Performance Analysis of Convolutional Neural Network Models for Breast Cancer Diagnosis

Anonymous Authors

***Abstract*—Breast cancer remains one of the leading causes of mortality among women. This study introduces a CNN-based model for breast cancer diagnosis and provides a comprehensive performance analysis, comparing it with two state-of-the-art pre-trained models: DenseNet121 and ResNet50. The models were trained on a dataset comprising three classes—normal, benign, and malignant breast cancer images. To enhance the dataset and improve model generalization, data augmentation techniques were applied, increasing the total number of images to over 9,000. The study evaluates the models' effectiveness in a multi-class classification setting, with results showing that the CNN model, ResNet50, and DenseNet121 achieved classification accuracies of 96.20%, 97.74%, and 93.94%, respectively. These findings highlight the proposed CNN model's potential for medical applications, offering an optimal balance between computational efficiency, high accuracy, and minimal false positive and false negative rates.**

***Index Terms*—Breast Cancer, Convolutional Neural Network, DenseNet121, ResNet50, Transfer learning, Mammography images.**

I. Introduction

According to the American Cancer Society (ACS), the advancement in oncology owes credit to global collaborations among diverse medical practitioners and researchers who have made notable strides in areas like anatomy, physiology, chemistry, epidemiology, and related disciplines. Breast cancer stands as one of the most lethal forms of cancer, claiming numerous lives annually, primarily among women [1]. According to the World Health Organization's (WHO) statistics, over 400,000 individuals succumb to breast cancer every year [2]. Halting the progression and severity of breast cancer necessitates early detection and accurate prognosis to ascertain whether the condition is benign or malignant. This approach facilitates swift and effective treatment planning, thereby reducing the mortality rate associated with this disease [3]. This study offers a performance analysis of three CNN-based models, aiming to identify more effective, precise, and advanced approaches for diagnosing breast cancer. The specific contribution of this study is as follows: this study introduces a custom CNN model for classifying breast cancer images into three categories: normal, benign, and malignant. The model is evaluated alongside two state-of-the-art pre-trained architectures, DenseNet121 and ResNet50, through a comprehensive performance analysis. Data augmentation techniques were applied to expand the dataset to over 9,000 images, improving model generalization. The results demonstrate high classification accuracy, with the CNN model achieving 96.20%, ResNet50 97.74%, and DenseNet121 93.94%, highlighting the effectiveness of deep learning in medical diagnostics.

II. Related Work

Numerous investigations have explored the utilization of deep-learning models for breast cancer identification. Patro et al [4] tilized a hybrid convolutional neural network (HCNN) and LeNet-5 to classify breast cancer images. Shahid et al. [5] assessed the merits and limitations of crucial imaging techniques for detecting breast cancer. Huang and Lin [6] employed AlexNet, DenseNet, and ShuffleNet to categorize breast density into benign and cancerous categories. Yadav, R. K. et al, in [7] developed a hybrid Convolutional Neural Network and assessed its performance using the MIAS and Digital Database for Screening Mammography (DDSM) datasets, achieving an accuracy of 98.44% on the MIAS dataset and 98.07% on the DDSM dataset. Mechria et al. [8] leveraged two deep CNN (DCNN) architectures, a shallow DCNN, and Alexnet, to detect breast cancer using 8000 mammography images sourced from the Digital Database of Screening Mammography (DDSM) dataset. The technique achieved a classification accuracy of 89.23%. El and Yassin [9] introduced an algorithm designed to differentiate between malignant and non-malignant breast tumors, achieving an impressive classification accuracy of 94%. In another study, Nawaz et al. [10] developed a CNN-based technique to classify breast cancer into multiple categories utilizing DL techniques, and it produced an accuracy rate of 95.4%. Chouhan et al. [11] developed a breast cancer technique based on emotional learning through hybrid characteristics. This approach, trained on SVM, achieved a ROC-AUC value of 84.6%. In [12] Chakravarthy et al. proposed fusion of hybrid deep features (FHDF) model for breast cancer detection using CBIS-DDSM data set for multiclass classification and achieved an accuracy of 97 70%. In [13], [14], and [15] the authors used classical machine learning techniques for multiclass classification and achieved accuracies of 75.5%, 92.60%, and 97.00% respectively, respectively, on the CBIS-DDSM dataset.

III. Methodology

The three models used in this study were developed using the Keras API of TensorFlow [16], an open-source Python library from which we imported the pre-trained ResNet50 and DenseNet121 networks. The models were implemented on Kaggle with GPU P100. ReLu and Softmax functions were utilized for activation; categorical cross-entropy was used as the loss function; and Adam was the optimizer. We also utilize the Keras tuner [16] to find the optimal parameter for our models.

A. General Flowchart for the designed models

To summarize all the steps followed in our experiment, we have designed a flowchart that contains all the details of our experiment. The flowchart is presented in Figure 1.

B. Description of the Proposed CNN

The custom CNN architecture consists of two convolutional layers and one fully connected layer. The input image dimensions are set at 224×244 pixels. The first layer is a 2D convolutional layer with 64 filters and a 3×3 kernel size. To introduce non-linearity, the model employs the ReLU activation function. Subsequently, the second layer reduces the spatial dimensions, employing a max pooling layer with a 5×5 pool size. The third layer comprises a 2D convolutional layer incorporating 32 filters and a 5×5 kernel size. Following this is a pooling layer and a dropout layer, introduced to randomly set a fraction of input units to zero during each training update. The final layer is a dense layer comprising three units and applies a softmax activation function, generating a probability distribution across the output classes for the input image. The architecture of this model is depicted in Figure 2.

C. Description of the modified ResNet50 and DenseNet121

The two models we have designed here are based on transfer learning and pre-trained models. We use the pre-trained models ResNet50 and DenseNet121 to leverage their capabilities for extracting information from images. DenseNet121 and ResNet50 architectures consist of 121 and 50 layers, respectively. The output obtained from both DenseNet121 and ResNet50 was passed through a sequence of layers, including one convolutional layer with 32 filters for both ResNet50 and DenseNet121. The models also consist of one fully connected layer, one dropout layer, and one output layer. In particular, the fully connected layer associated with DenseNet121 comprises 512 units, whereas the corresponding layer in ResNet50 comprises 128 units. Utilizing dropout regularization, the dropout rates applied were 0.4 for DenseNet121 and 0.2 for ResNet50. It is essential to note that only the added layers

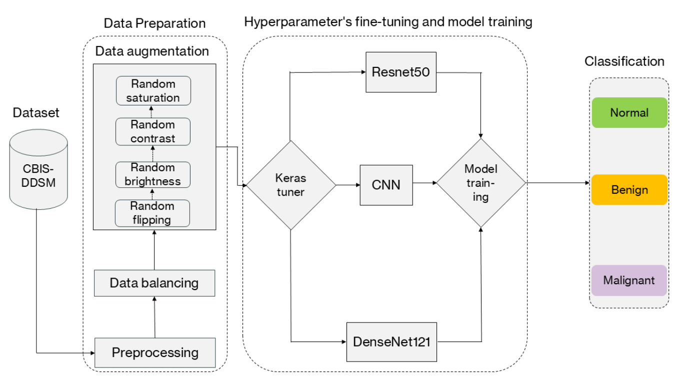


Fig. 1: General Flowchart for the designed models.

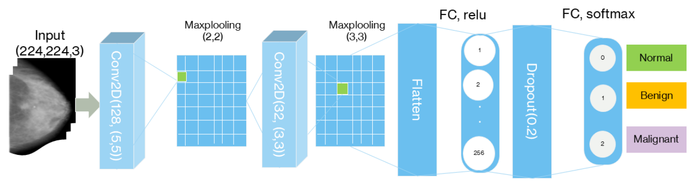


Fig. 2: Custom CNN model used in this study.

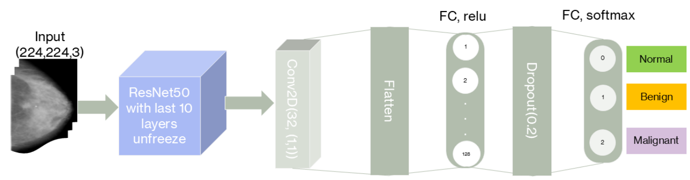


Fig. 3: ResNet50 model used in this study.

were fine-tuned, and the parameters utilized for training these three models were chosen through experimental selection, as described in Table I. Figure 4 and 3 show the architectures of those models pre-trained models.

D. Dataset Description

The dataset utilized for this experimental study is the Curated Breast Imaging Subset of the Digital Database for Screening Mammography (CBIS-DDSM) [17]. Specifically, we focused on a subset of the dataset containing mammographic images of breast masses. This subset consists of 1,644 images categorized into three classes: Normal (171 images), Benign (707 images), and Malignant (766 images). Each image in the dataset has three variations: Full Image, Cropped Image, and Region of Interest (ROI). The dataset contains 38 columns detailing patient-related information, but only the image path and pathology type were used in this study. To ensure accuracy in image retrieval, we utilized the DICOM dataset [17], which provides the correct paths for each image. These verified paths were saved and divided into three subsets for further processing.

a) Data Pre-Processing: The data pre-processing phase consisted of two key stages: data augmentation and dataset balancing.

b) Data Augmentation: To enhance the dataset size and improve the model's ability to generalize, we applied the following augmentation techniques to the normal and benign images.

- Random Brightness Adjustment: The brightness of each image was randomly altered within the range of [-0.3, 0.3].

TABLE I: Parameters used to train models.

Model	Training time (mins)	Drop rate	Number of epoch	Batch size	Learning rate
CNN	19.07	0.4	100	32	0.0001
ResNet50	23,711	0.3	100	32	0.001
DenseNet121	23,358	0.3	100	32	0.0001

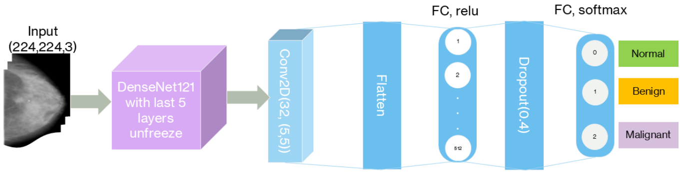


Fig. 4: DenseNet121 model used in this study.

- Random Contrast Adjustment: The contrast of each image was randomly scaled by a factor within the range [0.8, 1.2].
- Random Saturation Adjustment: The saturation of each image was randomly modified within the range [0.8, 1.2].

After the above step, we augmented each class with a number of images equal to two times the number of images in the dataset. These augmentations helped introduce variability in the dataset while preserving essential features crucial for classification. Figure 5 below shows the data augmentation process.

c) Dataset Balancing: The original dataset exhibited class imbalance, particularly in the normal category, which contained significantly fewer samples compared to the benign and malignant categories. To address this issue, we reduced the number of augmented images using formula (1):

$$\text{NA}_{c_i} = \max\{Nc_1, Nc_2, Nc_3\} - N_{c_i}; \tag{1}$$

Where:

- NA_{c_i} is the number of augmented images for class c_i;
- Nc_1, Nc_2, Nc_3 represent the image counts for the three classes;
- N_{c_i} is the current number of images in class c_i.

This strategy ensured a balanced dataset, preventing model bias toward overrepresented classes.

d) Dataset Statistics: Table II shows a summary of the image distribution before and after augmentation. We improved the dataset quality by applying these pre-processing steps, ensuring better model performance and robustness during classification.

IV. Result and Discussion

This section presents the performance analysis of the proposed CNN model, ResNet50, and DenseNet121.

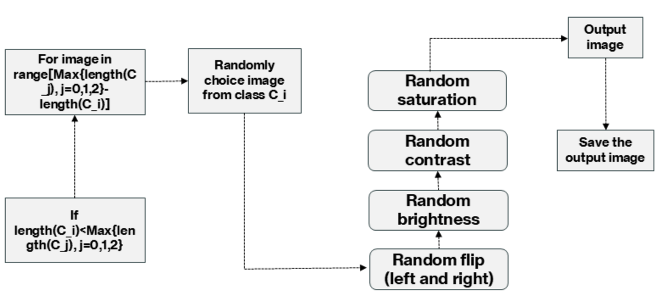


Fig. 5: Data augmentation process.

TABLE II: Dataset information.

Class	Before data augmentation	Number of images after balancing	After data augmentation
Normal	107	664	3,388
Benign	591	664	3,388
Malignant	664	664	3,388
Total	1,644	1,992	10,160

A. General Performance Analysis and Interpretation

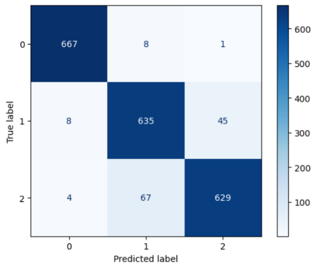
The experimental results demonstrate that all three models ResNet50, DenseNet121, and CNN achieved high accuracy in breast cancer classification using the CBIS-DDSM dataset. Among them, ResNet50 exhibited the highest accuracy (97.81%), followed by CNN (96.97%) and DenseNet121 (96.82%). Precision and recall values also aligned with accuracy scores, indicating consistent model performance across various metrics.

From a clinical perspective, these results are highly significant. Breast cancer diagnosis using mammography remains a critical screening method, and automated deep learning-based approaches can enhance diagnostic efficiency. High recall values across all models suggest that they effectively identify malignant cases, reducing the likelihood of false negatives, which is crucial in medical diagnostics. Moreover, high precision indicates fewer false positives, minimizing unnecessary biopsies and psychological distress for patients.

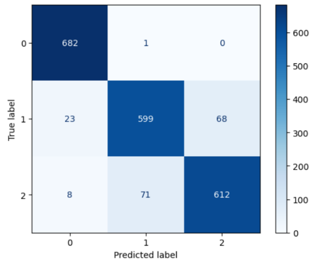
B. Confusion Matrix Analysis

The confusion matrix is presented in Figure 6. The confusion matrices reveal that all models performed well in distinguishing normal cases, with very few misclassifications. ResNet50 misclassified only 3 normal cases as benign and 3 as malignant, while DenseNet121 and CNN each misclassified 2 normal cases as benign and 2 as malignant. This demonstrates strong specificity in identifying normal cases across all models. However, a noticeable challenge was the differentiation between benign and malignant cases. ResNet50 misclassified 48 benign cases as malignant and 61 malignant cases as benign, while DenseNet121 and CNN exhibited similar trends, with DenseNet121 having the highest number of benign cases incorrectly classified as malignant (74).

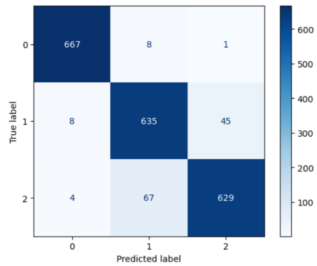
Among benign cases, the models varied in their ability to distinguish them correctly. ResNet50 misclassified 7 benign cases as normal, DenseNet121 misclassified 14, and CNN misclassified 10, indicating that DenseNet121 had the highest tendency to confuse benign cases with normal cases. Additionally, the models struggled with false negatives in malignant cases, with ResNet50 misclassifying 61 malignant cases as benign, DenseNet121 misclassifying 57, and CNN misclassifying 55. These misclassifications are clinically significant, as failing to identify malignant cases correctly could lead to delayed diagnosis and treatment. Overall, ResNet50 demonstrated the best performance among the three models, correctly classifying the highest number of malignant cases while maintaining strong accuracy across all categories. However, the tendency of all models to misclassify benign as malignant cases suggests the



((a)) CNN Model.



((b)) DenseNet121 Model.



((c)) ResNet50 Model.

Fig. 6: Confusion matrix for ResNet50, DenseNet121, and CNN.

need for further refinement, possibly through advanced feature extraction techniques or hybrid deep learning approaches. These findings highlight the importance of improving benign-malignant differentiation to enhance diagnostic reliability in breast cancer detection.

a) Comparative Performance Analysis of the Three Models: While all three models demonstrated strong classification performance III, ResNet50 emerged as the best-performing model, achieving the highest accuracy, precision, and recall (97.74%). The confusion matrices further highlight the model's superiority, as it had the lowest misclassification rates compared to DenseNet121 and CNN. DenseNet121, despite being a robust model, exhibited slightly lower performance, particularly in distinguishing between benign and malignant cases, as indicated by its higher number of misclassified malignant cases (74 compared to 48 in ResNet50). Similarly, CNN performed slightly better than DenseNet121 but still trailed ResNet50 in overall classification accuracy. The differences in model performance can be attributed to architectural variations. ResNet50's deep residual connections facilitate better gradient flow and mitigate the vanishing gradient problem, enabling more effective learning. In contrast, while DenseNet121's dense connectivity structure enhances feature propagation, it might lead to redundant feature learning, slightly impacting performance. CNN, being a shallower network, likely lacks the depth required to capture complex mammographic features as effectively as the other two models. As shown in Table III, the proposed CNN model achieved a classification accuracy of 96.30%, indicating that 96.30% of the test dataset images were correctly classified. Additionally, the CNN model attained precision and sensitivity scores of 96.30% for both metrics. The sensitivity score of 96.30% indicates the model's capability to correctly identify a substantial number of malignant images within the dataset. The precision of 96.30% reflects the quality of the model's positive predictions. Moreover, examining the confusion matrix in Fig. 6(c) reveals that the CNN model accurately classified 98.67% of normal images, 92.61% of benign images, and 89.86% of malignant images in the test dataset.

C. Comparison with Related Studies

To further contextualize our findings, we compared the best-performing model (ResNet50) with other recent studies that utilized the CBIS-DDSM dataset for breast cancer classification. As shown in Table IV, our best model outperformed existing approaches, including the FHDF method (97.7%). Compared to traditional machine learning approaches, such as Apriori Dynamic Selection with SVM, which achieved 75.81%, Optimized kernel ELM architecture with 92.6%, and Support vector machine (SVM) with boundary descriptor feature inputs with 97.00% in three different studies, ResNet50 demonstrated a substantial improvement.

The superior performance of ResNet50 in this study can be attributed to the model's advanced residual learning capabilities, effective feature extraction, and robust training strategies, including data augmentation and class balancing. These findings further validate the effectiveness of deep learning architectures in medical image classification and highlight the potential of ResNet50 as a state-of-the-art model for mammographic image analysis.

D. Clinical Implications

The results indicate that ResNet50 is the most effective model for breast cancer classification, outperforming the other models in accuracy, precision, and recall. However, the confusion matrices reveal a persistent challenge: distinguishing benign from malignant cases. Misclassifying malignant tumors as benign (false negatives) is particularly concerning, as it can lead to missed diagnoses and delayed treatment. This highlights the need for further refinement in deep learning models to enhance their ability to differentiate between these critical categories. Despite this challenge, the models demonstrated high specificity in identifying normal cases, with very few normal images misclassified. This strong performance suggests that deep learning models can effectively rule out non-cancerous cases, reducing unnecessary follow-up tests. However, the tendency to confuse benign and malignant cases suggests that additional strategies, such as incorporating advanced preprocessing techniques or using an ensemble approach, may further improve classification performance. Addressing these misclassifications is essential to ensure that AI-based

TABLE III: Performances analysis of the proposed models.

Models	Precision (%)	Sensitivity (%)	Accuracy(%)
CNN	96.30	96.30	96.30
ResNet50	97.81	97.74	97.74
DenseNet121	94.25	93.87	93.94

TABLE IV: Comparison with previous studies.

Models	Accuracy (%)	Reference
ResNet50	97.74	This Study
Fusion of Hybrid Deep Features (FHDF)	97.70	Chakravarthy et al. [12]
Apriori Dynamic Selection with SVM	75.81	Khaoula et al. [13]
Optimized kernel extreme learning machine	92.60	Figlu et al. [14]
Support vector machine (SVM) with boundary descriptor feature inputs	97.00	Safdarian and Hedyezadeh [15]

diagnostic tools provide reliable support in clinical settings. The implications of these findings underscore the potential of deep learning models in assisting radiologists by providing a second opinion. Given the high performance of these models, particularly ResNet50, integrating such AI-based tools into clinical workflows can improve diagnostic accuracy, expedite decision-making, and ultimately contribute to better patient outcomes. AI-driven breast cancer classification systems could help alleviate the workload of radiologists while enhancing the consistency and reliability of diagnoses. In summary, while all three models exhibit strong predictive power, ResNet50 stands out as the most reliable for breast cancer classification. Nonetheless, improvements in handling benign-malignant misclassification remain an important area for future research. By refining these models through enhanced training data, feature extraction methods, and hybrid approaches, deep learning can continue to evolve as a valuable asset in the early detection and diagnosis of breast cancer.

V. Conclusion

This study explored the performance of three deep learning models ResNet50, DenseNet121, and a custom CNN in classifying breast cancer using the CBIS-DDSM dataset. The results demonstrated that all models achieved high accuracy, precision, and recall, highlighting the effectiveness of deep learning in medical image analysis. Among the three models, ResNet50 outperformed the others, achieving the highest accuracy of 97.74%, making it the most reliable model for breast cancer classification in this study. However, the confusion matrix analysis revealed challenges in distinguishing benign from malignant cases, which is a critical aspect of breast cancer diagnosis. The high specificity observed in classifying normal cases suggests that deep learning models can be effective in reducing unnecessary follow-up examinations. However, the misclassification of malignant tumors as benign (false negatives) remains a significant concern, as it may delay treatment and negatively impact patient outcomes. These findings highlight the need for further refinement in deep learning models, particularly in handling borderline cases where benign and malignant features overlap. Strategies such as enhanced data augmentation, improved feature extraction, and hybrid model approaches could help address this issue and further improve classification accuracy. The implications of this study extend beyond model performance metrics. The integration of AI-based diagnostic tools into clinical workflows has the potential to enhance early detection efforts, support radiologists in decision-making, and ultimately improve patient care. By reducing diagnostic variability and providing consistent assessments, deep learning models can serve as valuable assistive tools in breast cancer screening. However, their deployment in real-world clinical settings requires rigorous validation, interpretability, and integration with existing radiology systems to ensure reliability and clinical trust. Future work should focus on improving the generalizability of these models by incorporating larger and more diverse datasets, employing ensemble learning techniques, and exploring explainability methods to make AI-driven predictions more transparent for medical professionals. Additionally, investigating the integration of multimodal data, such as histopathology images and patient demographics, could further enhance diagnostic accuracy. Despite the challenges, this study reinforces the potential of deep learning in breast cancer classification and provides a foundation for further advancements in AI-assisted medical imaging.

References

[1] D. Koboldt, R. Fulton, M. McLellan, H. Schmidt, J. Kalicki-Veizer, J. McMichael, et al. Comprehensive molecular portraits of human breast tumors. Nature. 2012;490:61–70. doi: 10.1038/nature11412

[2] Z. Anastasiadi, G. D. Lianos, E. Ignatiadou, H. V Harissis, M. Mitsis. (2017). Breast cancer in young women: an overview. Updates in surgery, 69:313–317.

[3] Giaquinto, A. N., Sung, H., Newman, L. A., Freedman, R. A., Smith, R. A., Star, J., ... & Siegel, R. L. (2024). Breast cancer statistics 2024. CA: a cancer journal for clinicians, 74(6), 477-495.

[4] Patro, P., Fathima, S. H., Harikishore, R., & Sahu, A. K. (2024). Breast cancer image classification by using HCNN and LeNet5. Discover Sustainability, 5(1), 491.

[5] S. M. Shah, R. A. Khan, S. Arif, and U. Sajid, "Artificial intelligence for breast cancer analysis: Trends & directions." Computers in Biology and Medicine, 142:105221, March 2022.

[6] M. Huang and T. Lin. "Considering breast density for the classification of benign and malignant mammograms." Biomedical Signal Processing and Control, 67:102564, 2021.

[7] Yadav, R. K., Singh, P., & Kashtriya, P. (2023). Diagnosis of breast cancer using machine learning techniques-a survey. Procedia Computer Science, 218, 1434-1443.

[8] V. Nemade, S. Pathak, and K. D. Ashutosh. "A systematic literature review of breast cancer diagnosis using machine intelligence techniques." Archives of Computational Methods in Engineering, 29(6):4401–4430, 2022.

[9] E. M. E. Houby and N. I. Yassin. "Malignant and nonmalignant classification of breast lesions in mammograms using convolutional neural networks." Biomedical Signal Processing and Control, 70:102954, 2021.

[10] M. Nawaz, A. A Sewissy, and T. H. A Soliman." Multi-class breast cancer classification using deep learning convolutional neural network." Int. J. Adv. Comput. Sci. Appl, 9(6):316–332, 2018.

[11] N. Chouhan, A. Khan, J. Z. Shah, M. Hussnain, and M. W. Khan. "Deep convolutional neural network and emotional learning based breast cancer detection using digital mammography." Computers in Biology and Medicine, 132:104318, 2021.

[12] Chakravarthy, Sannasi, et al. "Multi-class breast cancer classification using cnn features hybridization." International Journal of Computational Intelligence Systems 17.1 (2024): 191

[13] Soulami, Khaoula Belhaj, Naima Kaabouch, and Mohamed Nabil Saidi. "Breast cancer: Three-class masses classification in mammograms using Apriori dynamic selection." Concurrency and Computation: Practice and Experience 34.24 (2022).

[14] Mohanty, Figlu, Suvendu Rup, and Bodhisattva Dash. "Automated diagnosis of breast cancer using parameter optimized kernel extreme learning machine." Biomedical Signal Processing and Control 62 (2020): 102108.

[15] Jiang, Zi-Rong, et al. ”Identification of novel cuproptosis-related lncRNA signatures to predict the prognosis and immune microenvironment of breast cancer patients.” Frontiers in Oncology 12 (2022): 988680.

[16] T. O’Malley, E. Bursztein, J. Long, F. Chollet, H. Jin, I. Luca, et al. Kerastuner, 2019.

[17] R. S. Sawyer-Lee, F. Gimenez, A. Hoogi, and D. L. Rubin. “Curated breast imaging subset of digital database for screening mammography (CBIS-DDSM).” The Cancer Imaging Archive, 2016. https://doi.org/10.7937/K9/TCIA.2016.7O02S9CY.

[18] F. F. Ting, Y. J. Tan, and K. S. Sim. “Convolutional neural network improvement for breast cancer classification.” Expert Systems with Applications, 120:103–115, 2019.