs in theoretical mathematics," *J. Discrete Math. Sci. Cryptogr.*, vol. 29, no. 2-B, pp. 913–921 (2026), doi: 10.47974/JDMSC-2542.

[15] A. K. Rajpoot, D. Dhabliya, V. N. Trupti, T. A. Wani, K. Gupta, and L. Lakshmanan, "Quantum cryptography: Assessing the security and practicality of quantum key distribution in modern communication networks," *J. Discrete Math. Sci. Cryptogr.*, vol. 29, no. 2-B, pp. 979–986 (2026), doi: 10.47974/JDMSC-2549.

[16] G. Kaur, M. Kaur, J. Singh, M. M. Laishram, S. Lakhanpal, and M. Kumar, "Effective strategies for writing maintainable code in cryptographic software development," *J. Discrete Math. Sci. Cryptogr.*, vol. 29, no. 2-A, pp. 609–616 (2026), doi: 10.47974/JDMSC-2503.

Received April, 2025