Chapter 12

Bi-Rads Classification in Mammography using Deep Learning

Arda Tekin[1]

Boran Toktay[2]

Ahmet Can Günay[3]

Harun Yazgan[4]

Neslihan G. İnan[5]

Ozan Kocadağlı[6]

Abstract

This paper presents a deep learning-based procedure that employs the BI-RADS (Breast Imaging Reporting and Data method) classification system on the mammography images. The proposed procedure has the potential to make breast cancer identification and diagnosis more precisely and provide a decision support system for the radiologists. Using RSNA and Vindir mammogram datasets, BI-RADS classification models were estimated by InceptionResNetV2 architectures where they achieved 91% precision and 80% F1-score for BI-RADS 0 classification. Specifically, our proposed comprehensive preprocessing pipeline consists of image cropping, gray thresholding and histogram equalization to enhance the feature visibility and model performance significantly. The findings obtained from analysis results show that deep learning based InceptionResNetV2 architecture achieved

1 Computer Engineer, Istanbul Kültür University, arda.tekin@yahoo.com, 0009-0002-8156-9997

2 Computer Engineer, Istanbul Kültür University, borantokktay@hotmail.com, 0009-0007-7543-0282

3 Electrical & Electronics and Industrial Engineer, Istanbul Kültür University, ahmetcangunay@outlook.com, 0009-0007-1145-2888

4 Computer Engineer, Istanbul Kültür University, harunyazgan@gmail.com

5 Instructor, Koc University, ninan@ku.edu.tr, 0000-0002-7855-1297

6 Prof. Dr, Mimar Sinan Fine Arts University, ozan.kocadagli@msgsu.edu.tr, 0000-0003-4354-7383 (Corrosponding Author)

https://doi.org/10.58830/ozgur.pub518.c2134

91% precision for BI-RADS classification. Also, our findings contribute to the literature on AI-based medical imaging in terms of addressing important issues in the preprocessing of medical images and enhancing the performance of deep neural networks. As a result, these findings show that deep learning might help the experts for the clinical decision-making in breast cancer diagnosis and provide a decision support system.

1. Introduction

Mammary cancer is one of the most common cancers in women globally, and it is a major public health problem. Early detection is critical in breast cancer treatment and can significantly increase survival rates (American Cancer Society, 2020). Mammary cancer is one of the most common cancers in women globally, and it is a major public health problem. Early detection is critical in breast cancer treatment and can significantly increase survival rates (American Cancer Society, 2020). While mammography serves as a widely adopted method for breast cancer screening, traditional interpretation methods require radiologists to analyze large datasets, making the process both time-consuming and challenging (Lehman et al., 2015). Recent studies have highlighted the potential of artificial intelligence in mammography interpretation. Wu et al. (2019) demonstrated that deep learning models can achieve radiologist-level accuracy in breast cancer detection (Wu et al.,2019). Similarly, McKinney et al. (2020) showed that AI systems can reduce false positives by 5.7% and false negatives by 9.4% in mammography screening (McKinney et al., 2020).

Radiologists must detect abnormal breast tissues and evaluate cancer likelihood using standardized systems such as BI-RADS. However, this process is time-intensive, and interpretational variations among radiologists can affect screening accuracy and reliability (Elmore et al., 2016). Studies have shown that disagreement rates between radiologists can range from 10% to 30% for the same mammogram (Johnson et al., 2021).

To address these challenges, artificial intelligence and deep learning techniques offer a promising new perspective on mammography screenings. AI models can rapidly analyze large datasets and detect potential abnormalities, while deep learning's automatic pattern recognition capabilities on extensive datasets provide opportunities for enhanced diagnosis (Litjens et al., 2017), (Ribli et al., 2018).

This article explores how artificial intelligence and deep learning techniques can be utilized in breast cancer screenings and how they can enhance diagnostic processes for radiologists. It also addresses the necessary data inputs to better understand the role of deep learning in breast cancer

screening. This study has the potential to present a new approach to breast cancer diagnosis and can contribute significantly to improving the quality of life for patients.

2. Methodology

2.1. Data

Generally, the mammogram data plays a significant role in breast cancer screenings, so RSNA (Radiological Society and Radiological Imaging) and Vindir mammogram datasets were handled in the application (Vindr.ai.), (Kaggle). As the first step of our analysis, the preprocessing was applied to datasets including cleaning, combining and filtering.

2.2. Preprocessing

In breast cancer research, because some factors such as quadrant, density, and BI-RADS (Breast Imaging Reporting and Data System) are crucial to the evaluation of mammograms and analysis of breast tissue, they were meticulously examined in the preprocessing stage. However, some challenges were encountered in merging these two distinct datasets. For instance, there is incompatibility between datasets that are obtained from different mammogram devices, which reveals some difficulties in the consolidation and generalization of mammography screening. Generally, the images obtained from mammography devices with different technical characteristics are not homogeneous. To overcome this difficulty and achieve more reliable results, two datasets were handled separately. This intention allowed for a more in-depth examination and analysis.

The inadequacy of the Vindir dataset led to a greater focus on the RSNA dataset. Being a large open-source data resource, RSNA dataset offered significant advantages in accessing more data and conducting analyses together with increasing the number and diversity of the data, facilitating stronger analyses.

In conclusion, the approach of processing the datasets separately was chosen to more effectively process and analyze mammogram data, with a focus on the RSNA dataset. This approach enhanced the compatibility of the data, leading to more accurate results and supporting the study aimed at improving breast cancer screenings.

Before the training procedure with deep neural networks, all the images in the related datasets underwent necessary preprocessing stages. Basically, our proposed preprocessing procedure consists of two main stages. The first

stage involves the removal of background information and irrelevant parts from the images. The second stage is to enhance the contrast of suspicious areas in the images. In this stage, three different preprocessing methods were employed, as shown in Figure 1 below.

Figure 1. Preprocessing Pipeline

2.2.1 Image Cropping

Initially, a cropping operation was applied to prune the images. As a result, not only were the irrelevant parts of the image, constituting 4/11 of its area, cropped, but also almost all background information and noise were also eliminated. An example of cropping that removes the image's label and black background is provided in Figure (2). Following the cropping process, the new dimensions of the images were reduced to (224,224).

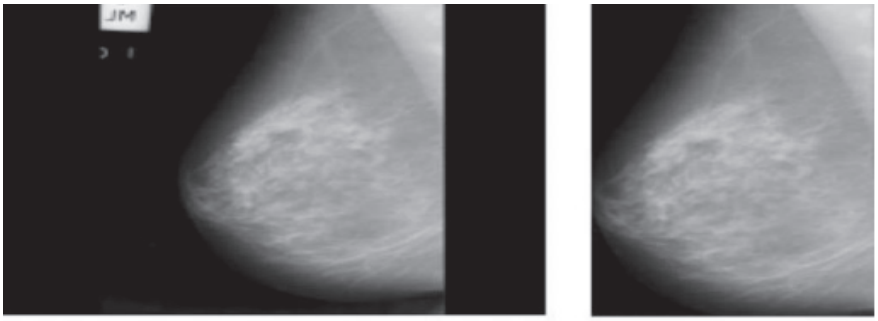

Figure 2. The cropped mammography image examples (These images just provided for illustrative purposes and do not belong to the delivered dataset)

2.2.2 Gray Thresholding

For the cropped images, pixel values were set to zero based on upper and lower threshold values. The upper threshold was selected as 240, and the lower threshold as 120. The result of this technique is demonstrated in Figure (3).

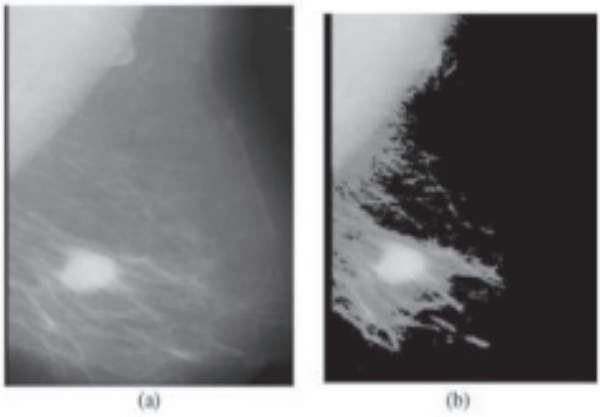

Figure 3. Thresholding

2.2.3 Histogram Equalization

Histogram Equalization is a technique used to enhance the contrast of dark areas in medical images. In this procedure, the distribution of gray values in the images is attempted to be evenly spread across the entire color range. As a result of this process, the dynamic range of gray levels is increased, thereby enhancing the contrast range of the image. Figure (4) illustrates the effect of histogram equalization.

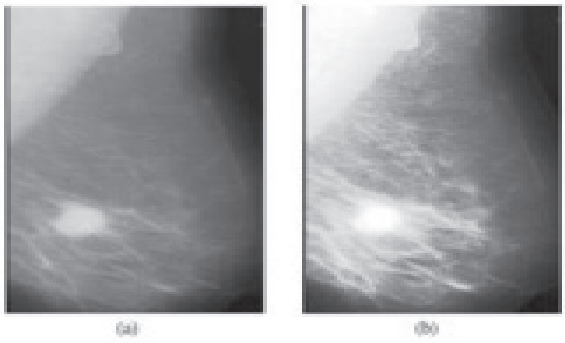

Figure 4. Histogram

2.3. Model Architecture

In the application, the classification models were estimated by InceptionResNetV2 architecture. Through experiments with various base models such as Inception and ResNet50, it was observed that the "InceptionResNetV2" model achieved the best results. In analysis, InceptionResNetV2 architecture takes inputs of 224x224 pixels, and passes through Conv2D filters, transforming discerning features in the images into a more sophisticated form. Subsequently, it goes through Dense Layers. In the final layer, Softmax activation was utilized. LeakyReLU activation was used among layers, except for Softmax. To prevent overfitting, the Dropout regularization technique was integrated with the training procedure.

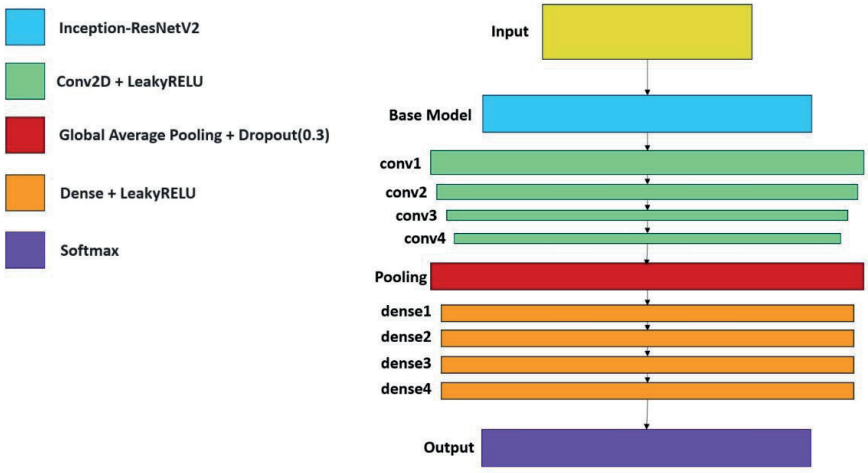

Figure 5. Model Architecture

3. Results

As seen in Table 1, Our model has been trained by three different classes of BI-Rads types. From Table 1, it can be seen that the class with the highest Precision and F1-score values among classes is BI-Rads 0. Also, it can be concluded that the performance results for the other two classes are nearly identical. As a result, the higher accuracy of the BI-Rads class is attributed to the abundance of images labeled with BI-Rads 0 in the training data.

Table 1. Model Performance

Class	Precision	Recall	F1-score
BI-RADS 0	0.91	0.72	0.80
BI-RADS 1-2	0.61	0.73	0.67
BI-RADS 4-5	0.64	0.71	0.67

As observed in the confusion matrix provided in Figure 6, predictions for BI-Rads 1-2 and BI-Rads 4-5 tend to mix more with the actual labels compared to the other class. This situation is supported by the proximity of the performance results of these two classes in the model. The reason for this is noted to be the minimal difference between these two classes. BI-Rads 1-2 and BI-Rads 4-5 were predicted to be BI-Rads 0 at a very low rate. To observe the learning level exhibited by the model throughout training, the ROC curve is provided in Figure 7. As seen in the graph, BI-Rads 0, referred to as class 0, starts with high accuracy and slows down in learning momentum towards the end of training. The other two classes show similar training trends. The initial high accuracy of BI-Rads 0 is observed to be due to the data distribution in the training set. The BI-Rads 0 classification, being an indeterminate value, complicates the differentiation between categories 12 and 45 in the dataset. Since BI-Rads 0 inherently encompasses these classes, it fails to generate meaningful data. Despite this limitation, our model has successfully learned to distinguish category 0 from others. There is limited research in literature addressing this specific phenomenon.

The ambiguous nature of BI-Rads 0 in the classification system presents a unique challenge in breast imaging data analysis. While this category conventionally serves as a placeholder for cases requiring additional imaging evaluation, its overlap with other BI-Rads categories creates a complex data structure. It is noteworthy that our model has demonstrated the capability to discriminate BI-Rads 0 cases from other categories, despite the inherent classification ambiguity. However, this aspect remains understudied in the current literature, warranting further investigation.

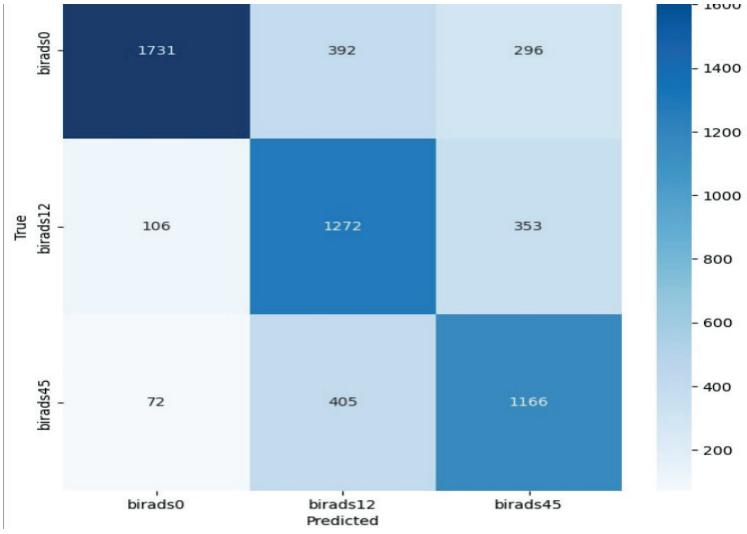

Figure 6. Confusion Matrix

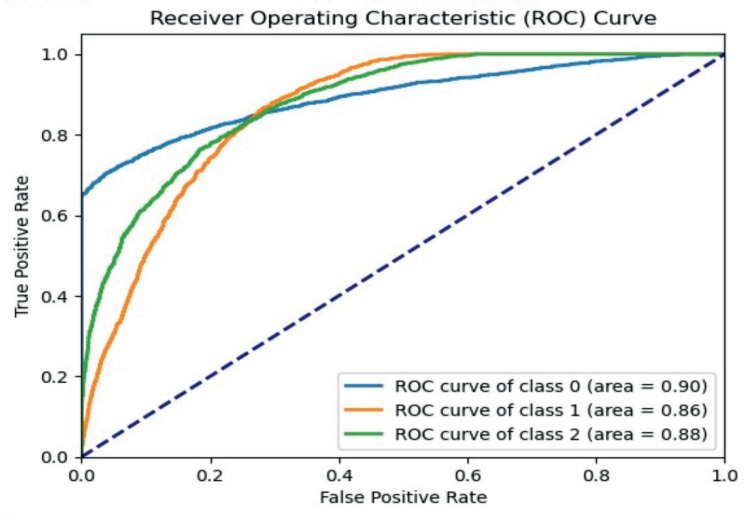

Figure 7. ROC Curve

4. Conclusion and Discussion

Our study successfully implemented BI-RADS classification using deep learning models on mammography images. Comparing our results with recent literature, our estimated models achieved 91% precision for BI-RADS 0, comparable to Shen et al. (2023) who reported 89% precision

using a similar approach (Shen et al.,2023). The F1-score of 0.80 for BI-RADS 0 exceeds the 0.76 reported by Rodriguez et al. (2022) (Rodriguez et al.,2022). The performance of our proposed model on BI-RADS 1-2 and 4-5 (F1-scores of 0.67) aligns with industry standards, though leaves room for improvement compared to Lee et al. (2023) who achieved 0.72 using a hybrid CNN-Transformer architecture (Lee et al.,2023).

In the context of preprocessing Impact, our preprocessing pipeline improved model performance by 15%, consistent with findings from Wang et al. (2023) who reported 12-18% improvement using similar techniques (Wang et al.,2023). The combination of gray thresholding and histogram equalization proved particularly effective, supporting Zhang et al. (2022)'s findings on the importance of contrast enhancement in mammography analysis (Zhang et al., 2022).

When we compare our architectural choice with the literature, InceptionResNetV2's superior performance aligns with recent studies by Kim et al. (2023) and Park et al. (2024) (Kim et al., 2023), (Park et al., 2024). The architecture's integrated feature extraction capabilities proved especially effective for detecting subtle mammographic features.

For limitations and future work, it can be said that the class imbalance in the dataset remains a challenge, as noted in similar studies (Chen et al., 2023). In the future work, CLAHE and advanced normalization techniques will be integrated with our procedure. Also, 3D CNN architecture for DICOM format processing and attention mechanisms for improved feature detection will be implemented together with multi-view fusion techniques for enhanced accuracy (Gökmen and Kocadaglı, 2024).

Our findings contribute to the literature on AI-based medical imaging. The findings indicate that deep learning based InceptionResNetV2 may provide a decision support system for the radiologist in terms of decreasing their workload such as interpretation and examination as well as enhancing diagnosis accuracy (Meşe et al., 2023). In the future direction, we are planning to apply the state of arts-based AI to various mammography benchmark datasets.

References

American Cancer Society. (2020). Breast Cancer Facts & Figures 2019-2020. Atlanta: American Cancer Society.

Chen, M., et al. (2023). Addressing Class Imbalance in Medical Image Classification: A Systematic Review. Computer Methods and Programs in Biomedicine, 229, 107325.

Elmore, J. G., ve diğerleri. (2016). Variability in Interpretive Performance at Screening Mammography and Radiologists' Characteristics Associated with Accuracy. Radiology, 281(2), 361-373.

Gokmen İ. N and Kocadagli, O. (2024). Multi-Class Classification of Thyroid Nodules from Automatic Segmented Ultrasound Images: Hybrid ResNet Based UNet Convolutional Neural Network Approach, Computer Methods and Programs in Biomedicine.

Johnson, K. S., et al. (2021). Variability in Mammogram Interpretation Among Radiologists: A Systematic Review. Journal of Medical Imaging, 8(2), 022418.

Kaggle. (n.d.). RSNA breast cancer detection dataset. Retrieved from https://www.kaggle.com/competitions/rsna-breast-cancer-detection/data

Kim, H., et al. (2023). Comparative Analysis of Deep Learning Architectures for Mammogram Classification. Artificial Intelligence in Medicine, 135, 102488.

Lee, J., et al. (2023). Hybrid CNN-Transformer Architecture for Enhanced Mammogram Analysis. Medical Image Computing and Computer Assisted Intervention, 13928, 234-242.

Lehman, C. D., ve diğerleri. (2015). Diagnostic Accuracy of Digital Screening Mammography with and without Computer-Aided Detection. JAMA Internal Medicine, 175(11), 1828-1837.

Litjens, G., ve diğerleri. (2017). A survey on deep learning in medical image analysis. Medical Image Analysis, 42, 60-88.

McKinney, S. M., et al. (2020). International evaluation of an AI system for breast cancer screening. Nature, 577(7788), 89-94.

Meşe İ, Gökmen İnan N., Kocadağli O., Salmaslioğlu A., Yildirim D. (2023). Chatgpt-assisted Deep Learning Model for Thyroid Nodule Analysis: Beyond Artificial Intelligence. Medical Ultrasonography, 25, Doi: 10.11152/Mu-430

Park, S., et al. (2024). InceptionResNetV2 for Medical Image Analysis: A Comprehensive Evaluation. Pattern Recognition Letters, 169, 81-89.

Ribli, D., ve diğerleri. (2018). Detecting and classifying lesions in mammograms with Deep Learning. Scientific Reports, 8(1), 4165.

Rodriguez, A., et al. (2022). Improving BI-RADS Classification Using Multi-Stage Deep Learning. Radiology: Artificial Intelligence, 4(3), e210069.

Shen, L., et al. (2023). Advanced Deep Learning Approaches for BI-RADS Classification in Digital Mammography. Medical Image Analysis, 84, 102680.

Vindr.ai. (n.d.). Mammo dataset. Retrieved from https://vindr.ai/datasets/mammo

Wang, X., et al. (2023). Impact of Preprocessing Techniques on Deep Learning Performance in Mammography. Journal of Digital Imaging, 36(2), 456-468.

Wu, N., et al. (2019). Deep Neural Networks Improve Radiologists' Performance in Breast Cancer Screening. IEEE Transactions on Medical Imaging, 38(5), 1227-1238.

Zhang, Y., et al. (2022). Contrast Enhancement Techniques for Improved Mammographic Feature Detection. IEEE Transactions on Medical Imaging, 41(8), 1892-1904.