System based on convolutional neural networks for breast cancer detection

Sánchez Fuentes Luis Fernando - 2142610, Monsalve Durán Santiago Enrique - 2200520, Pertuz Said

Connectivity and Signal Processing Research Group (CPS)
Universidad Industrial de Santander

Abstract

This research analyzes how domain shift affects CNN-based mammography classifiers and proposes mitigation strategies. A baseline was established with pre-trained End2End architectures from public repositories. Among the architectures, the best-performing model was selected and adapted through transfer learning and fine-tuning to a private dataset. Subsequently, a radiomic approach was taken at the patient level, integrating images with four views, manual descriptors and a gradient boosting classifier were constructed. An initial evaluation of the models was performed using the private dataset, and AUC, sensitivity, and specificity indicators were reported under an optimized threshold with G-mean. When evaluated out-of-domain without adaptation, substantial drops in model performance were identified. Two model configurations were trained using transfer learning followed by fine-tuning. The results of both configurations substantially improved performance over the baseline. Configuration 1 improved AUC and specificity; configuration 2 improved AUC and sensitivity. Radiomic analysis simultaneously improved sensitivity and specificity compared to the baseline with a slight decrease in AUC. In conclusion, adapting the model to domain shift is effective in reducing metric losses, while radiomics is an alternative in cases with specific constraints and conditions.

Keywords— convolutional neural networks, mammography, radiomics, transfer learning, fine-tuning, sensitivity, specificity, threshold.

I. Introduction

Breast cancer is the most common cancer in women and one of the leading causes of cancer morbidity. According to the Global Cancer Observatory, in 2022, more than 2,296,840 new cases and 666,103 deaths from breast cancer were diagnosed worldwide, with a high incidence in low- and middle-income regions [1]. In Colombia, it is the most common cancer in both sexes, with 17,018 new cases and 4,752 deaths in 2022 [2]. Due to the relevance of this pathology, different methods for early detection are becoming increasingly important in the global context, where mammography continues to be the most widely used standardized modality for screening, especially for women over 40 or those with a genetic predisposition [3]. However, real-world screening still faces persistent interpretation limits and variability, yielding high false-positive rates (56.3% Digital Mammography; 49.6% Digital Breast Tomosynthesis over 10 years [4]) and significant inter-reader variability that affects quality and costs [5].

With the aim of improving detection rates and successful screenings, Computer-Aided Detection/Diagnosis (CAD) systems have been established, but they have not demonstrated consistent improvements in diagnostic accuracy compared to readings by specialists [4]. Recently, technology has evolved towards Deep Learning with convolutional neural networks (CNN), which, when trained with large data sets, have shown an improvement in predictive and detection diagnosis in mammograms [5]. In addition, the scientific community has consolidated public repositories that contribute to research in segmentation, detection, and classification, that also serve for models and algorithm comparison [6], [7]. These technologies have evolved into state-of-the-art (SOTA) models, which represent the most advanced and best-performing approaches nowadays. However, these advances in CNNs face important limitations in performance due to domain shifts. Domain shift is when the data distribution at deployment differs from the training distribution due to changes in datasets, which causes performance indicators to deteriorate due to the difference in the adjustment of the new data to the model [8].

In this work, we aim to address domain shifts evaluating the performance of SOTA models for cancer detection in public datasets and subsequently adapting the strongest baseline to a target domain. We first quantify the performance gap of the pre-trained models on our target domain to establish a baseline. We then perform a preprocessing step to better align data distributions and apply transfer learning/fine-tuning to adapt the best model. For comparison purposes, we construct a patient-level radiomic representation that integrates four-view predictions with complementary image descriptors to assess its added value. All approaches are compared against the baseline on an independent test set using AUC, accuracy, sensitivity, and specificity with 95% confidence intervals. Our hypothesis is that domain-aware adaptation can compensate for performance lost due domain shifts between the source datasets and the target domain.

II. Problem statement

Although mammography is the standard screening test for breast cancer detection, the interpretation of results continues to present a high percentage of false positives and false

negatives, which ultimately reduces the final benefit of the test and increases medical costs for patients. Although various technologies have been implemented to improve diagnostic accuracy, significant challenges remain in terms of the ability to generalize models to unseen domains and the heterogeneity in transfer-learning sources and strategies. Current SOTA models perform well using the original dataset, but when using a model trained and validated in another context, there is a degradation in AUC, accuracy, and sensitivity due to the different distribution of data, miscalibration, among other factors [11]. However, it is possible to adapt these models to a small dataset through domain-aware adjustments and careful validation, while quantifying uncertainty. Therefore, it is necessary to test whether a model trained on another dataset can be adapted to the target domain with minimal changes to improve performance without increasing false positive detection.

III. Previous related work

The use of CNNs has gradually replaced CAD models based on hand-crafted features in the detection of breast cancer through mammogram analysis. Large-scale research demonstrated that a CNN trained with thousands of mammograms outperformed a reference CAD system in terms of sensitivity and specificity indicators, concluding the importance of deep learning-based models for mammogram analysis [12]. Subsequent research showed that adopting Faster R-CNN for simultaneous detection and classification of masses and microcalcifications achieves SOTA performance in INbreast and competitiveness in the Digital Mammography DREAM Challenge, reinforcing the viability of object detection frameworks for this task [13]. The importance of using CNNs can be seen by adopting performance measurement indicators, as well as the existence of various techniques that can be used for reading and interpreting mammograms with outstanding results.

From a multi-view and high-resolution approach, a study proposed a multi-view model (CC/MLO and both breasts) that can process high-resolution images while combining contralateral and ipsilateral signals. This work showed significant improvement by preserving fine details and integrating the four views [14]. Another study proposed an end-to-end scheme that contains annotations only in the initial phase and then continues with image-level labels, which resulted in high detection performance while reducing dependence on exhaustive annotations [15]. This line of research led to GMIC, a global local and multiple instance classifier that blends global saliency with local patches for high-resolution mammograms, resulting in interpretable and robust attention maps [16]. Previous research justifies the patient unit composed of the four views, the use of high resolution, and end-to-end strategies, as well as helping to provide a basis for interpretability in error analysis and threshold adjustments.

In relation to generalization and adaptation between domains, studies have been conducted on the use of transfer learning to strengthen detectors, in which domain generalization for mass detection in multivendor/multicenter environments was systematically analyzed, resulting in the generation of single-source training strategies with the aim of mitigating domain shift [17]. Meanwhile, in full-field digital mammography (FFDM), transfer adaptation has been studied, such as fine-tuning of Faster-CNN trained in OPTIMAM Mammography Image Database (OMI-DB), for the detection of cases in databases with a smaller dataset, which has been shown to perform competitively among manufacturers [18]. This provides evidence of inter-domain dropout, which in turn supports the transfer learning and fine-tuning strategy to recover performance with a dataset with few data.

As support for clinical validation and evaluation in real trials, studies have been conducted in which an AI system was developed and showed a reduction in false positives and false negatives compared to external readers, while also achieved competitive performance in double-reading protocols. This model was established by retrospectively evaluating clinical cohorts however, it serves as a step toward multicenter and prospective validations [19]. In addition, a major collective effort was the Digital Mammography DREAM Challenge, in which multiple AI algorithms were evaluated in scenarios simulating clinical practice. The results showed that no model was consistently better than radiologists in all settings; however, the AI ensemble and single reader significantly improved specificity at fixed sensitivity, which in practical terms means that AI could be used as a "second digital reader" [20]. On the other hand, the Mammography Screening with Artificial Intelligence trial (MASAI) conducted in Sweden compared AI-assisted reading with standard double reading in population screening with the aim of finding prospective evidence and evaluating workflow. The results of the pre-specified safety analysis showed no lower detection rates with reduced reading burden, demonstrating that there is some safety in incorporating AI into the screening circuit [21]. Previous research helps to guide clinical criteria and provides prospective evidence to argue for safety and reading workload.

Finally, efficient annotation and generalization to Digital Breast Tomosynthesis (DBT) have been addressed based on an approach that combines image-level pre-training with the flagging of suspicious regions with minimal annotation. The system designed was able to extend from 2D mammography to DBT, as well as detect cancers in previously negative exams and increase sensitivity compared to specialists without an increase in false positives [22]. In relation to DBT, a large-scale data research project called JAMA Network Open was conducted in 2023 with 22,032 DBT volumes, codes, and an evaluation platform, which resulted in biopsied tumor sensitivities of around 0.96 in the most effective methods and, in turn, released tools to contribute to reproducible research [23]. In addition, a study investigated the ability of an algorithm to detect in 2D those cases only detected in DBT, demonstrating the importance and potential of AI to detect cancers in situations where they were not identified by digital mammography [24]. As can be seen from these studies, a future line of research into DBT can be opened up, demonstrating large-scale results, while providing evidence that a good 2D model can detect cases with

good performance.

IV. Materials and methods

A. Imaging and patient data

For reproducibility, we used two public datasets: The CBIS-DDSM and INbreast. The CBIS-DDSM (Curated Breast Imaging Subset of DDSM) is a public dataset of mammograms derived from the original DDSM database, which contains thousands of digitized mammography images from both breast cancer patients and those with negative results, with annotations verified by pathology [25]. Due to its large amount of data, it is a commonly used dataset for training and evaluating breast cancer detection algorithms. In contrast, the INbreast is a public database of full-field digital mammograms, containing 410 high-quality images from 115 patient studies with benign and malignant findings identified and detailed by radiologists [9]. In order to study the impact of domain shift, we used a private dataset collected at Tampere University Hospital (Tampere, Finland). This dataset was collected as part of the project "Software for Parenchymal Analysis of Mammographic Images for Breast Cancer Risk Estimation" (code 110284467139 – Minciencias). The collection and use of imaging and clinical data was reviewed and approved by the Institutional Review Board (IRB) of Tampere University Hospital. This dataset contains data from a retrospective age-matched case–control cohort with 382 patients (191 cancer cases and 191 controls). Each exam includes craniocaudal (CC) and mediolateral-oblique (MLO) views of the left and right breast (four views in total per patient, see Fig. 1), for a total of 1,528 images. The dataset is balanced at the patient level (50%/50% cases–controls) but imbalanced at the image level because not every view of a cancer case shows a malignant finding (382 positive vs 1,146 negative images). This private dataset was used in the final stage as the target domain for training, validation, and testing. Imaging data were randomly split into train, validation and test sets stratified by patient age. The split was performed at the patient level, preserving matched case–control pairs: both members of each pair (and all images from each patient) were assigned to the same subset. Of the 1528 mammograms, 70% (1064) were used for training, 10% (152) were used for validation and 20% (312) were used for testing.

B. Image Processing

Standard preprocessing was performed on all mammograms to standardize the format and highlight key information. Each mammogram image, which was originally in grayscale, was resized to a fixed size of 1152 x 896 pixels. Subsequently, the single gray channel image was replicated into three identical channels, as the models require 3-channel (RGB) images for processing. In addition, pixel intensities were normalized by scaling the intensity values to [0-255] (from 16 to 8 bits), and the pixel mean for each image, calculated over the public dataset on the training split, was subtracted. This last step allows the data to be centered and coincides with the methodology used in the pre-trained CNN-based models. The preprocessing scripts used in this work can be found in [15].

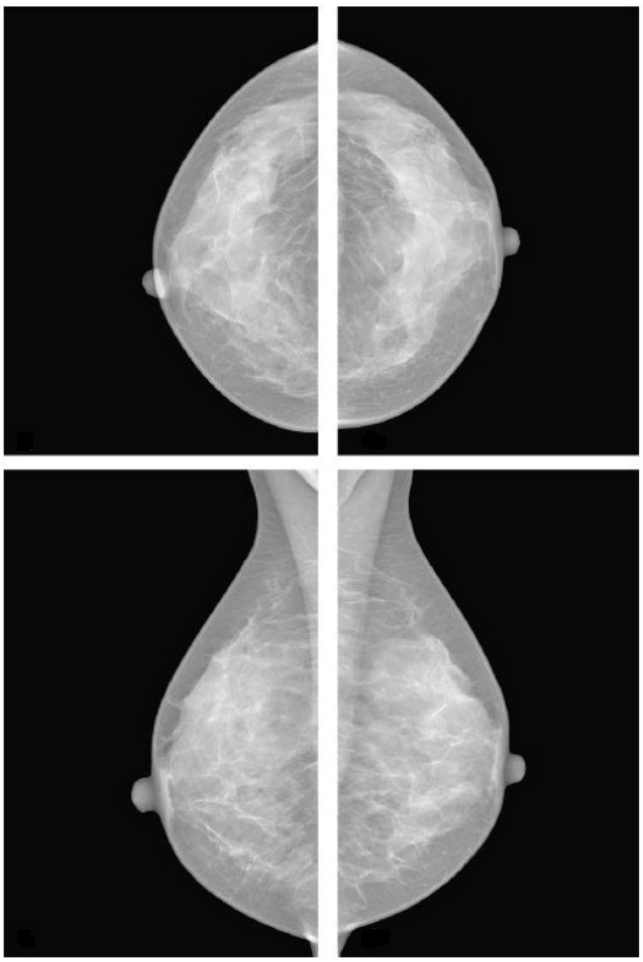

Fig. 1. Mammogram examples (top-left to bottom-right): right CC, left CC, right MLO, left MLO.

C. Baseline CNN-based detection models

We selected the End2End architecture proposed by Li Shen et al. [15]. This architecture was previously compared against three other SOTA systems for cancer detection in our research group and yielded the highest performance [26]. In addition, the End2End architecture provides open-source, reproducible implementations facilitating subsequent domain-adaptation tasks. We evaluated the End2End CNN framework to identify the architecture best suited to our requirements. Specifically, we adopt a VGG16-based model, using the widely adopted 16-layer backbone (which processes the mammogram to produce spatial feature maps) in computer vision, noted for its sequential convolutional blocks and strong performance on generic image classification tasks [27]. This backbone expects three-channel (RGB) input. On top of the backbone, the prediction head comprises two stacked Conv2D–Batch Normalization–Activation–Dropout–MaxPooling2D blocks, followed by GlobalAveragePooling2D and a final fully connected dense layer that refines these features to generate image-level predictions (see Fig. 2). Because the End2End framework offers multiple implementations, in this work we retain only the

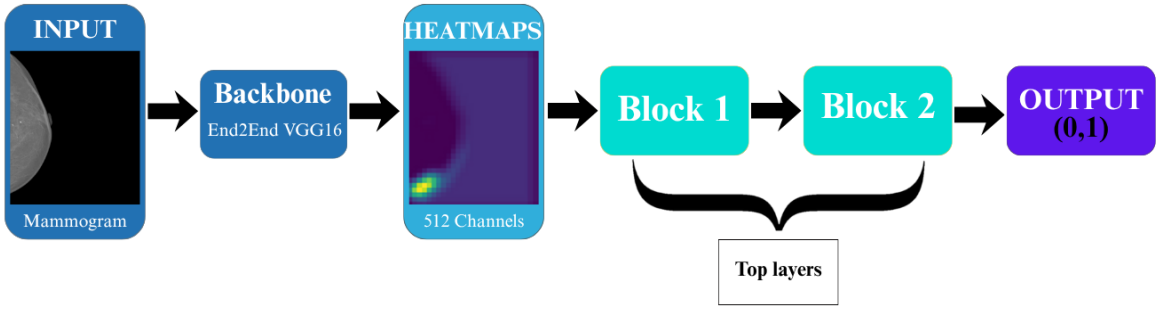

Fig. 2. Block diagram of VGG16 End2End architecture

variant that uses VGG16 as the backbone for the subsequent experiments. All models used were pre-trained on public datasets (CBIS-DDSM and INbreast).

D. Transfer learning and fine-tuning

For domain adaptation, we used transfer learning followed by a two-step fine-tuning process using only the private dataset. The pre-trained End2End model was used as a starting point to retrain its final layers with the private training images. In the first stage, only the upper layers were unfrozen and retrained, while the deeper convolutional backbone remained frozen. In the second stage, all layers of the model were unfrozen for full fine-tuning, allowing a more comprehensive adaptation of the network to the specific characteristics of the medical data. In this way, low-level features such as edges and textures learned during pre-training were preserved at the beginning, before progressively adapting the entire network. During training, callbacks were implemented to improve robustness: early stopping to prevent overfitting by monitoring the validation loss, reduction of the learning rate when the validation performance plateaued, and checkpoint saving of the best weights achieved. In addition, careful data augmentation was applied, consisting of small and smooth transformations to increase the variability of training images without altering their anatomical meaning. These transformations included slight rotations, horizontal shifts, vertical shifts, zoom adjustments to simulate closer or farther views, and horizontal flips. Care was taken to avoid excessive distortions that could negatively affect orientation and recognition. The fine-tuning process was carried out by monitoring the performance indicators in the validation set at each epoch.

E. Radiomic analysis

In addition to transfer learning and fine-tuning, we explored the use of radiomics descriptors to improve the robustness to domain shift. Specifically, an analysis of traditional radiomic features was performed on the mammograms, with the aim of increasing detection with additional descriptors. For each patient in the private dataset, a feature vector was constructed containing four malignancy scores predicted by the End2End base model from the four images per patient, plus 34 radiomic image features for each mammogram calculated using the OpenBreast repository tools [28]. These radiomic features include intensity distributions, texture measures, morphological features of the breast and glandular tissue, among other descriptors used in cancer detection [28]. Each vector consisted of 141 attributes including the patient's age.

In this approach, grid search was applied using GridSearchCV to explore parameter ranges, shown in Table I, for traditional machine learning model (Gradient Boosting) and identify the best hyperparameters. After tuning, a Principal Component Analysis (PCA) was used to reduce the dimensionality of the feature space, concentrating most of the variance in a smaller set of components that retained the essential information.

TABLE I
RANGES EXPLORED IN GRIDSEARCH FOR MACHINE LEARNING MODEL (GRADIENT BOOSTING)

Parameter	Range (Min-Max)
N_components (PCA)	10-30
Learning_rate	0.050-0.200
Max_depth	2-4
Min_samples_leaf	1-20
N_estimators	200-10000
Subsample	0.600-1.000

V. Performance Measures and Statistical Analysis

A decision threshold converts the probabilities predicted by the models into class labels (malignant vs non-malignant). Higher thresholds increase specificity and typically lower sensitivity. There are techniques for optimally calculating this value, and those that maximize sensitivity are of particular interest for this project. The technique chosen was G-mean, which is defined as the root of sensitivity times specificity; the threshold that maximizes G-mean on the validation split is then fixed and used for reporting on the independent test set. Given image-level class imbalance, with 1,528 images, of which 1,146 are non-malignant (75%) and 382 are malignant (25%), it is not possible to select the threshold arbitrarily, and G-mean is used to calculate the optimal threshold that maximizes sensitivity.

We report the Area Under the ROC Curve (AUC), sensitivity and specificity. To quantify uncertainty, 95% confidence intervals for AUC on the test set were obtained via a nonparametric bootstrap (resampling with replacement).

TABLE II
BASELINE INFERENCE ON INBREAST (IN-DOMAIN)

Architecture	Pre-Trained with	AUC (95% CI)	Sensitivity	Specificity	Threshold
VGG16	CBIS-DDSM	0.721[0.650-0.787]	0.510	0.951	0.520
VGG16	CBIS-DDSM+INbreast	0.967[0.946-0.986]	0.870	0.974	0.282

TABLE III
BASELINE INFERENCE ON CBIS-DDSM (IN-DOMAIN)

Architecture	Pre-Trained with	AUC (95% CI)	Sensitivity	Specificity	Threshold
VGG16	CBIS-DDSM	0.881[0.869-0.894]	0.791	0.802	0.332
VGG16	CBIS-DDSM+INbreast	0.789[0.772-0.806]	0.749	0.651	0.247

TABLE IV
BASELINE INFERENCE ON PRIVATE DATASET (OUT-OF-DOMAIN)

Architecture	Pre-Trained with	AUC (95% CI)	Sensitivity	Specificity	Threshold
VGG16	CBIS-DDSM	0.732[0.653-0.810]	0.641	0.794	0.306
VGG16	CBIS-DDSM+INbreast	0.755[0.686-0.827]	0.666	0.790	0.256

VI. Experiments and results

A. Computing Environment

A development environment was set up to run deep learning models using a workstation with the Ubuntu Linux operating system, connected to the Connectivity and Signal Processing (CPS) Research Group's computing cluster to take advantage of GPU processing. The main hardware included eight 16GB NVIDIA Tesla V100 GPUs working in parallel, with CUDA v11.4 drivers and libraries installed to accelerate intensive mathematical operations. The code was developed in Python (3.7.12), using the TensorFlow (2.8.1) and Keras frameworks (2.8.0) to ensure the import and execution of pre-trained models from the End2End repository [15].

B. Initial Inference with Pre-trained Models.

We used the baseline pre-trained models to make inference on public and private datasets. The environment was adapted to load the models and perform initial inference tests on the datasets. The mammograms from datasets were processed individually through each model, obtaining a prediction or estimated probability of malignancy for each image.

Tables II and III report the in-domain performance of the baseline models (evaluation on the same public datasets used for training). In contrast, Table IV summarizes the out-of-domain results on our private dataset, evidencing the expected performance drop due to domain shift. Within the VGG16 family we evaluated two variants: (i) pre-trained on CBIS–DDSM only and (ii) pre-trained on CBIS–DDSM and then INbreast. The CBIS–DDSM + INbreast variant generalized better across datasets, so we selected this second variant as the baseline model for the subsequent domain adaptation tasks.

C. Performance of baseline CNN-based model plus radiomic features.

After selecting the base model, we built the patient-level feature vector and ran a grid search to tune a Gradient Boosting classifier; the selected hyperparameters were learning_rate = 0.05, max_depth = 2, min_samples_leaf = 20, n_estimators = 200, subsample = 0.8, random_state = 42, n_components(PCA) = 20. Results can be seen in Table V.

TABLE V
IMPACT OF RADIOMIC FEATURES IN THE INFERENCE ON THE PRIVATE DATASET

Model	AUC (95% CI)	Sensitivity	Specificity	Threshold
VGG16 Baseline + Radiomic Analysis (RA)	0.752 (0.641-0.863)	0.692	0.820	0.552

D. Performance of fine-tuned models.

After selecting the baseline model and the Stage-1/Stage-2 training hyperparameters, shown in Table VI, together with the data-augmentation setup, we proceeded with training. The data augmentation settings can be seen in Table VII.

TABLE VI
HYPERPARAMETERS FOR TRANSFER LEARNING AND FINE-TUNING

Hyperparameter	Value Stage 1	Value Stage 2
Batch Size	16	16
Optimizer	Adam	Adam
Weight Decay	0.0001	0.0015
Dropout	0.2	0.5
Learning Rate	2e-04	2e-07
Learning Rate Patience	2	2
Early Stopping Patience	3	3
Epoch	50	30

TABLE VII
DATA AUGMENTATION SETTINGS

Hyperparameter	Value
Rotation Range	$\pm 10^{\circ}$
Width Shift	$\pm 5\%$
Height Shift	$\pm 5\%$
Zoom Range	$\pm 10\%$
Horizontal Flip	True

We evaluated two fine-tuning configurations shown in Table VIII.

Configuration 1: Unfreeze from Block 2 upward.
In the first setup, during the initial training stage we only let the last part of the network learn: Block 2 and the output layer. Everything before that—Backbone + Block 1—stays fixed. (Later, in the next stage, the whole model is fine-tuned.)

Configuration 2: Unfreeze from Block 1 upward.
In the second setup, in the initial stage we let a larger portion learn: Block 1, Block 2, and the output layer. The Backbone remains fixed at first. (Then, as before, the full model is fine-tuned.)

TABLE VIII
FINE-TUNING TEST RESULTS

Configuration	AUC (95% CI)	Sensitivity	Specificity	Threshold
Config1	0.802 (0.736-0.860)	0.653	0.833	0.284
Config2	0.809 (0.746-0.864)	0.692	0.794	0.253

VII. Discussion

A. Baseline in-domain vs. out-of-domain performance

The comparison between sets showed a deterioration in performance due to domain shift. When comparing the results of the pre-trained End2End model with public datasets in terms of performance when moving outside the domain (see Tables II, III and IV), its AUC performance decreased by 21.92%, as did sensitivity, which fell by 23.45%, and specificity, which decreased by 18.89%. The above results demonstrate the need to adapt the datasets to the target domain to better match the data and improve model performance.

B. Impact of radiomic analysis

As shown in Table IX, the performance of the model trained with public datasets as a baseline is compared with the results of the radiomic analysis and the two model configurations that applied transfer learning and fine-tuning. With regard to radiomic analysis, simultaneous improvements in sensitivity (3.9%) and specificity (3.8%) are observed, which contrasts with a slight decrease of 0.04 in the AUC. This suggests that radiomic analysis improves domain adaptation due to the improvements in specificity and sensitivity.

C. Impact of fine-tuning in out-of-domain performance

Starting from the baseline End2End model, we studied the impact of fine-tuning on domain adaptation using two different configurations. Both configurations yielded an improvement of out-of-domain performance. However, there are marked differences between the performances of each of the model configurations. Config-1 shows a marked improvement in results compared to the baseline of 6.23% in AUC and 5.44% in specificity, but at the expense of sensitivity, which decreases by 1.95%. Config-2, on the other hand, shows superior performance when compared to the baseline, as the AUC increases by 7.15% and sensitivity by 3.9%, but with only a slight improvement of 0.51% in specificity. Both model configurations show outstanding performance results but have different applications. Config-1 is useful in cases where the goal is to reduce false positives due to excessive recalls and unnecessary biopsies. In contrast, Config-2 could be applied in cases where false negatives need to be reduced, and the priority is to not miss patients with cancer who have been misdiagnosed. Training dynamics are consistent with these patterns: in Config-1 (Fig. 3, about 40 epochs), both losses drop rapidly at the beginning and the validation loss levels off near 0.92 after roughly 15 epochs while the training loss keeps decreasing to around 0.83, opening a modest generalization gap. In Config-2 (Fig. 4, about 55 epochs), the validation loss follows a slow downward trend toward 0.89 and stays closer to the training curve, indicating more stable generalization during longer training.

VIII. Conclusions

In this work we studied the impact of domain shift in the performance of CNNs for breast cancer detection from mammography images. For this purpose, we first identified a baseline CNN-based model trained in public datasets. Subsequently, we fine-tuned the model in a new target domain, using a private dataset. Moreover, a patient-level radiomic pipeline was built that integrates four-view CNN scores with handcrafted descriptors, providing an interpretable, low-cost alternative. Evaluation on an independent test set shows that domain-aware fine-tuning improves over the non-adapted baseline, while radiomics offers competitive sensitivity and practical value when computation or labeled data are limited.

Radiomic analysis showed improvements compared to baseline performance by 3.9% in sensitivity and by 3.8% in specificity, with a little decrease in AUC of 0.4%. However, it was outperformed by the utilization of fine-tuning. Nevertheless, radiomic analysis may be an interesting option under strong computational constraints, when local labeled data are limited, or when simple multimodal integration is required and the data are already adjusted for the model.

Therefore, it is evident that a combination of approaches can help manage model performance when a domain shift occurs. CNNs capture complex patterns, while radiomic analysis provides stability and interpretability. In addition, managing domain shift through image normalization, stable feature

TABLE IX
COMPARISON OF PERFORMANCE BEFORE AND AFTER DOMAIN ADAPTATION IN THE PRIVATE DATASET

Configuration	Dataset	AUC	(95% CI)	Sensitivity	Specificity	Threshold
VGG16 Baseline	Private	0.755	(0.686-0.827)	0.666	0.790	0.256
VGG16 Baseline + RA	Private	0.752	(0.641-0.863)	0.692	0.820	0.552
Fine-tuning Config 1	Private	0.802	(0.736-0.860)	0.653	0.833	0.284
Fine-tuning Config2	Private	0.809	(0.746-0.864)	0.692	0.794	0.253

selection, and patient-level validation is just as important as the architecture of the base algorithm.

Fine-tuning the model on the private dataset improved performance over the baseline without adaptation. Training dynamics were stable in both cases (decreasing training loss and early plateau of validation loss), indicating controlled generalization. In practice, when screening priorities emphasize sensitivity, the second configuration may be preferable; when preserving overall discrimination is critical, the first configuration is suitable. Overall, domain-aware fine-tuning recovers performance lost to distribution shift and remains the best test-set performer, while radiomics remains a viable option when compute or labeled data are limited.

REFERENCES

[1] International Agency for Research on Cancer. "Breast Fact Sheet", Global Cancer Observatory, 2024. Accessed: Sep. 3, 2025. [Online]. Available: https://gco.iarc.who.int/media/globocan/factsheets/cancers/20-breast-fact-sheet.pdf

[2] International Agency for Research on Cancer. "Colombia Fact Sheet", Global Cancer Observatory, 2024. Accessed: Sep. 3, 2025. [Online]. Available: https://gco.iarc.who.int/media/globocan/factsheets/populations/170-colombia-fact-sheet.pdf

[3] U.S. Preventive Services Task Force, "Screening for Breast Cancer: U.S. Preventive Services Task Force Recommendation Statement," JAMA, vol. 331, no. 22, pp. 1918–1930, Apr. 2024, doi: 10.1001/jama.2024.5534.

[4] T. H. Ho, et al., "Cumulative probability of false-positive results after 10 years of screening with digital breast tomosynthesis vs digital mammography," JAMA Network Open, vol. 5, no. 3, Mar. 2022, Art. no. e222440, doi: 10.1001/jamanetworkopen.2022.2440.

[5] J. G. Elmore et al., "Variability in Interpretive Performance at Screening Mammography and Radiologists' Characteristics Associated with Accuracy," Radiology, vol. 253, no. 3, pp. 641–651, Dec. 2009, doi: 10.1148/radiol.2533082308.

[6] C. D. Lehman et al., "Diagnostic accuracy of digital screening mammography with and without computer-aided detection," JAMA Internal Medicine, vol. 175, no. 11, pp. 1828–1837, Nov. 2015, doi: 10.1001/jamainternmed.2015.5231.

[7] S. M. McKinney et al., "International evaluation of an AI system for breast cancer screening," Nature, vol. 577, pp. 89–94, Jan. 2020, doi: 10.1038/s41586-019-1799-6.

[8] R. S. Lee et al., "A curated mammography data set for use in computer-aided detection and diagnosis research," Scientific Data, vol. 4, Art. no. 170177, Dec. 2017, doi: 10.1038/sdata.2017.177.

[9] I. Moreira et al., "INbreast: Toward a Full-Field Digital Mammographic Database," Academic Radiology, vol 19, pp. 236-248, Feb. 2012, doi: 10.1016/j.acra.2011.09.014.

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

[11] X. Wang, G. Liang, Y. Zhang, et al., "Inconsistent performance of deep learning models on mammogram classification," J. Am. Coll. Radiol., vol. 17, no. 6, pp. 796–803, 2020.

[12] T. Kooi et al., "Large scale deep learning for computer aided detection of mammographic lesions," Medical Image Analysis, vol. 35, pp. 303–312, Jan. 2017, doi: 10.1016/j.media.2016.07.007.

[13] D. Ribli et al., "Detecting and classifying lesions in mammograms with Deep Learning," Scientific Reports, vol. 8, no. 4165, Mar. 2018, doi: 10.1038/s41598-018-22437-z.

[14] K. J. Geras et al., "High-Resolution Breast Cancer Screening with Multi-View Deep Convolutional Neural Networks," Mar. 2017, doi: 10.48550/arXiv.1703.07047

[15] L. Shen, L. R. Margolies, J. H. Rothstein, E. Fluder, R. McBride, and W. Sieh, "Deep learning to improve breast cancer detection on screening mammography," Scientific Reports, vol. 9, no. 12495, Aug. 2019, doi: 10.1038/s41598-019-48995-4.

[16] Y. Shen et al., "An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization," Medical Image Analysis, vol. 68, no. 101908, Feb. 2021, doi: 10.1016/j.media.2020.101908.

[17] L. Garrucho, K. Kushibar, S. Jouide, O. Diaz, L. Igual, and K. Lekadir, "Domain generalization in deep learning based mass detection in mammography: A large-scale multi-center study," Artificial Intelligence in Medicine, vol. 132, p. 102386, Aug. 2022, doi: 10.1016/j.artmed.2022.102386.

[18] R. Agarwal, O. Díaz, M. H. Yap, X. Lladó, and R. Martí, "Deep learning for mass detection in Full Field Digital Mammograms," Computers in Biology and Medicine, vol. 121, p. 103774, Jun. 2020, doi: 10.1016/j.compbiomed.2020.103774.

[19] S. M. McKinney et al., "International evaluation of an AI system for breast cancer screening," Nature, vol. 577, pp. 89–94, Jan. 2020, doi: 10.1038/s41586-019-1799-6.

[20] T. Schaffter et al., "Evaluation of combined artificial intelligence and radiologist assessment to interpret screening mammograms," JAMA Network Open, vol. 3, no. 3, Mar. 2020, Art. no. e200265, doi: 10.1001/jamanetworkopen.2020.0265.

[21] K. Lång 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 accuracy study," The Lancet Oncology, vol. 24, no. 8, pp. 936–944, Aug. 2023, doi: 10.1016/s1470-2045(23)00298-x.

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

[23] N. Konz et al., "A competition, benchmark, code, and data for using artificial intelligence to detect lesions in digital breast tomosynthesis," JAMA Network Open, vol. 6, no. 2, Art. no. e230524, Feb. 2023, doi: 10.1001/jamanetworkopen.2023.0524.

[24] V. Dahlblom, I. Andersson, K. Lång, A. Tingberg, S. Zackrisson, and M. Dustler, "Artificial intelligence detection of missed cancers at digital mammography that were detected at digital breast tomosynthesis," Radiology Artificial Intelligence, vol. 3, no. 6, Sep. 2021, doi: 10.1148/ryai.2021200299.

[25] R. Sawyer-Lee, F. Gimenez, A. Hoogi, and D. Rubin, "Curated Breast Imaging Subset of Digital Database for Screening Mammography (CBIS-DDSM) [Dataset]," The Cancer Imaging Archive, 2016, doi: 10.7937/K9/TCIA.2016.7O02S9CY.

[26] Pertuz et al., "Saliency of breast lesions in breast cancer detection using artificial intelligence," Scientific Reports, vol. 13, no. 1, Nov. 2023, doi: 10.1038/s41598-023-46921-3.

[27] K. Simonyan and A. Zisserman, "Very deep convolutional networks for Large-Scale image recognition," Sep. 2014, doi: 10.48550/arXiv.1409.1556

[28] S. Pertuz, and H. Shu, "OpenBreast [Dataset]," GitHub, 2021, Accessed: Aug. 25, 2025. [Online]. Available: https://github.com/spertuz/openbreast

ANNEXES

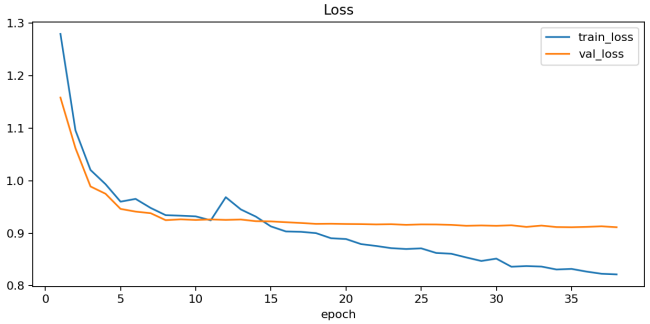

Fig. 3. Loss Curves Config 1.

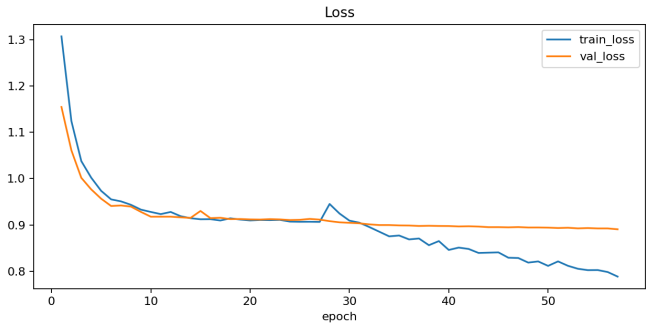

Fig. 4. Loss Curves Config 2.