Détection du cancer du sein à partir de la mammographie par l'utilisation de l’IA

CHIBI Mohammed
mohammed.chibi@um5r.ac.ma

ESSAMADI Oussama
oussama.essamadi@um5r.ac.ma

ZOUIR Amine
amine.zouir@um5r.ac.ma

Abstract

This research paper presents a deep learning approach to breast cancer detection using mammograms. The study develops a Convolutional Neural Network (CNN) model to classify mammogram images as either benign or malignant. The model is trained and evaluated on a dataset of mammogram images, achieving an accuracy of 90.24% on the test set. The study also investigates the model's performance using various metrics, including precision, recall, F1-score, and AUC. Additionally, the study presents a visual analysis of the model's predictions, highlighting its ability to accurately classify mammogram images. The findings suggest that deep learning models like CNNs hold significant potential for improving breast cancer detection accuracy.

I. Introduction

Breast cancer poses a significant global health challenge, ranking among the leading causes of cancer-related deaths in women. The importance of early detection in improving treatment outcomes and increasing survival rates cannot be overstated. Mammography stands as a widely utilized imaging technique for breast cancer screening. However, the interpretation of mammograms can be complex, and even skilled radiologists may occasionally overlook subtle indicators of malignancy [1].
Artificial intelligence (AI) has shown considerable promise in enhancing medical image analysis, including mammography. Deep learning, a subfield of AI, leverages artificial neural networks with multiple layers to extract intricate patterns from data [2]. This paper delves into the potential of deep learning for breast cancer detection using mammograms. A Convolutional Neural Network (CNN) model is developed to categorize mammogram images as benign or malignant [3]. The model is trained on an extensive dataset of mammogram images and its performance is rigorously evaluated using various metrics. The study's findings highlight the efficacy of deep learning in the automatic classification of mammograms. This capability has the potential to aid radiologists in achieving more precise diagnoses, ultimately contributing to improved patient care.

II. Methods

1. The Dataset

The study utilizes the Augmented *INbreast Dataset* [4], a collection of 7,632 grayscale mammogram images, an isolated version of the augmented INbread dataset [5] belonging to the Dataset of breast mammography images with masses [6], each categorized as either benign or malignant. These images were pre-processed to mask extraneous image content; each image in the dataset is a cutout focusing on the breast tissue. The dataset is split into three distinct subsets:

- *Training set*: This subset comprises 3,816 images and is used to train the CNN model, allowing it to learn the features associated with benign and malignant cases.
- *Validation set*: Consisting of 1,908 images, this subset is used during the training process to monitor the model's performance on unseen data and to fine-tune the model's parameters, helping to prevent overfitting.
- *Test set*: With 1,908 images, this subset is held back until after the model is fully trained. It provides an unbiased evaluation of the model's performance on completely new data, giving a realistic estimate of its generalization ability.

By dividing the data into these three subsets, the study ensures a rigorous evaluation of the model's performance and reduces the risk of overfitting, leading to more reliable results when applied to real-world cases.

2. Image Preprocessing

Normalizing pixel values is a crucial step in preparing image data for deep learning models. In this study, all images in the dataset undergo this preprocessing step to scale their pixel values to a range of 0-1.

Mammogram images, like most digital images, typically have pixel values ranging from 0 to 255 [7]. These values represent the intensity of light captured by the imaging sensor for each pixel. To normalize the pixel values, each value is divided by 255. This mathematical operation effectively scales the entire range of pixel intensities from 0-255 to a standardized range of 0-1.

This normalization process offers several benefits:

- *Improved Model Training*: Neural networks, including CNNs, generally perform better when the input data is scaled to a consistent range [3]. Normalization helps to prevent large variations in pixel values from disproportionately influencing the model's learning process. This can lead to faster convergence during training and improved overall model performance.

- *Preventing Vanishing/Exploding Gradients*: By scaling the pixel values to a smaller range, normalization helps to prevent issues with vanishing or exploding gradients during backpropagation, which is the process by which the model learns from its errors. This is particularly important in deep networks with multiple layers.

- *Enhanced Generalization*: Normalization can also contribute to better generalization, meaning the model is more likely to perform well on unseen data. This is because the model is less sensitive to the specific range of pixel values present in the training set.

In the context of this study, normalizing pixel values to a range of 0-1 was an important step in preparing the mammogram images for training the CNN model. This preprocessing technique likely contributed to the model's high accuracy in classifying benign and malignant cases.

3. CNN Model Architecture

The CNN model architecture used in the study employs a sequential design, where layers are stacked linearly, with each layer's output feeding directly into the next.
This specific architecture is well-suited for image classification tasks like identifying breast cancer in mammograms [2].

The model consists of the following layers:

- *Three Convolution-and-Pooling Blocks for Spatial Feature Extraction*:
    - *Convolutional layers*: These layers form the foundation of a CNN. They use small filters (also known as kernels) to scan the input image, extracting features like edges, textures, and patterns. In this model, the convolutional layers use 32 or 64 filters with a 3x3 kernel size. The number of filters dictates the depth or number of feature maps produced by the convolutional layer. Increasing the filter count allows the model to learn more complex features. The ReLU (Rectified Linear Unit) activation function is applied after each convolutional layer to introduce non-linearity [8], which is crucial for learning complex patterns.

    - *Max pooling layers*: These layers reduce the spatial dimensions of the feature maps generated by the convolutional layers [9]. They work by selecting the maximum value within a small region (here, 2x2 pool size) and discarding the rest. Max pooling helps the model become more robust to small variations in the position of features within the image and reduces computational complexity.

- *Flattening Layer*:
    - The output of the last convolution-and-pooling block is a 3-dimensional tensor representing multiple feature maps. To feed this into a dense layer, which expects a 1D vector, a flattening layer is used to convert the 3D feature maps into a 1D vector.

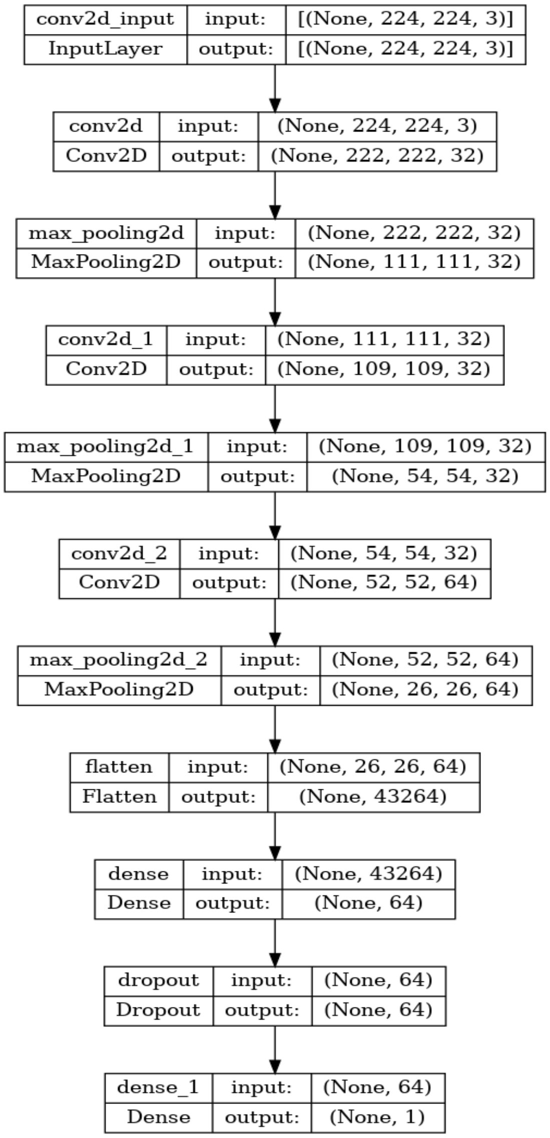
- *Two dense layers*:
    - First Dense Layer: This layer is a fully connected layer with 64 neurons. It processes the flattened feature vector, learning high-level combinations of features from the previous layers. It also uses the ReLU activation function.

    - Second Dense Layer (Output Layer): This layer has a single neuron and uses the sigmoid activation function to produce a probability between 0 and 1. This probability represents the model's confidence that the input mammogram image belongs to the malignant class.

- *Dropout Layer*:
  - This layer is placed after the first dense layer to reduce overfitting, a common problem in machine learning where a model learns the training data too well and performs poorly on new data. Dropout randomly deactivates a proportion of neurons [10] (here, 50%) during each training step, forcing the network to learn more robust and generalized features.

This architecture (visualized in *Figure 1*), combining convolutional and pooling layers for spatial feature extraction, a flattening layer to transition to fully connected layers, and dense layers for classification, is typical for CNNs used in image classification. The addition of dropout helps the model generalize better to unseen mammogram images. The output of the sigmoid activation function in the final layer provides a probability score, which can then be thresholded to classify the mammogram image as either benign or malignant.

Figure 1: A Convolutional Neural Network (CNN) *with three convolution + pooling blocks and two dense layers, designed for binary image classification*


conv2d_input | input: | [(None, 224, 224, 3)]
InputLayer | output: | [(None, 224, 224, 3)]
conv2d | input: | (None, 224, 224, 3)
Conv2D | output: | (None, 222, 222, 32)
max_pooling2d | input: | (None, 222, 222, 32)
MaxPooling2D | output: | (None, 111, 111, 32)
conv2d_1 | input: | (None, 111, 111, 32)
Conv2D | output: | (None, 109, 109, 32)
max_pooling2d_1 | input: | (None, 109, 109, 32)
MaxPooling2D | output: | (None, 54, 54, 32)
conv2d_2 | input: | (None, 54, 54, 32)
Conv2D | output: | (None, 52, 52, 64)
max_pooling2d_2 | input: | (None, 52, 52, 64)
MaxPooling2D | output: | (None, 26, 26, 64)
flatten | input: | (None, 26, 26, 64)
Flatten | output: | (None, 43264)
dense | input: | (None, 43264)
Dense | output: | (None, 64)
dropout | input: | (None, 64)
Dropout | output: | (None, 64)
dense_1 | input: | (None, 64)
Dense | output: | (None, 1)

4. Model Training

The training process for the CNN model involves iteratively adjusting the model's parameters to minimize the difference between its predictions and the actual labels of the mammogram images in the training set. This process uses the following key components:

- *Epochs*: An epoch represents one complete pass through the entire training dataset. The model is trained for 25 epochs, meaning it sees and learns from the full training set 25 times. The number of epochs is a hyperparameter that can be adjusted; training for more epochs can sometimes lead to better performance, but it also increases the risk of overfitting.

- *Optimizer*: An optimizer is an algorithm that determines how the model's parameters are updated based on the calculated gradients during backpropagation. The study employs the *rmsprop* optimizer, a widely used optimization algorithm that adapts the learning rate for each parameter based on the recent magnitude of gradients. Rmsprop is often effective in preventing oscillations and achieving faster convergence during training [11].

- *Loss function*: The loss function quantifies the difference between the model's predictions and the true labels. The model uses the binary cross-entropy loss function, which is a common choice for binary classification problems [12]. Binary cross-entropy measures the dissimilarity between the predicted probabilities and the actual binary labels (0 for benign, 1 for malignant), penalizing incorrect predictions more heavily.

- *Metrics*: Metrics are used to evaluate the model's performance during training and on the validation and test sets. Accuracy, which represents the proportion of correctly classified images, is used as the primary metric in this study. Other metrics, such as precision, recall, and F1-score, could also be used to gain a more comprehensive understanding of the model's performance, particularly its ability to correctly identify malignant cases.

By training the model with the *rmsprop* optimizer, the binary *cross-entropy* loss function, and monitoring *accuracy* as a metric over 25 epochs, the study aims to develop a CNN model that can effectively learn to distinguish between benign and malignant mammograms.
The choice of optimizer, loss function, and the number of epochs plays a crucial role in determining the model's final performance and its ability to generalize to unseen data.

III. Results

The evaluation of the CNN model's performance in classifying mammogram images as benign or malignant involves multiple metrics to provide a comprehensive assessment of its capabilities, as shown by *Table 1*:

- *Accuracy*: Accuracy is the most intuitive metric, representing the proportion of correctly classified images out of the total number of images in the test set. It provides a general overview of the model's performance, but it can be misleading if the dataset is imbalanced, meaning one class has significantly more samples than the other. In the sources, the model achieves a test accuracy of approximately 90%, meaning it correctly classifies around 90% of the test images.

- *Loss*: The loss function, in this case, binary cross-entropy, quantifies the errors made by the model during training and evaluation. Lower loss values indicate better performance, as it means the model's predictions are closer to the actual labels. The sources report a test loss of about 0.57, which is a moderate value, suggesting there is room for potential improvement.

- *Precision*: Precision measures the proportion of correctly classified positive samples (malignant cases) out of all samples predicted as positive. It answers the question: "Of all the mammograms the model classified as malignant, how many were actually malignant?". A high precision score indicates a low rate of false positives, meaning the model is not frequently misclassifying benign cases as malignant.

- *Recall* (Sensitivity): Recall measures the proportion of correctly classified positive samples (malignant cases) out of all actual positive samples. It answers the question: "Of all the actually malignant mammograms, how many did the model correctly identify?". High recall is particularly critical in medical diagnosis, as it means the model is effectively detecting true positive cases. A low recall would mean the model is missing a significant number of malignant cases, which could have severe consequences. The sources show a recall of 0.97 for malignant cases, indicating the model is highly sensitive in identifying these cases.

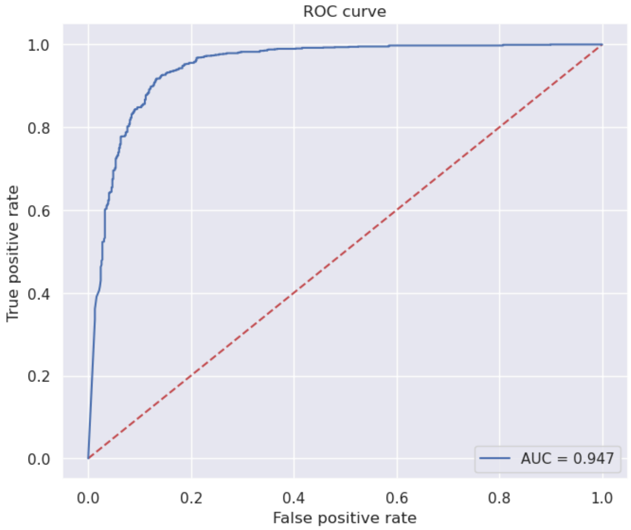
- *F1-score*: The harmonic mean of precision and recall. It provides a balanced measure that considers both false positives and false negatives. A high F1-score indicates a good trade-off between precision and recall. The F1-score for malignant cases in the sources is above 0.93, suggesting a strong performance in identifying malignant tumors.

Table 1: The classification performance metrics for a model distinguishing between benign (Class 0) and malignant (Class 1) samples

	Precision	Recall	F1-score	Support
Benign	0.93	0.76	0.84	617
Malignant	0.89	0.97	0.93	1239
Accuracy			0.90	1856
Macro Avg	0.91	0.87	0.88	1856
Weighted Avg	0.91	0.90	0.90	1856

- *Area Under the Receiver Operating Characteristic Curve* (AUC): The AUC is a metric that evaluates the model's ability to discriminate between the two classes (benign and malignant) across various probability thresholds [13]. An AUC of 1.0 represents perfect discrimination, while an AUC of 0.5 suggests the model is no better than random chance. The AUC value of 0.947 reported in the sources indicates the model has excellent discriminatory power. Refer to *Figure* 2.

Figure 2: This chart shows the Receiver Operating Characteristic (ROC) curve for a binary classifier. The x-axis represents the false positive rate, while the y-axis represents the true positive rate. The diagonal red line marks the performance of a random classifier. The blue curve lies well above the diagonal, indicating strong discriminative ability, with an area under the curve (AUC) of 0.947.



These metrics provide a multifaceted view of the model's performance. While accuracy gives a general idea of correct classifications, loss reflects the magnitude of prediction errors. Precision and recall assess the model's ability to correctly classify positive cases while minimizing false positives and false negatives, respectively. The F1-score balances these two aspects, and the AUC evaluates the model's overall discriminatory power. By considering all these metrics, the study can confidently assess the model's effectiveness in identifying malignant breast cancer from mammograms.

The trained CNN model achieved a test accuracy of 90.24%, demonstrating its effectiveness in classifying mammogram images. The high recall of 0.97 for malignant cases is particularly noteworthy, as it underscores the model's sensitivity in detecting these crucial cases, indicating a strong ability to correctly identify both positive and negative cases.
Further validation of the model's performance is provided by the AUC (Area Under the Receiver Operating Characteristic Curve), a metric that evaluates the model's discriminatory power across various probability thresholds [13]. The AUC value of 0.947, close to the ideal value of 1.0, reinforces the model's exceptional ability to distinguish between benign and malignant cases.

Analysis of the confusion matrix (*Figure* 3) revealed that the model performs well in identifying benign cases but has some misclassifications for malignant cases. This highlights the importance of further refinement to minimize false negatives. Visualization of predicted probabilities and actual labels demonstrated the model's confidence in its predictions.

Figure 3: This confusion matrix visualizes the results of a breast cancer classifier distinguishing between malignant and benign tumors. The rows represent actual labels, and the columns represent predicted labels. Here, the model correctly identifies 469 malignant tumors (top-left) and 1,206 benign tumors (bottom-right), while misclassifying 148 malignant tumors as benign (top-right) and 33 benign tumors as malignant (bottom-left)

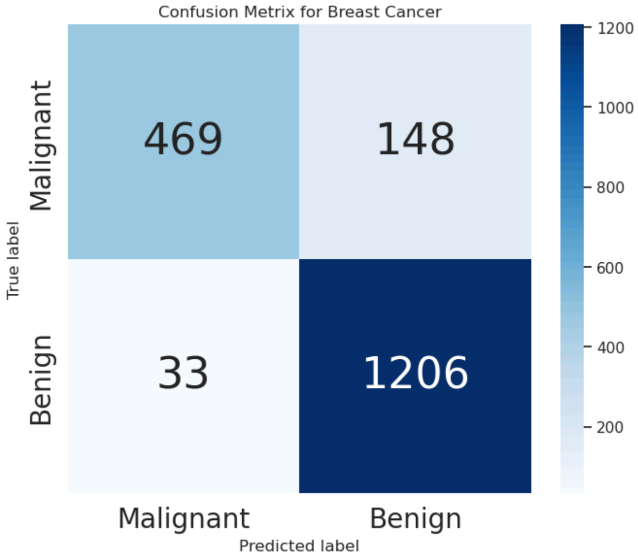
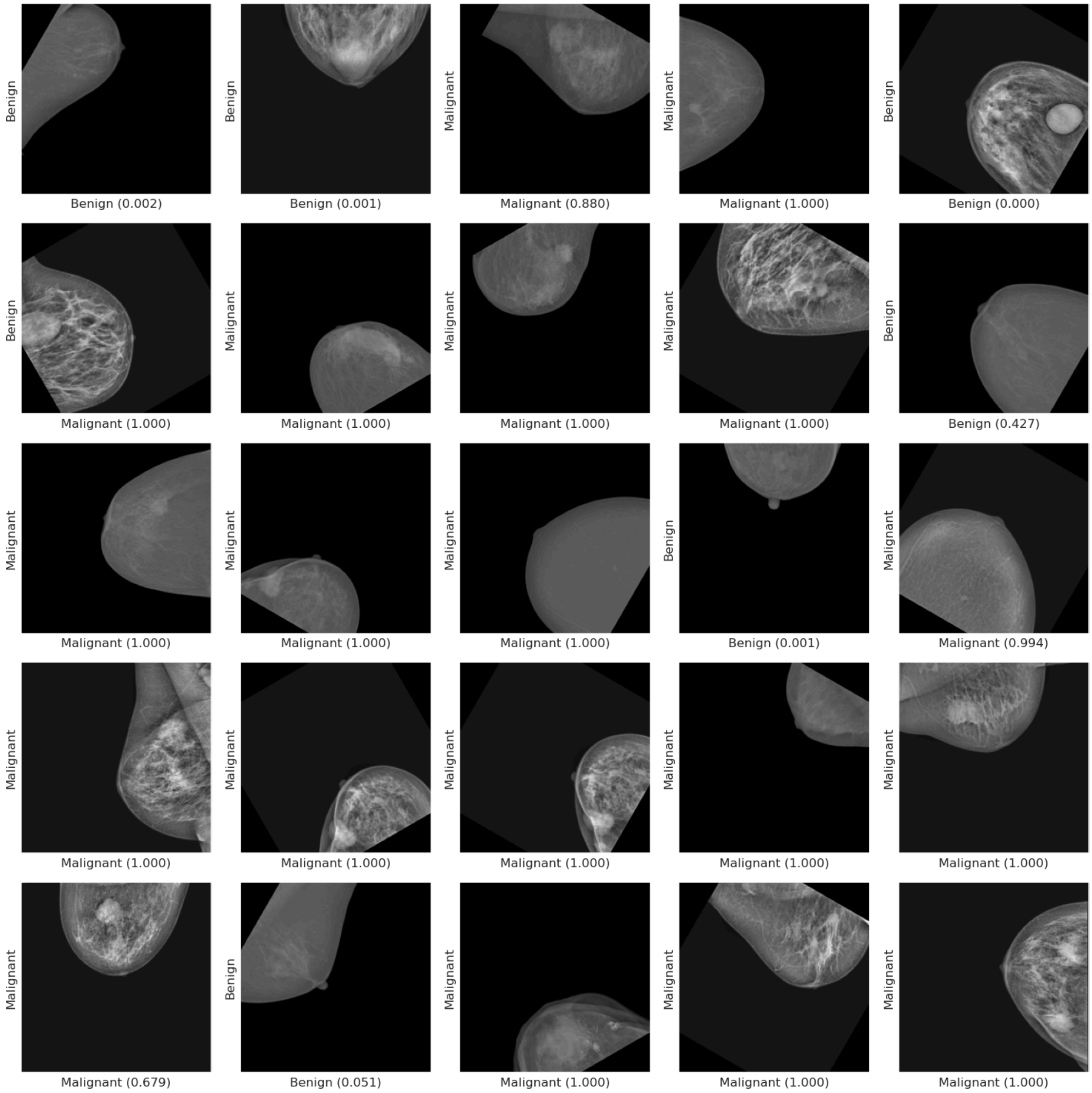


Figure 4 shows a set of mammogram patches classified by our convolutional neural network model. Each patch is labeled as “Benign” or “Malignant,” followed by the model’s predicted probability (in parentheses). These probabilities range from near 0 to 1, reflecting varying levels of confidence. Most malignant lesions have high confidence scores close to 1.0, while benign samples often receive predictions near 0.0. This visualization highlights the model’s capacity to discriminate between benign and malignant breast tissue, providing a clear illustration of how automated classification outputs can potentially aid radiologists in clinical decision-making. However, certain borderline cases suggest that careful expert review remains vital in real-world diagnostic workflows

Figure 4: A grid of mammogram patches, each labeled benign or malignant alongside the model’s confidence score.



IV. Discussion

The findings from this study clearly underscore the efficacy of deep learning models, particularly Convolutional Neural Networks (CNNs), in classifying mammogram images for breast cancer detection. By processing large volumes of image data, CNNs are able to learn intricate features and patterns—such as subtle textural or morphological changes—that may be indicative of malignancies. These capabilities enable the model to deliver high accuracy, precision, recall, and AUC scores, thereby offering a compelling evidence base for the utility of this approach in clinical settings.

One of the most significant advantages of applying CNNs to mammogram analysis lies in the model's ability to uncover details that can elude even experienced radiologists. For instance, faint calcifications, irregular masses, or minuscule architectural distortions could be overlooked during a rapid human review. In contrast, the CNN can systematically analyze every region of the image, highlighting potential areas of concern. This ability to detect subtle cues has profound implications for patient outcomes: earlier and more accurate diagnoses can facilitate timely interventions, thereby potentially improving survival rates and reducing the burden of invasive treatments.

Despite these promising results, it is crucial to emphasize that the CNN is intended as a complementary tool rather than a replacement for human expertise. Radiologists bring indispensable clinical judgment, contextual knowledge of patient history, and experience in interpreting a wide array of imaging modalities. By working in tandem with CNNs, radiologists could benefit from automated alerts or second-opinion assessments. This synergy may lead to enhanced diagnostic confidence, more efficient workflows, and reduced rates of missed diagnoses or false positives.

However, the current study has limitations that must be addressed in future research. First, the dataset used here, while demonstrating strong predictive performance, may not fully capture the diversity of real-world populations. Breast composition, imaging protocols, and patient demographics can vary widely, influencing both the appearance of breast tissue in mammograms and the prevalence of different types of lesions. Therefore, additional research on larger and more heterogeneous datasets is imperative for robust model generalization and reliability.

Moreover, incorporating a broader range of clinical data could substantially enrich the predictive power of the CNN. For example, integrating patient history, genetic predispositions, hormonal factors, and lifestyle risk factors could enable more personalized risk assessments. By combining image-based features with structured clinical data, researchers may further refine the model, potentially capturing correlations that remain hidden when only mammographic images are examined.

Looking forward, continued advancements in deep learning architectures and training techniques may further boost diagnostic accuracy and computational efficiency. Transfer learning from other medical imaging tasks, data augmentation strategies, and interpretability methods (such as Grad-CAM or saliency maps [14]) could not only improve performance but also increase transparency in the decision-making process. This transparency is vital for fostering trust among clinicians, patients, and regulatory bodies.

In conclusion, this study demonstrates the potential of CNNs to significantly enhance breast cancer detection through the automated classification of mammogram images. While the reported performance metrics—high accuracy, precision, recall, and AUC—are encouraging, caution is warranted until further large-scale validation is performed. Ongoing efforts should seek to address issues of dataset diversity, integration with clinical data, and model interpretability. Nonetheless, the work represents a crucial step forward in leveraging artificial intelligence for improved breast cancer screening, ultimately offering a promising pathway to better patient care and outcomes.

References

[1] Elmore, J. G., Wells, C. K., Lee, C. H., Howard, D. H., & Feinstein, A. R. (1994). Variability in radiologists' interpretations of mammograms. The New England journal of medicine, 331(22), 1493–1499. https://doi.org/10.1056/NEJM199412013312206

[2] Wang L. (2024). Mammography with deep learning for breast cancer detection. Frontiers in oncology, 14, 1281922. https://doi.org/10.3389/fonc.2024.1281922

[3] Sarvamangala, D. R., & Kulkarni, R. V. (2022). Convolutional neural networks in medical image understanding: a survey. Evolutionary intelligence, 15(1), 1–22. https://doi.org/10.1007/s12065-020-00540-3

[4] Essamadi, O. (2025, January). Augmented INbreast Dataset, Version 1. Retrieved January 20, 2025 from https://www.kaggle.com/datasets/eoussama/breast-cancer-mammograms.

[5] Moreira, I. C., Amaral, I., Domingues, I., Cardoso, A., Cardoso, M. J., & Cardoso, J. S. (2012). INbreast: toward a full-field digital mammographic database. Academic radiology, 19(2), 236–248. https://doi.org/10.1016/j.acra.2011.09.014

[6] Huang, M. L., & Lin, T. Y. (2020). Dataset of breast mammography images with masses. Data in brief, 31, 105928. https://doi.org/10.1016/j.dib.2020.105928

[7] Mohd Khuzi, A., Besar, R., Wan Zaki, W., & Ahmad, N. (2009). Identification of masses in digital mammogram using gray level co-occurrence matrices. Biomedical imaging and intervention journal, 5(3), e17. https://doi.org/10.2349/biij.5.3.e17

[8] Kulathunga, N., Ranasinghe, N. R., Vrinceanu, D., Kinsman, Z., Huang, L., & Wang, Y. (2020). Effects of the Nonlinearity in Activation Functions on the Performance of Deep Learning Models. ArXiv.org. https://arxiv.org/abs/2010.07359v1

[9] Gholamalinezhad, H., & Khosravi, H. (2020). Pooling Methods in Deep Neural Networks, a Review. ArXiv:2009.07485 [Cs]. https://arxiv.org/abs/2009.07485

[10] Park, S., & Kwak, N. (2017). Analysis on the dropout effect in convolutional neural networks. In Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13 (pp. 189-204). Springer International Publishing.

[11] Vijendra Babu, D., Karthikeyan, C., Shreya, & Kumar, A. (2020). Performance Analysis of Cost and Accuracy for Whale Swarm and RMSprop Optimizer. IOP Conference Series: Materials Science and Engineering, 993, 012080. https://doi.org/10.1088/1757-899x/993/1/012080

[12] Ruby, U., & Yendapalli, V. (2020). Binary cross entropy with deep learning technique for image classification. Int. J. Adv. Trends Comput. Sci. Eng, 9(10).

[13] Sulam, J., Ben-Ari, R., & Kisilev, P. (2017, September). Maximizing AUC with Deep Learning for Classification of Imbalanced Mammogram Datasets. In VCBM (pp. 131-135)

[14] Nunnari, F., Kadir, M. A., & Sonntag, D. (2021, August). On the overlap between grad-cam saliency maps and explainable visual features in skin cancer images. In International Cross-Domain Conference for Machine Learning and Knowledge Extraction (pp. 241-253). Cham: Springer International Publishing.