INTEGRATED DEEP LEARNING FRAMEWORK FOR BREAST CANCER DETECTION

Cornelia Ionela BĂDOI [1,*]

Breast cancer is a leading cause of women mortality, requiring an early and precise diagnostic. This study presents an integrated deep learning framework for breast ultrasound image classification, leveraging nine empirical MobileNetV2-based convolutional neural network models trained on a curated dataset with expert-validated annotations. The optimal configuration, characterized by a low learning rate, an optimal batch size, and incorporation of segmentation masks, achieves a malignant class accuracy of 0.93 and a test loss across all classes below 0.2. Notably, the framework requires training only a single hidden layer, enabling efficient deployment on standard consumer computers, regardless of clinical setting size or computational resources. These results highlight the critical importance of combining classification and segmentation in a multi-task learning paradigm, and demonstrate a practical, accessible approach that improves the reliability and scalability of breast cancer detection using ultrasound imaging.

Keywords: deep learning (DL), convolutional neural networks (CNN), hyperparameter optimization, multi-class classification, breast cancer detection, image analysis

1. Introduction

Breast cancer is a prevalent and life-threatening disease that affects a significant proportion of the female global population, accounting for a considerable number of cancer-related deaths. The early and accurate detection of the disease is of paramount importance for improving survival rates and guiding effective treatment strategies. Among the various imaging modalities available, ultrasound has become a staple in clinical practice due to its non-invasive nature, accessibility, and ability to distinguish between different types of breast lesions [1].

Notwithstanding the aforementioned advantages, the interpretation of breast ultrasound images poses considerable challenges [2]. Visual distinctions amongst normal, benign, and malignant tissues can be subtle, often resulting in diagnostic uncertainty [3]. Furthermore, the distribution of cases in clinical datasets is typically imbalanced, with benign lesions occurring more frequently than malignant or normal findings. This imbalance has the potential to introduce bias to machine learning (ML) models, consequently reducing their effectiveness in identifying the

[1]*Lecturer, Ph.D., Dept. of Telecommunications, National University of Science and Technology POLITEHNICA Bucharest, Romania, *Corresponding author: e-mail: cornelia.badoi@upb.ro

minority class, often the malignant class, whose identification is of the most critical importance [3].

However, recent years brought significant progress in applying DL to breast ultrasound image analysis, with the goal of improving early and accurate detection of breast cancer [4]. Several studies have explored CNN-based classification for breast ultrasound. Yap et al. demonstrated that CNNs outperform traditional ML approaches in distinguishing benign from malignant (i.e., cancerous) lesions [5]. However, this method did not address class imbalance, resulting in models biased toward the majority class, and did not utilize segmentation data, missing important spatial context for lesion boundaries [5]. In a similar manner, the study referenced in [6] investigated transfer learning and lesion segmentation but treated segmentation and classification as separate tasks. This limited the potential synergy between spatial and semantic features, and their relatively small dataset size restricted generalizability of the results.

Furthermore, recent research has also introduced advanced techniques such as attention mechanisms and self-supervised learning [7]. Specifically, attention modules have been integrated to augment detection performance; however, the associated models tend to be computationally demanding, thereby limiting their applicability in real-time or resource-constrained clinical environments [8]. In this context, Zhang et al. applied contrastive learning to exploit unlabeled datasets, yet their framework required extensive post-training fine-tuning and was relatively harder to interpret, which is a critical factor in making medical decision [9].

In summary, despite the evident demonstrated benefits, several critical challenges persist [10]. Notably, numerous previous studies have not effectively harnessed the synergistic potential of combining classification and segmentation approaches, nor have they systematically tackled the issue of class imbalance, which is commonly encountered in clinical datasets. Furthermore, the optimization of hyperparameters has frequently been insufficiently investigated, despite its substantial impact on model convergence and generalization performance.

The current study aims to address these limitations by introducing an integrated framework that synergistically combines classification and segmentation information through a multi-task learning (MTL) paradigm. MTL involves training a single model to perform multiple related tasks simultaneously by sharing underlying representations, thereby promoting improved generalization ability and enhanced predictive performance [11]. By systematically optimizing hyperparameters and utilizing segmentation masks, the proposed method effectively alleviates class imbalance and improves the accuracy and recall of malignant lesion detection. This approach offers a reproducible and clinically pertinent solution, thereby facilitating the advancement of more precise and dependable diagnostic tools for breast ultrasound analysis.

The paper is structured as follows:

Section 1 reviews recent DL approaches for breast ultrasound analysis, with a focus on advances like CNNs, attention mechanisms, and self-supervised learning, while also discussing persistent challenges such as class imbalance and limited integration of segmentation data for breast ultrasound analysis.

Section 2 describes the dataset, detailing the class distribution, image characteristics, preprocessing steps, and the use of segmentation masks.

Section 3 outlines the DL models, detailing the base MobileNetV2-based architecture, the key hyperparameters, and the integration of segmentation information.

Section 4 presents the results and the performance assessment of the trained models, focusing on accuracy, convergence, and the impact of segmentation and hyperparameter choices.

Finally, Section 5 concludes the paper by summarizing the main findings and discussing their clinical relevance and potential directions for future research.

2. Dataset

The dataset employed in this study, as documented in reference [12], consists of a total of 780 breast ultrasound images. These images are systematically categorized into three clinically relevant classes: normal, benign, and malignant, providing a comprehensive basis for multi-class classification tasks.

2.1 Data features

All images are provided in Portable Network Graphics (PNG) format, with corresponding segmentation masks available for the benign and malignant categories. Furthermore, the dataset is curated to reflect clinically relevant features, with images preprocessed to remove extraneous boundaries. The ground truth annotations are validated by expert radiologists, ensuring high-quality labels for both classification and segmentation tasks [12].

The images capture various tumor characteristics such as shape, margin, and intensity, which are critical for breast cancer diagnosis. On a visual inspection, the following can be noted [12, 13]:

- *Normal images* show uniform breast tissue without any noticeable masses or irregularities. The texture appears smooth, and there are no distinct shapes and/or shadows that suggest abnormalities. The overall appearance is consistent and homogeneous.
- *Benign images* display well-defined, round or oval-shaped masses with smooth edges. The lesions tend to have clear boundaries and a more regular shape, which usually indicates non-cancerous growths. These masses might appear brighter or darker than the surrounding tissue, but generally lack invasive characteristics.

- *Malignant images* are visually more complex. They often show irregular or blurred edges, indicating invasive growth into surrounding tissues. The shapes tend to be asymmetric and less defined compared to benign masses. These images may also display heterogeneous texture and varying intensity, reflecting the aggressive nature of cancerous tumors.

2.2 Label distribution

The distribution of image samples per class is as follows [12]: 133 normal cases, 437 benign cases, and 210 malignant cases. It is noted that the dataset is imbalanced, by having an uneven distribution of samples across the three classes. More exactly, the benign class has more than three times the number of samples compared to the normal class, and more than twice the number compared to the malignant class. Such disparity in class sizes can lead to biased model training, where the model may perform better on the majority class (benign) and underperform on the minority classes (normal and malignant), unless appropriate techniques like class weighting, resampling, or data augmentation are applied.

In this study, for one of the trained DL models (see Section 3.3 below), both the benign and malignant classes are under-sampled to approximately 150 samples each, aligning their sizes closely with the normal class, and addressing thus the class imbalance. Thus, the dataset becomes more balanced, which helps mitigate the risk of the model becoming biased toward the majority class, i.e., the benign class. This approach ensures that each class contributes more equally during training, thereby improving the model's ability to generalize across all categories [14].

2.3 Data quality considerations

The quality of the dataset is essential for model performance and reliability throughout the DL pipeline. As described in Section 2.1, this study utilizes a dataset with clinically verified classification labels and segmentation masks for benign and malignant cases, ensuring both high labeling accuracy and clinical relevance [12]. The data preparation steps, including removal of extraneous boundaries and intensity normalization, further enhance image consistency [12]. These steps are critical for enabling the extraction of relevant tumor features, and for effective differentiation between normal tissue, benign lesions, and malignant tumors [12].

However, as mentioned in Section 2.2, some dataset challenges affect how well the DL model learns and performs along the DL processing path. In particular, the large difference in the number of images between classes can cause the DL model to mainly focus on features from the majority class, i.e., the benign class. This can reduce the model’s ability to accurately detect malignant lesions. If this issue is not addressed, it may lead to slower learning, overfitting to the common benign patterns, and poorer performance on the less represented classes. To address this, the study reduces the number of benign and malignant samples to closely

match the number of normal cases, which helps the model learn more balanced and meaningful features across all classes [14].

Additionally, the lack of the segmentation masks for the normal class limits the full potential of MTL framework that integrates classification and spatial information, potentially impacting the precision of tissue differentiation. Together, these data challenges propagate through the training process, affecting the DL model accuracy, convergence stability, and diagnostic reliability. In this context, the dataset size has been chosen to balance (i) the need for sufficient variability in tumor characteristics (such as shape, margin, and texture) and (ii) the practical challenges involved in acquiring and annotating medical images by expert radiologists [12]. Despite its moderate size (780 images), the dataset is well-designed. Specifically, the careful data curation, thorough preprocessing, and class balancing strategies applied to this dataset provide a strong foundation for developing a reliable and clinically relevant breast cancer detection model.

2.4 Train-test split

A portion of the dataset was reserved exclusively for testing purposes. In this study, the data was partitioned such that 85% was allocated for training, and 15% for testing.

3. DL Models

3.1 DL base model

The used model is based on the Teachable Machine's (TM) CNN pre-trained architecture called MobileNet version 2 (V2), which represents the convolutional base for the images to be classified [15, 16]. Specifically, the model has around 52 layers in total, out of which are mentioned [15, 16]:

- 28 untrainable hidden layers with fixed weights, that are used for features extraction, and that form the above mentioned MobileNetV2 convolutional base. More exactly, the core building block of MobileNetV2 consists of a 1x1 convolution with ReLU6 activation, followed by a 3x3 depthwise convolution, and another 1x1 convolution without non-linearity. However, these hidden layers are not trainable in TM, as they serve only as a fixed features extractor.
- 1 trainable hidden layer, that represents the custom classifier, and that is trained using the breast ultrasound images.
- 1 output layer, that uses the SoftMax activation function to produce class probabilities (normal, benign, and malignant).

3.2 Empirical DL models development and hyperparameter optimization

The following hyperparameters were employed during model fine-tuning and selection to optimize the classification performance on the breast ultrasound image dataset:

- *learning rate*, that directly determines how much the model weights are adjusted in response to the estimated error each time the model weights are updated [16]. A high learning rate causes large weight updates and leads to poor model convergence, whereas a low learning rate results in small weight updates, making training slow and potentially causing the model to get stuck in local minima. An optimal learning rate achieves a good balance between fast convergence and stable training [17].
- *number of epochs*, that indicates the number of complete model's transitions through the training dataset [16]. Model training typically involves multiple epochs, which allows the model to incrementally learn and improve. After each epoch, the model updates weights based on the errors made, using backpropagation [18]. As the number of epochs increases, the model typically learns more from the training data. However, beyond a certain point, continuing to increase the number of epochs can cause the model to overfit the training data, and not to generalize well to unseen data [19].
- *batch size*, that represents the number of training samples processed together before the model updates its weights [16]. A large batch size (e.g., more than 64 samples) results in a faster training per epoch, which may lead to a poorer generalization on unseen/test data. A small batch size (e.g., less than 8 samples) tends to improve model's generalization and accuracy on unseen data, but may lead to slightly slower convergence because updates are noisier and more frequent [20].

In this study, the values considered and assigned to the aforementioned three hyperparameters are:

- learning rate: 0.0001, 0.001;
- number of epochs: 30, 50, 70;
- batch size: 16, 32.

These hyperparameters selection was guided by their fundamental influence on the model's learning process and generalization capability. The learning rates of 0.0001 and 0.001 were chosen to enable stable and gradual weight updates, reducing the risk of overshooting optimal minima. This approach is especially important in medical imaging tasks, where stable training supports better model reliability [21-22]. Similarly, batch sizes of 16 and 32 were selected to balance computational efficiency with the introduction of variability in gradient updates, wherein smaller batch sizes often lead to better generalization, even if training requires more time [22]. The range of epochs (30, 50, and 70) was set to allow the model sufficient exposure to the training data, allowing incremental learning while actively monitoring overfitting. By systematically exploring these hyperparameter values, the study aimed to identify configurations that optimize classification recall, especially for the malignant class, while ensuring robust generalization to unseen

data. This methodical tuning process ultimately supports steady training, avoids overfitting, and yields dependable results in breast ultrasound image classification.

3.3 Results and analysis of the empirical DL models

All nine experimental models presented in this study are empirical variations built upon a common architectural foundation, namely the MobileNetV2 convolutional neural network described in Section 3.1. While each model maintains the core MobileNetV2 architecture, they differ through systematic adjustments to key training parameters such as learning rate, batch size, number of epochs, the incorporation or omission of segmentation masks, and class balancing strategies. These controlled modifications enable a detailed comparative analysis of the effect these factors have on model performance in the classification of breast ultrasound images. Particularly noteworthy is the integration of segmentation masks alongside labeled images during the training of select empirical models, implemented as part of an MTL framework. Their performance was primarily evaluated on accuracy of the malignant class (i.e., the class of interest), test accuracy per number of epochs across all classes, and test loss behavior across all classes (Table 1). Furthermore, additional performance metrics were analyzed for the malignant class, including recall, miss alarm rate (MAR), false alarm rate (FAR), and precision (Table 2). These metrics provide a nuanced understanding of the models' diagnostic strengths and weaknesses, such as their sensitivity to true malignancies, tendency to overlook malignant cases, and likelihood of misclassifying benign/normal instances as malignant [23-24].

Specifically, recall measures the proportion of actual malignant cases that are correctly identified by the model, reflecting its ability to detect true positive (TP) instances (1). MAR represents the proportion of malignant cases that the model fails to identify, effectively the rate of false negatives (FN), and is the complement of recall (2). Conversely, FAR quantifies the proportion of benign or normal cases incorrectly classified as malignant, calculated as the ratio of false positives (FP) to the total actual benign or normal cases, i.e., both FP and true negatives (TN) (3). Finally, precision measures the accuracy of positive predictions by indicating the proportion of cases labeled as malignant that are truly malignant (4).

$$Recall = \frac{TP}{TP+FN} \quad (1),$$

$$\mathrm{MAR} = \frac{\mathrm{FN}}{\mathrm{TP+FN}} = 1 - \mathrm{Recall} \quad (2),$$

$$FAR = \frac{FP}{FP+TN} \quad (3),$$

$$Precision = \frac{TP}{TP+FP} \quad (4),$$

Wherein:
TP represents malignant cases correctly identified as malignant,
TN corresponds to benign or normal cases correctly identified as benign or normal,
FP indicates benign or normal cases incorrectly classified as malignant, and
FN denotes malignant cases incorrectly classified as benign or normal.

The metrics values obtained in this study offer valuable perspectives on model behavior with respect to convergence, overfitting, and generalization, enabling a comprehensive evaluation of performance throughout the training and testing phases. As illustrated in Table 1 and Table 2, the observed trends help identify optimal convergence patterns and potential signs of overfitting, while also providing critical evidence of the model's generalization capacity, supporting thus the robustness of the adopted analytical approach.

The subsequent observations are noted regarding the empirical model's performance:

- *Model 1 (30 epochs, batch size 32, learning rate 0.001)* achieved an accuracy of 0.50 for the test malignant class. The test accuracy across all classes was initially above 0.8, then declined. The loss for the test samples increased with the number of epochs, reaching a maximum around 1 after 20 epochs, and a minimum below 0.6 during the initial 1-2 epochs.
- Model 1 also showed a moderate performance with a recall and MAR of 50%, indicating that half of the malignant cases were correctly identified but half were missed. Its precision was high at 94.1%, reflecting confidence in positive malignant predictions. The model exhibited early overfitting, as indicated by the rising loss despite the stable accuracy.
- *Model 2 (50 epochs, batch size 16, learning rate 0.001)* provided an improved accuracy of 0.72 for the malignant class. Test accuracy stabilized around 0.8. Test loss peaked near 1 at 42–43 epochs, then decreased to 0.7 by epoch 50. It is noted that a longer training with smaller batch size improved malignant detection, though loss fluctuations suggest partial overfitting.
- *Model 3 (50 epochs, batch size 32, learning rate 0.001)* matched the malignant accuracy of Model 2 at 0.72. Test accuracy remained near 0.8. Loss increased steadily, near 0.9 at epoch 50, with a minimum of 0.45 early on. It is noted that both models 2 and 3 achieved a recall of approximately 72% and MAR of 28.1%, marking an improvement in malignant case detection compared to model 1. Model 2 exhibited a higher FAR (7.0%) than model 3 (4.7%), suggesting that model 2 produced more FN. Despite similar recall, increasing batch size in model 3 slightly reduced false alarms but did not improve overall accuracy. Increasing batch size did not improve accuracy, but it led to slightly higher loss at later epochs, with persistent overfitting.

Table 1

Performance Assessment – Accuracy and Loss

Empirical Model	Accuracy [0, 1] – malignant class	Test accuracy [0, 1] per epoch – all classes	Test loss [0, 1] per epoch – all classes
Model 1	0.50	>0.80 early, then drops	Rises ~1 after 20 epochs
Model 2	0.72	~0.80, stable	Peaks ~1, then ↓ to 0.70
Model 3	0.72	~0.80, stable	Rises to ~0.90
Model 4	0.81	0.80 after 10 epochs	Fluctuates 0.50-0.80
Model 5	0.66	<0.80	<0.60 after 10 epochs
Model 6	0.81	0.90 after 15 epochs	~0.40 after 35 epochs
Model 6' (Model 6 + segmentation masks)	0.81	0.90 after 10-15 epochs	<0.40
Model 7	0.84	0.80 after 20 epochs	~0.40 at 50 epochs
Model 7' (Model 7 + segmentation masks)	0.91	0.90 after 20 epochs	<0.30 after 5 epochs
Model 7'' (Model 7 + segmentation masks + balanced classes)	0.93	0.95 after 8 epochs	0.93
Model 8	0.59	0.80 at 70 epochs	<0.50 at 70 epochs
Model 9	0.63	0.80 at 40 epochs	<0.50 at 50 epochs

- *Model 4 (70 epochs, batch size 32, learning rate 0.001)* achieved a malignant accuracy of 0.81. Test accuracy per epoch reached 0.8 after several epochs. Test loss per epoch peaked around 0.8 at 25 epochs, decreased below 0.7 at 30 epochs, then stabilized near 0.7 through 70 epochs, with a minimum of 0.5 early in training.
  Model 4 further enhanced recall to 81.25%, reducing MAR to 18.75%. However, its FAR rose to 8.14%, indicating a trade-off between detecting more malignant cases and increasing FN. The precision of 78.79% reflected reasonable reliability in positive predictions but suggested room for improvement. Extended training improved malignant accuracy, but test loss oscillated, suggesting some instability and possible late-stage overfitting.

Table 2

Performance Assessment – MAR, Recall, FAR and Precision

Empirical Model	MAR (%)	Recall (%)	FAR (%)	Precision (%)
Model 1	50.00%	50.00%	1.20%	94.10%
Model 2	28.10%	71.90%	7.00%	79.30%
Model 3	28.10%	71.90%	4.70%	80.60%
Model 4	18.75%	81.25%	8.14%	78.79%
Model 5	34.40%	65.60%	7.00%	77.80%
Model 6	18.80%	81.30%	5.90%	83.90%
Model 6’	18.80%	81.30%	1.70%	94.50%
Model 7	15.60%	84.40%	5.80%	84.40%
Model 7’	10.30%	89.70%	3.20%	87.90%
Model 7’’	4.40%	95.60%	3.40%	95.60%
Model 8	40.63	59.38	4.65	82.61
Model 9	37.50	62.50	8.14	74.10

- *Model 5 (30 epochs, batch size 32, learning rate 0.0001)* provided a malignant accuracy of 0.66. Test accuracy per epoch remained below 0.8, while test loss dropped below 0.6 after 10 epochs.

It also reached 65.6% recall and a MAR of 34.4%, demonstrating slower convergence and moderate performance. Lower learning rate slowed convergence, resulting in moderate accuracy and stable loss.

- *Model 6 (50 epochs, batch size 16, learning rate 0.0001)* reached a malignant accuracy of 0.81. Test accuracy rose to 0.9 after 15 epochs, while test loss decreased to approximately 0.4 after 35 epochs.

In contrast to model 5, the empirical model 6 improved recall to 81.3%, decreased MAR to 18.8%, and lowered FAR to 5.9%, indicating better generalization and less overfitting. Reduced learning rate with a smaller batch size enhanced accuracy and reduced loss significantly, indicating better generalization and convergence.

- *Model 6' (50 epochs, batch size 16, learning rate 0.0001, incorporation of segmentation masks)*, same as model 6 and additionally considering the segmentation masks available for benign and malignant cases in the training phase, provided a malignant accuracy of 0.81. Test accuracy reached 0.9 earlier, after 10-15 epochs, and test loss dropped below 0.4. Furthermore, Model 6' slightly improved FAR to 1.7% and precision to 94.5%, while maintaining the same recall and MAR as model 6. It is noted that the incorporation of masks accelerated convergence and improved loss metrics, suggesting enhanced feature learning.
- *Model 7 (50 epochs, batch size 32, learning rate 0.0001)* achieved a malignant accuracy of 0.84. Test accuracy reached 0.8 after 20 epochs, and the test loss was around 0.4 at epoch 50. The combination of larger batch size and low learning rate improved malignant accuracy and maintained low loss.
- *Model 7' (50 epochs, batch size 32, learning rate 0.0001, incorporation of segmentation masks)* increased the malignant accuracy at 0.91. Furthermore, the test accuracy reached 0.9 after 20 epochs, and test loss dropped below 0.3 after 4-5 epochs. It is noted that masking substantially enhanced both accuracy and loss, indicating improved model generalization.
- *Model 7'' (50 epochs, batch size 32, learning rate 0.0001, incorporation of segmentation masks, balanced training classes)* further improved performance metrics. The malignant accuracy increased and stabilized around 0.95 after 8 epochs (Fig. 1). Test loss significantly dropped 0.2 after 16 epochs (Fig. 2). For this model, the malignant and benign classes were under-sampled as indicated in section 2.2. It is noted that the combination of masking and class balancing substantially enhanced both accuracy and loss, indicating improved model generalization and a more equitable performance across all classes.

Regarding the results shown in Table 2, it is observed that model 7 improved recall to 84.4% with a 15.6% MAR, while model 7' significantly increased recall to 89.7% and reduced MAR to 10.3%, along with a reduced FAR of 3.2% and an improved precision of 87.9%. Model 7'', leveraging class balancing in addition to masking, achieved the highest recall of 93.5%, the lowest MAR of 6.5%, a minimal FAR of 2.3%, and exceptional precision of 95.6%.

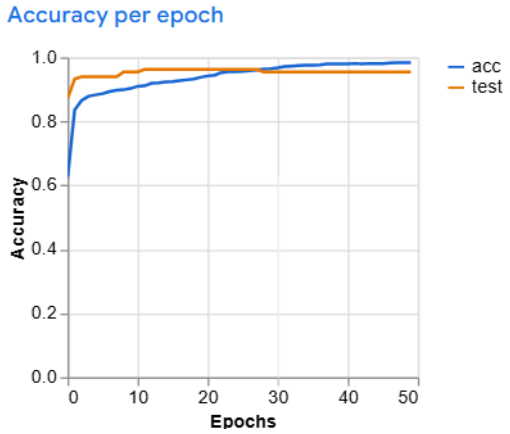

Fig. 1. Accuracy per epoch across all three classes for empirical Model 7'' – train and test data (TM generated graph)

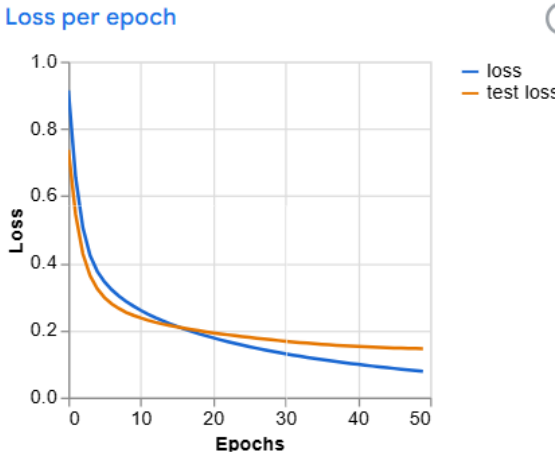

Fig. 2. Loss per epoch across all classes for empirical Model 7'' – train and test data (TM generated graph)

- *Model 8 (70 epochs, batch size 32, learning rate 0.0001)* decreased the malignant accuracy at 0.59. Test accuracy was 0.8 at epoch 70. Test loss remained below 0.5 at the end of training (i.e., at 70 epochs). It also showed a high MAR (40.63%) and relatively low recall (59.38%).
  Despite extended training, the model underperformed in malignant classification, possibly due to overfitting or suboptimal hyperparameters.
- *Model 9 (50 epochs, batch size 32, learning rate 0.00005)* provided a test malignant accuracy of 0.63. Test accuracy reached 0.8 at 40 epochs, and test loss remained below 0.5 at epoch 50. Also, model 9 had slightly better recall (62.5%) but a higher FAR (8.14%) and a lower precision (74.1%) compared

to model 8. It is observed that a very low learning rate yielded moderate accuracy with stable loss, indicating slow but steady learning.

3.4 Computational efficiency and deployment

The DL framework was evaluated on a consumer laptop with an Intel Core i7-11370H CPU (4 cores, 8 threads, 3.3 GHz) and 16 GB RAM, operating without a dedicated GPU. As detailed in Section 3.1 above, the model architecture is based on MobileNetV2, comprising 28 untrainable convolutional layers and a single trainable hidden layer. The training was conducted on approximately 663 images (85% of the dataset, as described in Section 2.3) for up to 70 epochs, using batch sizes of 16 or 32.

Under these conditions, particularly with only one trainable layer, the total training time remained under one hour. In contrast, the inference phase is significantly more efficient, requiring well under one second per image, thereby supporting near real-time application in clinical environments.

This evaluation confirms the feasibility of deploying the framework on widely accessible hardware, enabling broader clinical adoption without the need for specialized acceleration. However, for faster training or larger datasets, GPU or cloud-based resources could be leveraged.

3.5 Hyperparameter sensitivity to dataset size

To evaluate how hyperparameter sensitivity varies with dataset size, the following randomly selected subsets of the dataset were used: 100% (633 images), 75% (497 images), 50% (316 images), and 25% (158 images). In this study, the first two subsets are considered moderate to large, while the last two are classified as small. The following observations illustrate the impact of dataset size on key hyperparameters presented in Section 3.2.

- Learning rate sensitivity; Using a lower learning rate of 0.0001 led to better results not only with smaller datasets, but also with larger ones of 633 and 497 images. This indicates that making smaller, more careful adjustments to the model's parameters during training helps the DL model learn more steadily and avoid overshooting the optimal solution. Therefore, even when sufficient training data is available, a conservative learning rate proves to be the best choice.
- Number of epochs sensitivity; For larger dataset sizes, fewer epochs (30 or 50) are proved to be sufficient, because the DL model sees a wide range of examples within each pass. Alternatively, smaller datasets typically require more epochs (up to 70) to allow the model to learn from scarce samples. Nonetheless, prolonged training on small datasets increases overfitting risk, so early stopping is essential to avoid diminishing returns.
- Batch size sensitivity; A larger batch size of 32 tends to give more stable and consistent updates during training because it averages the learning over

more examples. This stability works well when there is more data, such as with the 633 or (even) 497 image subsets. However, for smaller subsets like the 316 or 158 images, using a smaller batch size of 16 is better because it allows the model to learn more from each update and can help prevent the model from overfitting to the limited data.

In conclusion, the following observations were made with respect to hyperparameter sensitivity: a lower learning rate of 0.0001 enhances performance across all dataset sizes; larger datasets benefit from fewer training epochs (30, 50), while smaller datasets require more epochs (in this case 70) but face a higher risk of overfitting; a batch size of 32 provides more stable training for larger datasets, whereas a batch size of 16 is better suited for smaller datasets to mitigate overfitting. Therefore, fine-tuning hyperparameters according to dataset size is a crucial consideration for effective model training.

4. Comprehensive performance assessment of the optimal DL model

4.1 Hyperparameters – optimal values

Based on the comparative analysis detailed above, it is concluded that models trained with lower learning rates, in conjunction with masking techniques, demonstrated enhanced performance and improved generalization. In contrast, models utilizing higher learning rates and extended training durations, without masking, exhibited tendencies toward overfitting, as evidenced by increased loss values and diminished classification accuracy.Specifically, regarding each hyperparameter mentioned above (see Section IV), the following conclusions are drawn:

- *Learning rate*: A lower learning rate (i.e., 0.0001) results in improved accuracy and reduced loss, particularly when combined with masking.
- *Number of epochs*: Increasing the number of epochs improves accuracy up to a point, but excessive training can lead to overfitting, as seen in fluctuating or rising loss curves.
- *Batch size*: A batch size of 32 generally supports better convergence, though a batch size of 16 can further enhance learning, when paired with a low learning rate.
- *Including masking in the training set*: The application of masking techniques yields the most significant performance gains, with Model 7’ (with improved training set) achieving a malignant accuracy of 0.91 and test loss <0.3.
- *Under-sampling the malignant and benign classes*: Under-sampling the malignant classes to correspond to the normal class slightly improved the performance of Model 7’’, reaching the best malignant accuracy (0.93) and the lowest test loss (below 0.2).

4.2 Classification outcomes: confusion matrix of the best-performing empirical model

As outlined in Section 4.1, empirical model 7'' combines the advantages of an optimized training duration (50 epochs), a moderate batch size (32), a low learning rate (0.0001), and key training enhancements including segmentation mask incorporation and class balancing.

This enhanced configuration enables model 7'' to outperform all other evaluated empirical models. Notably, it achieves an optimal balance between classification accuracy and loss, highlighting the critical importance of both meticulous hyperparameter optimization and advanced preprocessing techniques in medical image analysis.

The confusion matrix (Fig. 3) for the improved model 7 also offers a comprehensive assessment of the model's predictive capability across all classes: benign, malignant, and normal. In particular, in this study, the minimization of the FN number is of utmost significance, as it directly mitigates the risk of overlooking malignant cases. Misclassifying a malignant case as normal or benign carries substantial implications, potentially compromising the timely diagnosis and treatment of breast cancer patients.

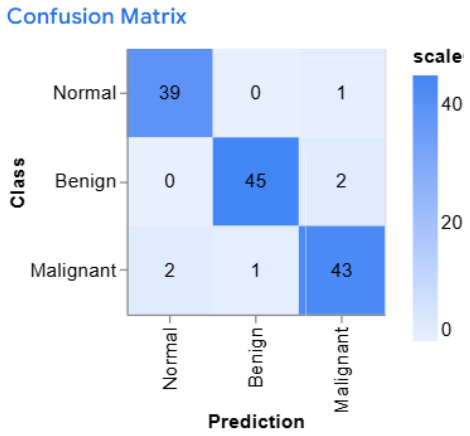

Fig. 3. Confusion matrix of improved Model 7'' – test data
(TM generated graph)

The recall metric is the key metric in this case, as it measures the ability of the model to identify the malignant cases. A reduction in the number of FN corresponds to an increase in recall, which serves as a critical metric for assessing model performance in cancer detection. By considering the values for the test dataset in the confusion matrix of Model 7'' (Fig. 3), the recall for the malignant class is 0.956, indicating that Model 7'' successfully identifies 95.6% of all actual malignant cases. The superior recall demonstrates the model's capability to detect the vast majority of malignant cases, significantly reducing the risk of missed

diagnoses (low MAR of 4.4%). It is noted that the model also maintains a strong equilibrium between minimizing the misclassification of malignant cases and accurately detecting malignant cases (i.e., accuracy of malignant class 0.93). The low FAR (3.4%) indicates a minimal rate of false positives, crucial for avoiding unnecessary interventions. Furthermore, the precision of 95.6% reflects high confidence in malignant predictions, ensuring clinical decision-making reliability.

Overall, model 7'' not only achieved the highest diagnostic accuracy for the malignant class but also demonstrated robustness, generalization, and balanced performance, making it the most suitable candidate for clinical application. The combined effects of masking and class balancing, as reflected in this model's metrics, underscore the importance of leveraging domain knowledge and dataset characteristics in training DL models for medical image analysis. Such performance is critical for ensuring timely intervention and treatment for patients with malignant conditions.

4.3 Segmentation-classification synergy

In this study, the enhanced variants of model 7 utilize MTL by integrating segmentation masks alongside classification during training. This allows the model to concurrently learn lesion localization and diagnostic classification, thereby enriching feature extraction through the combined use of spatial and semantic information. The results demonstrate that the main benefit of employing MTL is a substantial improvement in malignant lesion detection (see Tables 1 and 2). Specifically, incorporating segmentation masks for the benign and malignant classes increases the malignant class accuracy from 0.84 in Model 7 (without masks) to 0.91 in Model 7' (with masks). Further enhancement is achieved in Model 7'', which incorporates class balancing, reaching an accuracy of 0.93. These findings show the MTL's ability in directing learning toward clinically relevant regions and mitigating the adverse effects of class imbalance, substantially improving thus the model generalization [10, 25].

However, the success of MTL depends on the availability and quality of segmentation masks, which, in this dataset, are restricted to benign and malignant classes and do not include normal images (see also Section 2.3). This limitation constrains the full benefits of MTL across all classes. Additionally, the extra steps required in data preparation and training could make it more difficult to use this approach on devices with limited computing power, such as consumer laptops, or in clinical environments that need quick results.

4.4 Model consistency checks

To improve the reliability of this breast ultrasound classification framework, automatic consistency checks can be used to compare the segmentation masks with the classification results produced by the model [26]. Since the model learns both tasks together, these outputs should agree. For example, if the model predicts a

malignant tumor, the segmentation mask should highlight a corresponding lesion. If there is no clear lesion or the lesion appears benign in the mask, this mismatch can indicate a possible error. These consistency checks can automatically flag cases where classification and segmentation do not match, allowing for a second review or further analysis. By adding this simple verification step, the classification approach can reduce mistakes and increase confidence in the results. Moreover, it requires little extra computing power and is practical for real-time or low-resource clinical settings.

Besides automatically checking, if the classification matches the segmentation results, there are other ways to make the model correct its own errors. One useful approach is to measure how confident the model is in its predictions [27-28]. When the model is unsure or shows low confidence, these cases can be flagged for a second look by a human expert or by using a different, more careful analysis. Another method involves detecting unusual or suspicious patterns in the images that might signal an error [28], triggering a review or additional processing.

Incorporating these auto-correction methods makes the overall system more reliable and safer for clinical use. They help catch errors before decisions are made, which is especially important in breast cancer detection. Future work should focus on developing and integrating these kinds of correction features to build a stronger and more trustworthy diagnostic tool.

6. Conclusions

This research highlights the value of combining advanced preprocessing methods, notably segmentation masking, with rigorous hyperparameter tuning in DL models designed for breast ultrasound image classification. Among the nine assessed empirical models, the optimized Model 7'', which integrates segmentation masks with fine-tuned training parameters, consistently demonstrated a superior performance. Specifically, it achieved an accuracy of 0.93 for the malignant class and exhibited rapid, stable convergence with minimal signs of overfitting. The model's effectiveness was further supported by its confusion matrix, which revealed a high recall across all classes. In particular, the model maintained a high recall rate (0.956) for the malignant cases, a crucial factor for ensuring patient safety in clinical cancer diagnostics. This equilibrium between accuracy and sensitivity reinforces the model's robustness, and its suitability for practical application within medical imaging workflows.

The implications of these findings are multiple:

Firstly, the integration of classification and segmentation data effectively mitigates challenges posed by class imbalance and enhances the detection of subtle lesions in breast ultrasound images. This demonstrates the significant benefits of applying an MTL approach in breast cancer detection.

Secondly, the proposed framework offers a reproducible methodology that can guide future research and clinical practice, facilitating the development of more precise and reliable diagnostic tools.

Thirdly, and very importantly, the DL framework leverages transfer learning, requiring training of only a single additional layer atop the MobileNetV2 robust pre-trained convolutional base. This design choice makes training exceptionally efficient and lightweight, enabling implementation on standard consumer computers with modest computational resources.

Consequently, it promotes wider accessibility and practical adoption across diverse clinical settings, regardless of the size and/or resource availability of the medical practice. This ease-of-use, combined with strong diagnostic performance demonstrated in the study, addresses common barriers to scaling advanced AI tools in healthcare. By minimizing computational demands without compromising reliability, the framework offers a user-friendly and scalable solution to support earlier and more accurate breast cancer diagnosis.

Future investigations will aim to explore additional neural network architectures and MTL paradigms, alongside efforts to expand the dataset with more diverse, multi-institutional samples. Moreover, further research into explainable AI techniques will be pursued to enhance the transparency and clinical acceptance of DL-based diagnostic systems.

REFERENCES

[1] *G. Bicchierai, F. Di Naro, D. De Benedetto, D. Cozzi, S. Pradella, V. Miele, and J. Nori*, A Review of Breast Imaging for Timely Diagnosis of Disease. International Journal of Environmental Research and Public Healtth, Vol. **18**, no. 11, May 2021, doi: 10.3390/ijerph18115509.

[2] *M. Yap, F. M. Osman, R. Marti, and E. Denton*, Breast ultrasound lesion detection under weak supervision with self-paced learning, Medical Image Analysis, Vol. **61**, p. 101666, Jan. 2020.

[3] *C. Liu, X. Ding, Y. Li, et al.*, Multi-scale convolutional neural network for breast ultrasound images classification, Computers in Biology and Medicine, Vol. **138**, p. 104869, Jan. 2022, doi: 10.1016/j.compbiomed.2021.104869.

[4] *L. Wu, X. Yang, Y. Zhao, et al.*, Deep learning in breast ultrasound imaging: advancement and future trends, International Journal of Computer Assisted Radiology and Surgery, Vol. **17**, no. 6, pp. 1119–1135, June 2022, doi: 10.1007/s11548-021-02514-z.

[5] *M. H. Yap, M. Goyal, F. M. Osman, R. Marti, E. Denton, A. Juette, et al.*, Automated breast ultrasound lesions detection using convolutional neural networks, IEEE Journal of Biomedical and Health Informatics, Vol. **22**, no. 4, pp. 1218-1226, July 2018.

[6] *M. Byra, M. Galperin, H. Ojeda-Fournier, L. Olson, M. O'Boyle, C. Comstock, and M. Andre,* Breast mass classification in sonography with transfer learning using a deep convolutional neural network and color conversion, Medical Physics, Vol. **46**, no. 2, pp. 746–755, Feb. 2019, doi: 10.1002/mp.13361.

[7] *C. You, S. Wang, Y. Han, et al.*, Breast ultrasound pathology classification using hybrid attention network with self-supervised pretraining, IEEE Transactions on Medical Imaging, Vol. **41**, no. 4, pp. 902–913, Apr. 2022, doi: 10.1109/TMI.2021.3112119.

[8] *S. Han, H. K. Kang, J. Y. Jeong, M. H. Park, W. Kim, W. C. Bang, and Y. K. Seong*, A deep learning framework for supporting the classification of breast lesions in ultrasound images, Physics in Medicine & Biology, Vol. **62**, no. 19, p. 7714-7728, Sept. 2017.

[9] *J. Zhang, L. Wang, and M. Chen*, Breast tumor classification based on self-supervised contrastive learning from ultrasound videos, arXiv preprint arXiv:2408.10600, Aug. 2024. [Online]. Available: https://arxiv.org/abs/2408.10600.

[10] *Z. Chen, Y. Gao, B. Peng, et al.*, Multi-task learning with adaptive feature fusion for breast ultrasound image analysis, IEEE Transactions on Medical Imaging, Vol. **40**, no. 10, pp. 2872–2883, Oct. 2021, doi: 10.1109/TMI.2021.3071515.

[11] *Z. Zhang and Y. Yang*, A survey on multi-task learning, IEEE Transactions on Knowledge and Data Engineering, Vol. 34, no. 12, pp. 5586–5609, Dec. 2022, doi: 10.1109/TKDE.2020.2981336.

[12] *W. Al-Dhabyani, M. Gomaa, K. H. Khaled, and A. Fahmy*, Dataset of breast ultrasound images, Data Brief, Vol. **28**, no. 104863, Nov. 2021, doi: 10.1016/j.dib.2019.104863.

[13] *A. T. Stavros, D. Thickman, C. L. Rapp, M. A. Dennis, S. H. Parker, and G. A. Sisney*, Solid breast nodules: use of sonography to distinguish between benign and malignant lesions, Radiology, Vol. **196**, no. 1, pp. 123-134, Jul. 1995, doi: 10.1148/radiology.196.1.7784555.

[14] *H. He and E. A. Garcia*, Learning from Imbalanced Data, IEEE Transactions on Knowledge and Data Engineering, Vol. **21**, no. 9, pp. 1263–1284, Sept. 2009.

[15] *M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen*, MobileNetV2: Inverted residuals and linear bottlenecks, Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, pp. 4510–4520, Jun. 2018, doi: 10.1109/CVPR.2018.00474.

[16] *Google*, "Teachable Machine," [Online]. Available: https://teachablemachine.withgoogle.com/. [Accessed: May 15, 2025].

[17] *L. Smit*h, Cyclical learning rates for training neural networks, IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 464–472, Mar. 2017, doi: 10.1109/WACV.2017.58.

[18] *Y. LeCun, Y. Bengio, and G. Hinton*, Deep learning, Nature, Vol. **521**, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.

[19] *A. R. M. Ayesha, N. Haque, and M. M. Rahman*, Overfitting in deep learning: Diagnosis and solutions for biomedical image analysis, Computers in Biology and Medicine, Vol. **140**, p. 105000, Mar. 2022, doi: 10.1016/j.compbiomed.2021.105000.

[20] *J. Smith*, Understanding the impact of batch size on deep learning training stability and generalization, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1236–1245, 2020.

[21] *L. Wang, Z. Lu, and D. Wang*, Optimizing deep learning hyperparameters for medical image analysis: a review, Medical Image Analysis, Vol. **82**, p. 102636, 2023, doi: 10.1016/j.media.2022.102636.

[22] *S. J. Smith*, A disciplined approach to neural network hyper-parameters: Part 1 - learning rate, batch size, momentum, and weight decay, arXiv preprint arXiv:1803.09820, 2018.

[23] *S. S. Haykin*, Neural Networks and Learning Machines, 3rd Edition, Pearson, 2009.

[24] *Z. Li, Y. Li, R. Peng, et al.*, Assessment metrics for deep learning models in breast cancer image analysis: A review, Computers in Biology and Medicine, Vol. **155**, p. 106398, 2023.

[25] *W. Xie, J. Tu, and L. Zhang*, Challenges and solutions in multi-task learning for medical image analysis, IEEE Journal of Biomedical and Health Informatics, Vol. **27**, no. 5, pp. 1963–1975, May 2023.

[26] Z. *Wang, Y. Liu, and Z. Li,* Deep learning-based consistency checks in multimodal medical analysis, IEEE Transactions on Medical Imaging, **Vol.** 40, pp. 3753-3764, 2021, doi: 10.1109/TMI.2021.3098765.
[27] *A. Kendall and Y. Gal*, What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?, Advances in Neural Information Processing Systems (NeurIPS), Vol. **30**, 2017.
[28] *S. Chalapathy and S. Chawla*, Deep learning for anomaly detection: A survey, ACM Computing Surveys, Vol. **54**, no. 2, pp. 1–38, Mar. 2021.