JADIMI:
JOURNAL OF AI-DRIVEN INFORMATICS AND MANAGEMENT INFORMATION
Vol. 1 No. 1, August 2026, pp 12-19

Comparison of Lightweight CNN and Transformer-Based Models for Mammogram Image Classification on CBIS-DDSM

Abdullah[1]*, Muhammad Varel Arifianta[1], Frenwin[1], Gena Darma[1]

[1] *School of Computing, Telkom University, Bandung, Indonesia, 40257*

[*] *Corresponding author: gensdarmaa@student.telkomuniversity.ac.id*

Abstract*—Detecting malignancy in mammograms requires models capable of recognizing local lesion features while preserving broader breast-tissue context. This study compares DenseNet121, EfficientNetV2-S, and MobileViT-S for binary classification of benign and malignant lesions in the Curated Breast Imaging Subset of the Digital Database for Screening Mammography (CBIS-DDSM). Region-of-interest images were converted to single-channel grayscale inputs of 224 × 224 pixels, enhanced using contrast-limited adaptive histogram equalization, and augmented through geometric and intensity transformations. Patient-level stratified group five-fold cross-validation was used to prevent patient overlap between training and validation folds. Probabilities from the five fold-specific models were averaged through CV-bagging before evaluation on an independent test set of 298 images from 145 patients. EfficientNetV2-S achieved the highest accuracy, macro-F1, and specificity at 0.6913, 0.6837, and 0.7412, respectively. DenseNet121 achieved the highest sensitivity and ROC-AUC at 0.6953 and 0.7446, respectively, and produced the fewest false-negative malignant predictions. MobileViT-S achieved 0.6510 accuracy and 0.7368 ROC-AUC with 4.94 million parameters, 2.8250 GFLOPs per model, and the lowest directly measured five-model CV-bagging inference latency at 11.2721 ms/image, compared with 14.3409 ms/image for EfficientNetV2-S and 15.5444 ms/image for DenseNet121. No architecture was best on every criterion: EfficientNetV2-S provided the strongest overall classification balance, DenseNet121 favored malignant-case sensitivity, and MobileViT-S had the most favorable computational profile. These findings constitute experimental evidence on CBIS-DDSM ROI data and require external, multi-institutional validation before clinical interpretation.

Keywords—mammogram classification; breast cancer; lightweight models; CNN; Transformer

Manuscript received 11 Aug. 2026; revised 21 Aug. 2026; accepted 23 Aug. 2026. Date of publication 23 Aug. 2026.

I. INTRODUCTION

Breast cancer is a major global health concern and the most frequently diagnosed cancer among women. GLOBOCAN estimates for 2022 reported approximately 2.3 million new cases and 666,000 deaths from female breast cancer, accounting for 11.6% of all cancer cases worldwide [1]. Prognosis is strongly influenced by disease stage because survival is substantially more favorable when cancer remains localized than after it has metastasized. These observations underscore the importance of early detection and screening programs capable of reaching broad populations[2].

Mammography is a key component of early detection; however, its interpretation requires considerable time and expertise. In a simulation involving 114,421 examinations, an AI-based protocol maintained comparable sensitivity, increased specificity from 98.1% to 98.6%, and reduced the reading workload by 62.6%[3]. The randomized MASAI trial likewise reported comparable cancer detection with an approximately 44% reduction in workload[4], while ScreenTrustCAD, involving 55,581 women, showed that the combination of one radiologist and AI was noninferior to two radiologists[5]. AI is therefore positioned as a decision-support tool rather than a replacement for clinical judgment.

These advances have largely been driven by convolutional neural networks (CNNs). Recent studies indicate that both CNNs and Vision Transformers are promising for mammogram classification, although their performance depends on the architecture, input representation, and evaluation protocol[6] . Mammo-

Light demonstrates the relevance of compact CNNs[7], while annotation-efficient approaches have been applied to mammography and digital breast tomosynthesis[8]. The Mirai model extends the use of mammograms to multihorizon risk prediction[9]. DenseNet employs dense connections to promote feature reuse[6], whereas EfficientNetV2 emphasizes training efficiency[10]. Transformers have been applied to multiview mammography[11][12][13] , while MobileViT-S combines global modeling with convolution-based local feature extraction[14].

CBIS-DDSM is widely used in mammogram classification research because it provides curated images, lesion annotations, and pathology labels[2][6]. Nevertheless, a single patient may be represented by multiple images, projections, breast sides, and annotations. If the data are split at the image level, anatomical information from the same patient may appear in both the training and test sets, leading to overly optimistic performance estimates. Such leakage is a recognized contributor to the reproducibility crisis in machine-learning-based research[15]. Patient-level partitioning must therefore be performed before augmentation so that all images from the same patient remain within a single partition[2].

Direct comparison across mammography studies is difficult because reported performance depends on input representation, preprocessing, data partitioning, and evaluation protocol. The gap addressed here is therefore not a new architecture, but the lack of a controlled comparison between efficient CNNs and a lightweight CNN–Transformer model under the same patient-level protocol. DenseNet121 was selected as an established CNN baseline with dense feature reuse, EfficientNetV2-S as an efficiency-oriented modern CNN, and MobileViT-S as a lightweight hybrid that combines local convolutional processing with global Transformer modeling. In this study, lightweight refers to models selected with deployment cost in mind rather than an identical parameter budget; parameter count and FLOPs are reported explicitly because the three models differ in size.

Accordingly, this study compares benign–malignant discrimination, performance–efficiency trade-offs, and fold-level variability under a common protocol using 224×224 grayscale ROIs, patient-level stratified group five-fold cross-validation, the same training configuration, and an independent test set. The contribution is a leakage-aware comparison with CV-bagging and joint reporting of predictive and computational metrics. The scope is limited to ROI classification and does not include segmentation, localization, clinical decision-making, or external validation.

II. RELATED WORK

Deep learning has significantly advanced mammographic image analysis, particularly in lesion classification and cancer detection. These tasks remain challenging because malignant signs such as masses, irregular margins, and microcalcifications can be subtle or obscured by variations in breast density. The availability of curated datasets has supported progress in this area. CBIS-DDSM provides standardized mammograms, lesion annotations, metadata, and pathology labels that enable researchers to develop and compare learning-based classification systems [2], [6].

Image preprocessing has also played an important role in improving the visibility of mammographic abnormalities. Because mammograms are grayscale images acquired under varying conditions, their intensity and contrast distributions may differ considerably. Recent EfficientNetV2-based mammography research has continued to use contrast-limited adaptive histogram equalization (CLAHE) to enhance local lesion contrast while limiting excessive noise amplification [16]. Such findings support the use of contrast enhancement, image normalization, and data augmentation as practical steps for preparing mammograms for deep-learning models. Transfer-learning studies have also shown that combining learned deep features with complementary mammographic descriptors can strengthen lesion classification when labeled data are limited [17].

Convolutional neural networks (CNNs) have formed the foundation of much of the research on automated mammogram classification. Comparative experiments involving CNNs and Transformers have shown that architecture selection, image representation, and evaluation design can substantially influence the reported results [6]. DenseNet is a notable CNN in this context because its densely connected layers encourage feature reuse and improve information flow, allowing the network to retain both low-level texture information and higher-level lesion characteristics. DenseNet121 therefore provided an established CNN baseline for evaluating mammogram classification performance. DenseNet121 has also been used as a backbone for benign–malignant breast-lesion classification in a multicentre imaging study, supporting its use as a comparative CNN baseline [18].

The development of EfficientNetV2 shifted attention toward models that balance predictive performance with training and inference efficiency. EfficientNetV2 combines fused convolutional blocks and mobile inverted bottleneck blocks with a scaling strategy that considers network depth, width, and input resolution [10]. These design choices make EfficientNetV2-S relevant to mammography studies that require strong feature extraction without relying on an excessively large architecture. An EfficientNet-based two-view model has also been evaluated on CBIS-DDSM using transfer

learning and five-fold cross-validation, further supporting the use of efficient convolutional architectures for mammographic analysis [19].

Transformer-based approaches subsequently broadened mammogram analysis by modeling relationships across distant image regions and multiple mammographic views. Chen et al. reported an AUC of 0.818 for a Transformer-based Multiview model, compared with 0.784 for its CNN baseline, demonstrating the potential benefit of incorporating global and cross-view information [11]. Later studies extended this direction through a multiview Swin Transformer [12] and graph and Transformer-based architectures designed to capture relationships among mammographic views [13]. Vision-Transformer-based transfer learning has likewise been investigated for benign–malignant mammogram classification, indicating the value of global-context modeling in this task [20]. Although these approaches can model broad contextual information, their computational requirements may limit their use on resource-constrained hardware.

MobileViT was introduced to reduce this limitation by combining convolutional processing of local image patterns with Transformer-based modeling of global relationships in a lightweight architecture [14]. This combination is particularly relevant to mammograms, in which a classifier must recognize fine lesion details while retaining information about the surrounding tissue. The MobileViT-S variant consequently offers an appropriate Transformer-based counterpart to lightweight CNNs when both classification performance and deployment cost are considered.

Recent CBIS-DDSM studies have evaluated CNNs, EfficientNet-based models, and Vision Transformers, but direct comparison remains difficult because task definitions, view configurations, preprocessing, and data partitions differ [6], [19], [20]. Results from two-view or differently partitioned setups are therefore not directly comparable with the present ROI-based patient-level protocol. This study focuses on a within-protocol comparison of DenseNet121, EfficientNetV2-S, and MobileViT-S using the same grayscale preprocessing, patient-level five-fold cross-validation, transfer-learning strategy, and CV-bagging. Predictive performance is considered together with parameter count, FLOPs, and inference cost rather than as a state-of-the-art claim.

III. RESEARCH METHODOLOGY

This section describes the methodology in accordance with the implementation and experimental artifacts of the three models.

A. Research Design

This study employed comparative experimental design. The workflow comprised data preparation, preprocessing, construction of five patient-level folds, training of the three models, selection of the best checkpoint, inference on the test data, probability aggregation, and metric analysis. Each model was trained independently using the same random seed and hyperparameters. The complete experimental workflow is shown in Fig. 1.

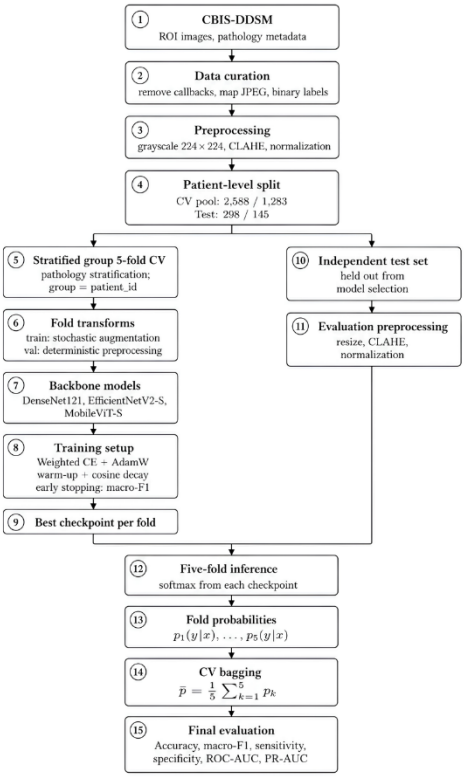

Figure. 1 Workflow of the research methodology and experimental procedure

B. Dataset and Data Partitioning

The data was obtained from CBIS-DDSM. The original training and validation sets were combined into a cross-validation (CV) pool, whereas the test set was retained and excluded from model selection. Samples labeled benign_without_callback were removed, and only samples with available JPEG ROI files were included. The resulting dataset is summarized in Table I.

TABLE I
SUMMARY OF DATASET PARTITIONING

Data Partition	No. of Images	No. of Patients
CV pool	2,588	1,283
Independent Test Set	298	145
Total	2,886	1,428

The independent test set comprised 170 benign and 128 malignant images as shown in Fig. 2. All architectures and fold-specific models were evaluated using the same test-set composition.

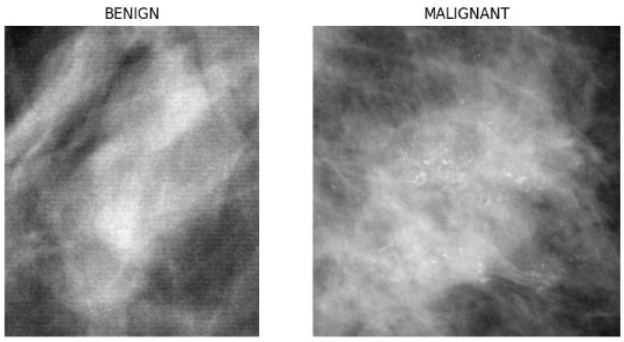

Figure. 2 Image samples from the CBIS-DDSM dataset

C. Preprocessing and Augmentation

Images were loaded in grayscale, resized to 224 × 224 pixels, processed using CLAHE, normalized, and converted into tensors. Augmentation was applied only to the training data. The transformation configuration is presented in Table II.

TABLE II
PREPROCESSING AND AUGMENTATION CONFIGURATION

Transformation	Main Parameter	Prob.
Resize	224x224	1.0
CLAHE	Default configuration	1.0
Horizontal flip	-	0.6
Rotation	Limit of 25°	0.7
Brightness/ Contrast	Limit of 0.25	0.7
Gaussian noise	Default configuration	0.4
Blur	Maximum kernel size of 5	0.4
Perspective	Scale of 0.05–0.10	0.3
Normalization	Default configuration	1.0

D. Fold Construction

Image indices in the CV pool were divided into five folds using *StratifiedGroupKFold*, with shuffling enabled and a random seed of 42. Pathology labels were used as the stratification variable, while *patient_id* defined the groups. In each iteration, four folds served as the training set and one as the validation set. No patient appeared in both partitions.

E. Model Configuration

The three models were constructed using the timm library with pretrained weights and two output neurons. DenseNet121 and EfficientNetV2-S were configured as CNN-based models, whereas MobileViT-S was configured as a Transformer-based model. The EfficientNetV2-S implementation used the tf_efficientnetv2_s identifier, while MobileViT-S used mobilevit_s. The first convolutional layer was adapted from three input channels to one using (3). The final classification layer produced logits for the benign and malignant classes. The measured model characteristics are presented in Table III; the reported FLOPs correspond to one image processed by one model.

TABLE III
MODEL CHARACTERISTICS

Model	Parameters (million)	GFLOPs/image	Artifact size (MB)
DenseNet121	6.9496	5.5089	28.41
EfficientNetV2-S	20.1796	5.6817	81.61
MobileViT-S	4.9386	2.8250	19.93

F. Training Strategy

Each model was trained using weighted cross-entropy and the AdamW optimizer. The initial learning rate was 5×10^{-5}, and the weight decay was 10^{-3}. The first three epochs employed linear warm-up with an initial factor of 1/3, followed by cosine annealing to a minimum learning rate of 10^{-6}. The models were trained for up to 30 epochs with batch sizes of 16 for training and 32 for validation and testing. Validation macro-F1 was monitored after each epoch. The checkpoint with the highest observed validation macro-F1 in each fold was retained, and training stopped after five consecutive epochs without improvement. The shared hyperparameter configuration was used to keep the comparison controlled, although it may not be optimal for every architecture.

TABLE IV
TRAINING HYPERPARAMETERS

Hyperparameter	Value
Number of folds	5
Maximum epochs	30
Training/validation batch size	16/32
Optimizer	AdamW
Initial learning rate	5×10^{-5}
Weight decay	10^{-3}
Early stopping patience	5 epochs
Loss function	Weighted cross-entropy
Best-model criterion	Validation macro-F1
Seed	42

G. Cross-Validation Bagging

The best model from the kth fold produced class probabilities for the test data. The final probability for class c was obtained by averaging the fold-specific probabilities, as defined in (1):

$$p^{-}(y = c \mid x) = \frac{1}{K}\sum_{k=1}^{K} p_k(y = c \mid x) \qquad (1)$$

where $K = 5$. The predicted label was assigned to the class with the highest mean probability. This procedure

reduces the dependence of the predictions on any single data split. Ensemble-based mammography studies have similarly reported that aggregating multiple model outputs can improve robustness compared with individual systems [21].

H. Evaluation Metrics

The evaluation included accuracy, macro-F1, malignant-class precision, sensitivity, specificity, and ROC-AUC. Reporting both discrimination and class-wise metrics, together with clear data-partitioning and reproducibility details, is consistent with current medical-imaging AI reporting guidance [22]. Accuracy was calculated using (2):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

Sensitivity, which represents the proportion of actual positive samples correctly identified, was calculated using (3):

$$Sensitivity = \frac{TP}{TP + FN} \quad (3)$$

Specificity, which measures the proportion of actual negative samples correctly identified, was calculated using (4):

$$Specificity = \frac{TN}{TN + FP} \quad (4)$$

Precision for the malignant class was calculated using (5):

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

The ROC-AUC score, representing the area under the receiver operating characteristic curve, was computed using (6):

$$ROC - AUC = \int_0^1 TPR\left(FPR^{-1}(x)\right)dx \quad (6)$$

Finally, macro-F1 was calculated using (7) by averaging the class-wise F1 scores across all C classes:

$$Macro - F1 = \frac{1}{C}\sum_{c=1}^{C}\frac{2P_cR_c}{P_c + R_c} \quad (7)$$

Here, TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively, with malignant designated as the positive class. Accuracy measures the proportion of all correct predictions; sensitivity is the proportion of malignant cases correctly identified; specificity is the proportion of benign cases correctly identified; and precision is the proportion of predicted malignant cases that are actually malignant.

For ROC-AUC, TPR and FPR are the true-positive and false-positive rates, FPR^{-1} denotes the inverse false-positive-rate function, and x is the integration variable over the unit interval. For macro-F1, C is the number of classes, while P_c and R_c are the precision and recall of class c, respectively. The malignant class was designated as the positive class. Macro-F1 is the mean F1 score across both classes, giving each class equal weight. In addition to predictive metrics, the efficiency comparison considered the number of parameters, FLOPs, model size, training time, inference time per image, and peak GPU memory usage.

I. Experimental Environment and Reproducibility

The experiments were conducted in a Kaggle Notebook environment providing two NVIDIA Tesla T4 GPUs. However, the training implementation did not employ multi-GPU mechanisms such as *DataParallel* or *Distributed Data Parallel*; consequently, each training process used a single T4 GPU. The software environment comprised Python 3.12.13, PyTorch 2.10.0+cu128 with CUDA 12.8 (*available*: True), cuDNN 91002, timm 1.0.26, Albumentations 2.0.8, OpenCV 4.13.0, and scikit-learn 1.6.1. To improve experimental consistency, the Python, NumPy, and PyTorch random seeds were set to 42. The best weights, test-set probabilities, learning curves, confusion matrices, classification reports, and metric summaries for each fold were stored separately.

IV. RESULT AND DISCUSSION

A. Overall Performance Comparison

The three models were evaluated on 298 independent images comprising 170 benign and 128 malignant cases. EfficientNetV2-S achieved the highest accuracy, macro-F1, and specificity, while DenseNet121 achieved the highest sensitivity and ROC-AUC (Table V). DenseNet121 produced 39 false negatives, fewer than the other models, but also produced more false positives. Overall performance remained moderate. Possible contributing factors include ROI-only 224×224 grayscale inputs, variation in lesion appearance, leakage-aware patient-level partitioning, and a shared training configuration that was not tuned separately for each architecture. These results describe experimental trade-offs and should not be interpreted as clinical validation.

TABLE V
PERFORMANCE COMPARISON ON THE INDEPENDENT TEST SET

Metric	DenseNet121	EfficientNetV2-S	MobileViT-S
Accuracy	0.6611	0.6913	0.6510
Macro-F1	0.6597	0.6837	0.6487
Sensitivity	0.6953	0.6250	0.6641
Specificity	0.6353	0.7412	0.6412
ROC-AUC	0.7446	0.7030	0.7368
E2E latency (ms/image)	16.3840	15.0013	11.8658

Fold-level macro-F1 varied modestly across patient-level splits: DenseNet121 0.6376 ± 0.0224, EfficientNetV2-S 0.6113 ± 0.0182, and MobileViT-S 0.6403 ± 0.0163. These standard deviations describe sensitivity to fold composition rather than statistical significance or

external generalizability. CV-bagging improved macro-F1 relative to the mean of the individual fold models, most notably for EfficientNetV2-S (+0.0724).

TABLE VI
SUMMARY OF FOLD-LEVEL MACRO-F1 SCORES ON THE TEST SET

Model	Fold 0	Fold 1	Fold 2	Fold 3	Fold 4	Mean ± SD
DenseNet121	0.6423	0.6539	0.6236	0.6068	0.6614	0.6376 ± 0.0224
EfficientNetV2-S	0.6340	0.6257	0.5895	0.6006	0.6068	0.6113 ± 0.0182
MobileViT-S	0.6318	0.6430	0.6168	0.6540	0.6558	0.6403 ± 0.0163

Detailed results for each DenseNet121 fold-specific model on the same test set are presented in Table VII. The most pronounced interfold variation occurred in sensitivity, which ranged from 0.5938 to 0.7188, and specificity, which ranged from 0.5824 to 0.7059. Probability aggregation increased the ROC-AUC from the highest individual-model value of 0.7235 to 0.7446 and produced a macro-F1 score of 0.6597.

TABLE VII
FOLD-LEVEL AND CV-BAGGING TEST RESULTS FOR DENSENET121

Fold	Overall performance			Classification		
	Acc.	F1	AUC	Sens.	Spec.	Prec.
Fold 0	0.6443	0.6423	0.7007	0.6641	0.6294	0.5743
Fold 1	0.6544	0.6539	0.7090	0.7188	0.6059	0.57768
Fold 2	0.6242	0.6236	0.7077	0.6797	0.5824	0.5506
Fold 3	0.6107	0.6068	0.6864	0.5938	0.6235	0.5429
Fold 4	0.6678	0.6614	0.7235	0.6172	0.7059	0.6124
CV-bagging	0.6611	0.6597	0.7446	0.6953	0.6353	0.5894

B. Training-Process Analysis

The validation macro-F1 curves in Fig. 3 illustrate the convergence behavior of the models. Each panel shows the fold whose best validation macro-F1 score was closest to the five-fold mean; the complete curves for all folds were retained as experimental artifacts.

A numerical summary of the training process is reported in Table VIII. For DenseNet121, the cumulative training time across the five folds was 32.3 min, while the total process time, including overhead, was 32.7 min. Its mean best validation macro-F1 score was 0.6545 ± 0.0219. The best checkpoints for folds 0–4 were obtained at epochs 10, 13, 8, 7, and 11, respectively; training stopped at epochs 15, 18, 13, 12, and 16 after no improvement was observed for five epochs. EfficientNetV2-S required the longest training time at 40.1 min and achieved a mean score of 0.6212 ± 0.0365. The EfficientNetV2-S archive did not retain the epoch indices of the best checkpoints. MobileViT-S required 31.1 min and achieved a mean score of 0.6541 ± 0.0250; its best checkpoints for folds 0–4 were obtained at epochs 15, 6, 7, 18, and 20, respectively.

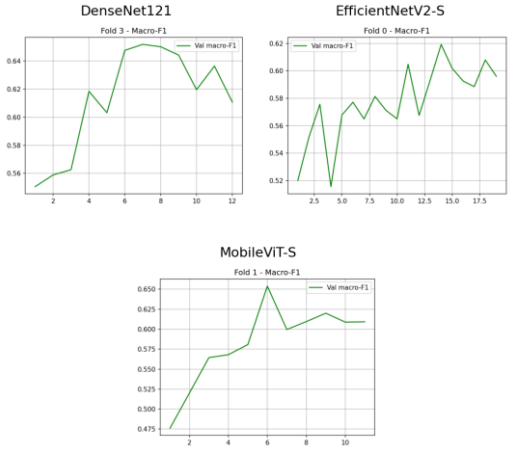

Figure. 3 Validation macro-F1 curves for representative folds of DenseNet121, EfficientNetV2-S, and MobileViT-S

TABLE VIII
SUMMARY OF THE TRAINING PROCESS

Model	Total time	Best val. F1
DenseNet121	32.7 min	0.6545 ± 0.0219
EfficientNetV2-S	40.1 min	0.6212 ± 0.0365
MobileViT-S	31.1 min	0.6541 ± 0.0250

The DenseNet121 training/validation set sizes for folds 0–4 were 2,093/495, 2,063/525, 2,052/536, 2,082/506, and 2,062/526 images, respectively. The corresponding best validation macro-F1 scores were 0.6186, 0.6754, 0.6676, 0.6518, and 0.6589, with training times of 6.5, 7.8, 5.7, 5.2, and 7.1 min.

C. Classification-Error Analysis

The CV-bagging confusion matrices of the three models are shown in Fig. 4. Each cell reports the number of cases in the independent test set using the green visualization generated during the experiments. DenseNet121 produced the fewest false negatives, EfficientNetV2-S produced the truest negatives and the fewest false positives, and MobileViT-S exhibited an error pattern between those of the two CNNs.

The class-wise metrics in Table IX illustrate the prediction tendencies of each model. DenseNet121 achieved the highest malignant recall at 0.6953, whereas EfficientNetV2-S achieved the highest malignant precision at 0.6452 and the highest benign recall at 0.7412. MobileViT-S achieved a malignant recall of 0.6641 with a precision of 0.5822.

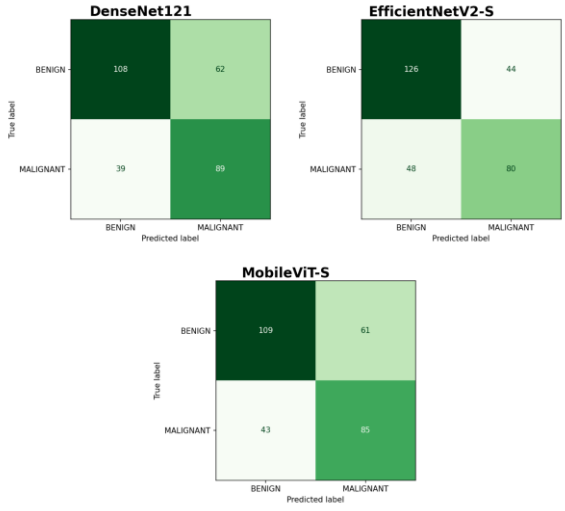

Figure. 4 CV-bagging confusion matrices on the independent test set for DenseNet121, EfficientNetV2-S, and MobileViT-S

TABLE IX
CLASS-WISE CLASSIFICATION PERFORMANCE

Model	Class	Precision	Recall	F1	Support
DenseNet121	Benign	0.7347	0.6353	0.6814	170
	Malignant	0.5894	0.6953	0.6380	128
EfficientNetV2-S	Benign	0.7241	0.7412	0.7326	170
	Malignant	0.6452	0.6250	0.6349	128
MobileViT-S	Benign	0.7171	0.6412	0.6770	170
	Malignant	0.5822	0.6641	0.6204	128

D. Efficiency Analysis and Model Selection

The computational profiles were measured on CUDA using FP32 precision, a batch size of 32, 298 test images, and five fold-specific models. FLOPs were calculated with torch.utils.flop_counter.FlopCounterMode; single-fold times in Table XI are estimates, whereas CV-bagging times were measured directly. MobileViT-S had the lowest parameter count, FLOPs, memory requirement, and inference time, while EfficientNetV2-S had the largest parameter count but the highest accuracy and macro-F1. DenseNet121 retained the highest sensitivity and ROC-AUC. CV-bagging evaluates five models per image and therefore increases compute and memory compared with single-model inference; this can improve robustness but is less attractive for constrained deployment.

TABLE X
COMPUTATIONAL COMPLEXITY OF THE THREE MODELS

Measurement	DenseNet121	EfficientNetV2-S	MobileViT-S
Parameters / model	6,949,634	20,179,618	4,938,626
FP32 memory / model	26.51 MB	76.98 MB	18.84 MB
GFLOPs / model	5.5089	5.6817	2.8250
Ensemble parameters	34,748,170	100,898,090	24,693,130
Ensemble FP32 memory	132.55 MB	384.90 MB	94.20 MB
Ensemble GFLOPs	27.5446	28.4084	14.1252

TABLE XI
INFERENCE COMPARISON ON 298 TEST IMAGES

Measurement	DenseNet121	EfficientNetV2-S	MobileViT-S
Est. single fold (ms/image)	3.1089	2.8682	2.2544
CV computation (s)	4.632235	4.273588	3.359071
CV latency (ms/image)	15.5444	14.3409	11.2721
CV (images/s)	64.33	69.73	88.72

The primary CV-bagging performance metrics are compared in Fig. 5.

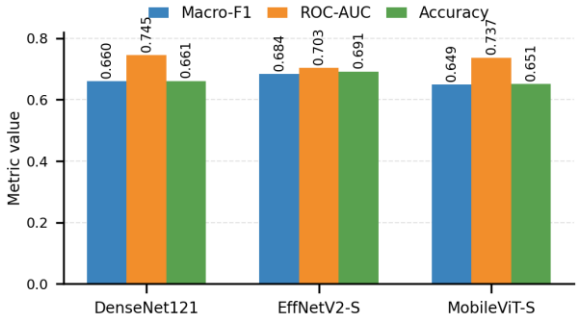

Figure. 5 Comparison of macro-F1, ROC-AUC, and accuracy obtained through CV-bagging for the three models

V. CONCLUSION

This study presents a controlled comparison of DenseNet121, EfficientNetV2-S, and MobileViT-S for benign–malignant mammogram classification on CBIS-DDSM using a common preprocessing pipeline, patient-level five-fold cross-validation, and CV-bagging. EfficientNetV2-S achieved the best overall classification balance with the highest accuracy and macro-F1 score, DenseNet121 achieved the highest sensitivity and ROC-AUC, and MobileViT-S provided the lowest computational cost. These results indicate that no single

model is optimal across all criteria, and model selection should consider both predictive performance and computational efficiency.

The findings remain limited to ROI-based CBIS-DDSM data and should be interpreted as experimental evidence rather than clinical validation. The shared training configuration may also not be optimal for every architecture. Future work should therefore include architecture-specific tuning and external evaluation on independent and multi-institutional datasets. Previous studies have shown the importance of cross-institutional validation for assessing the generalizability of mammography models [23], while multicenter validation platforms provide a framework for evaluating AI systems across different clinical settings [24]. Multi-institutional validation has also been shown to be important for assessing robustness across populations and institutions [25]. Such evaluation would provide stronger evidence regarding the generalizability and potential clinical applicability of the compared models.

REFERENCES

[1] F. Bray *et al.*, "Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA Cancer J. Clin.*, vol. 74, no. 3, pp. 229–263, 2024, doi: 10.3322/caac.21834.

[2] M. Makram, A. S. Aziz, M. M. Saeid, and M. T. Mohamed, "Breast cancer detection and classification via a robust deep learning approach," *Sci. Rep.*, vol. 16, Art. no. 22511, 2026, doi: 10.1038/s41598-026-62255-2.

[3] A. D. Lauritzen *et al.*, "An artificial intelligence-based mammography screening protocol for breast cancer: Outcome and radiologist workload," *Radiology*, vol. 304, no. 1, pp. 41–49, 2022, doi: 10.1148/radiol.210948.

[4] K. Lang *et al.*, "Artificial intelligence-supported screen reading versus standard double reading in the Mammography Screening with Artificial Intelligence trial (MASAI): a clinical safety analysis of a randomised, controlled, non-inferiority, single-blinded, screening acc," *Lancet Oncol.*, vol. 24, no. 8, pp. 936–944, 2023, doi: 10.1016/S1470-2045(23)00298-X.

[5] K. Dembrower, A. Crippa, E. Colon, M. Eklund, and F. Strand, "Artificial intelligence for breast cancer detection in screening mammography in Sweden: a prospective, population-based, paired-reader, non-inferiority study," *Lancet Digit. Health*, vol. 5, no. 10, pp. e703–e711, 2023, doi: 10.1016/S2589-7500(23)00153-X.

[6] M. Cantone, C. Marrocco, F. Tortorella, and A. Bria, "Convolutional networks and transformers for mammography classification: An experimental study," *Sensors*, vol. 23, no. 3, p. 1229, 2023, doi: 10.3390/s23031229.

[7] M. A. K. Raiaan, N. M. Fahad, M. S. H. Mukta, and S. Shatabda, "Mammo-Light: A lightweight convolutional neural network for diagnosing breast cancer from mammography images," *Biomed. Signal Process. Control*, vol. 94, p. 106279, 2024, doi: 10.1016/j.bspc.2024.106279.

[8] W. Lotter *et al.*, "Robust breast cancer detection in mammography and digital breast tomosynthesis using an annotation-efficient deep learning approach," *Nat. Med.*, vol. 27, no. 2, pp. 244–249, 2021, doi: 10.1038/s41591-020-01174-9.

[9] A. Yala *et al.*, "Toward robust mammography-based models for breast cancer risk," *Sci. Transl. Med.*, vol. 13, no. 578, p. eaba4373, 2021, doi: 10.1126/scitranslmed.aba4373.

[10] M. Tan and Q. V. Le, "EfficientNetV2: Smaller models and faster training," in *Proc. 38th Int. Conf. Mach. Learn. (ICML)*, vol. 139, pp. 10096–10106, 2021.

[11] X. Chen *et al.*, "Transformers improve breast cancer diagnosis from unregistered multi-view mammograms," *Diagnostics*, vol. 12, no. 7, p. 1549, 2022, doi: 10.3390/diagnostics12071549.

[12] S. Sarker, P. Sarker, G. Bebis, and A. Tavakkoli, "MV-Swin-T: Mammogram classification with multi-view Swin Transformer," in *Proc. 2024 IEEE Int. Symp. Biomed. Imaging (ISBI)*, pp. 1–5, 2024, doi: 10.1109/ISBI56570.2024.10635578.

[13] F. Manigrasso, R. Milazzo, A. R. Sicilia, L. Morra, and F. Lamberti, "Mammography classification with multi-view deep learning techniques: Investigating graph and transformer-based architectures," *Med. Image Anal.*, vol. 99, p. 103320, 2025, doi: 10.1016/j.media.2024.103320.

[14] S. Mehta and M. Rastegari, "MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2022.

[15] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," *Patterns*, vol. 4, no. 9, p. 100804, 2023, doi: 10.1016/j.patter.2023.100804.

[16] M. Ameen, "Explainable mammogram analysis with EfficientNetV2 and Grad-CAM++ for robust cancer diagnosis," *Diagnostics*, vol. 16, no. 1, p. 105, 2026, doi: 10.3390/diagnostics16010105.

[17] M. A. Jones, R. Faiz, Y. Qiu, and B. Zheng, "Improving mammography lesion classification by optimal fusion of handcrafted and deep transfer learning features," Phys. Med. Biol., vol. 67, no. 5, p. 054001, 2022, doi: 10.1088/1361-6560/ac5297.

[18] N. Mao et al., "Attention-based deep learning for breast lesions classification on contrast enhanced spectral mammography: a multicentre study," Br. J. Cancer, vol. 128, no. 5, pp. 793–804, 2023, doi: 10.1038/s41416-022-02092-y.

[19] D. G. P. Petrini, C. Shimizu, R. A. Roela, G. V. Valente, M. A. A. K. Folgueira, and H. Y. Kim, "Breast cancer diagnosis in two-view mammography using end-to-end trained EfficientNet-based convolutional network," IEEE Access, vol. 10, pp. 77723–77731, 2022, doi: 10.1109/ACCESS.2022.3193250.

[20] G. Ayana et al., "Vision-Transformer-based transfer learning for mammogram classification," Diagnostics, vol. 13, no. 2, p. 178, 2023, doi: 10.3390/diagnostics13020178.

[21] W. Hsu et al., "External validation of an ensemble model for automated mammography interpretation by artificial intelligence," JAMA Netw. Open, vol. 5, no. 11, p. e2242343, 2022, doi: 10.1001/jamanetworkopen.2022.42343.

[22] A. S. Tejani et al., "Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 update," Radiol. Artif. Intell., vol. 6, no. 4, p. e240300, 2024, doi: 10.1148/ryai.240300.

[23] D. Ueda et al., "Development and validation of a deep learning model for detection of breast cancers in mammography from multi-institutional datasets," PLoS One, vol. 17, no. 3, p. e0265751, 2022, doi: 10.1371/journal.pone.0265751.

[24] F. Cossío et al., "VAI-B: a multicenter platform for the external validation of artificial intelligence algorithms in breast imaging," J. Med. Imaging, vol. 10, no. 6, p. 061404, 2023, doi: 10.1117/1.JMI.10.6.061404.

[25] A. Yala et al., "Multi-institutional validation of a mammography-based breast cancer risk model," J. Clin. Oncol., vol. 40, no. 16, pp. 1732–1740, 2022, doi: 10.1200/JCO.21.01337.