Explainable AI Techniques for Interpretable Breast Cancer Classification

Tony K. Hariadi[1], Qodri Aziz[2], Slamet Riyadi[3]*, Kamarul Hawari Ghazali[4], Khairunnisa Binti Hasikin[5], Tri Andi[6]
Dept. of Electrical Engineering, Universitas Muhammadiyah Yogyakarta, Yogyakarta, Indonesia [1]
Artificial Intelligence and Robotic Center, Universitas Muhammadiyah Yogyakarta, Indonesia [2, 3]
Dept. of Information Technology, Universitas Muhammadiyah Yogyakarta, Yogyakarta, Indonesia [3, 6]
Faculty of Electrical and Electronic Engineering, Universiti Malaysia Pahang Al-Sultan Abdullah, Pekan, Pahang[4]
Dept. of Biomedical Engineering, Universiti Malaya, Kuala Lumpur, Malaysia [5]

Abstract—**Breast cancer is still a major health risk for women all over the world, and thus finding it early is very important for the patient's survival. Digital Breast Tomosynthesis (DBT) offers enhanced imaging capabilities relative to conventional mammography; yet, its quasi-3D characteristics provide distinct interpretability issues, often rendering deep learning models as black boxes. This work tackles the issue of transparency by testing three Explainable Artificial Intelligence (XAI) methods: Gradient-weighted Class Activation Mapping (Grad-CAM), Score-CAM, and Local Interpretable Model-Agnostic Explanations (LIME). The ResNet-50 architecture was utilized to analyse a dataset of 396 DICOM images that had been pre-processed in a unique way, including colour-mapping and balancing. The study used Insertion and Deletion Area Under the Curve (AUC) to carefully quantify how reliable the visual explanations were, in addition to usual criteria like accuracy, which achieved 94%. It was shown that LIME and Score-CAM generated attention maps that were dispersed or inconsistent, whereas Grad-CAM always showed lesion-specific areas with great accuracy. Grad-CAM was the best method for analysing DBT findings, since it had the highest Insertion AUC of 0.9078. These results provide radiologists with a way to trust and check automated diagnoses, which closes the gap between AI that works well and AI that is reliable in the clinic.**

Keywords—Breast cancer; DBT; Grad-CAM; ResNet-50; XAI

I. Introduction

Breast cancer is still the most common kind of cancer in women across the world and is the leading cause of cancer deaths. The World Health Organization (WHO) said that in 2020, there were more than 2.3 million new cases and 685,000 fatalities [1]. Breast cancer is the most common kind of cancer in Indonesia, making up around 30% of all cancer cases [2]. It is more common than cervical cancer. Finding the disease early is crucial for enhancing the chances of survival and the long-term outlook. Deep Learning (DL) on Convolutional Neural Networks (CNNs) has emerged as the most popular method for analyzing medical images to distinguish between benign and malignant breast cancers [3], [4]. Architectures like ResNet-50 have shown outstanding performance in classification tasks, but how they make decisions within is still not clear. Because the reasoning behind a model's categorization isn't always clear, medical practitioners generally don't trust it because it is a "black box" [5]. The Explainable Artificial Intelligence (XAI) method was developed to clarify and simplify deep learning (DL) models in order to address these issues. Specifically, XAI techniques were developed to visualize which image regions most significantly influence model predictions [6], [7]. Nonetheless, the majority of current XAI assessments concentrate on 2D imaging. Digital Breast Tomosynthesis (DBT) has distinct interpretative issues because of its quasi-3D attributes and overlapping tissue features, which markedly contrast with traditional mammography. At present, there is an absence of extensive research explicitly examining the performance of XAI methods—such as Grad-CAM, Score-CAM, and LIME—when used on DBT data with the ResNet-50 architecture.

This study seeks to evaluate and contrast three XAI methodologies—Grad-CAM, Score-CAM, and LIME—in analysing breast cancer classifications obtained from DBT images, addressing the current deficiency in the field. The evaluation employs visual saliency map analysis and quantitative metrics, namely the Deletion and Insertion Area Under the Curve (AUC). The aim of this research is to identify the most reliable XAI framework for 3D breast imaging, providing clinical decision-support insights and practical guidance for radiologists in selecting transparent AI technologies for medical diagnosis.

II. Literature Review

This section analyses previous studies that support this research. The literature reviewed includes topics related to breast cancer, DBT imaging, the use of deep learning in medical image classification, and the XAI methods used to understand CNN-based classification models. Buda et al. [8] released a large DBT dataset with 22,032 volumes from 5,060 patients and used DenseNet to create a basic model for finding breast cancer. This model had a sensitivity of 65% and two false positives per volume, but the study didn't say how easy it was to understand it. This work demonstrates the necessity of integrating XAI methods into DBT analysis to enhance the transparency and trustworthiness of medical AI systems.

Rodriguez-Ruiz et al. [9] showed that an autonomous AI system was able to detect cancer on mammograms with an Area Under the Curve (AUC) of up to 0.93, which is equivalent to human radiological performance. Meanwhile, Antropova et al. [10] used DCE-MRI images with the Maximum Intensity Projection (MIP) approach and CNN for the classification of benign and malignant lesions, resulting in an AUC of 0.88. Although neither study integrated XAI methods, their results

*Corresponding author.

open up important avenues for applying XAI to improve the clinical interpretability of CNN-based diagnostic systems.

A study by Kim et al. [10] more explicitly demonstrates the impact of visualisation on diagnostic confidence and accuracy. Using a CNN based on ResNet-34 and heatmap visualisation similar to Grad-CAM, the accuracy of mammography classification increased (AUC from 0.79 to 0.89), and the recall rate decreased from 60.4% to 49.5%. These findings confirm the importance of visual interpretability in increasing radiologists' trust in AI systems, although the study did not directly compare several XAI methods.

Di Martino et al. [5] examined various XAI methodologies employed in medicine, including SHAP, LIME, and Grad-CAM. Researchers found that Grad-CAM works well with CNNs because it can highlight important parts of medical images on its own, regardless of the architecture. LIME uses surrogate models to present local interpretations, and SHAP uses the theory of Shapley values to ensure that everything is consistent across the board. Many people prefer SHAP and Grad-CAM; however, the most effective XAI method depends on the specific model, context, and clarity of understanding. There have been no thorough studies that have directly compared the XAI methods Grad-CAM, LIME, and Score-CAM for classifying breast cancer using CNNs based on DBT.

Kursun et al. [11] applied Score-CAM to explain the classification results of deep learning models in leaf image-based plant disease detection. Score-CAM works without using gradients but instead calculates the weights of each activation channel based on the prediction score, resulting in more stable and less noisy visualisations than Grad-CAM. This study confirms that Score-CAM can improve the visual interpretability of CNN models without disturbing the model architecture. In the context of this study, Score-CAM is used as one of the XAI methods compared in assessing visualisation clarity and interpretation reliability in DBT image-based breast cancer classification.

Jusman et al. [12] assessed the effectiveness of GoogLeNet and ResNet-50 in the classification of X-ray images for COVID-19 detection. ResNet-50 always did better than GoogLeNet, with an average test accuracy of 94%, a precision and recall of 90%, and an F1-score of 89%. These results indicate that the residual learning architecture is better at finding complicated patterns in medical images. ResNet-50 was used as the CNN backbone for this study because it can deeply and accurately pull out important features, which are crucial for classifying breast cancer based on DBT images.

Based on a literature review, ResNet-50 has proven effective in medical image classification, including breast cancer, but interpretability challenges in DBT images still need to be addressed. These findings were utilized to select the appropriate model architecture and XAI methods for analysis during the implementation phase.

III. Methodology

This section outlines the methodologies and procedures employed in the research process, encompassing data processing, CNN model architecture, and the application and assessment of three XAI methods: Grad-CAM, LIME, and Score-CAM. This study uses the DBT image dataset for breast cancer classification and employs the ResNet-50 architecture as the primary CNN framework. Then, the XAI method analyses the classification results for clarity and consistency.

A. Research Flow

This research includes six stages, starting from data collection to the presentation of XAI results visualisation, as shown in Fig. 1.

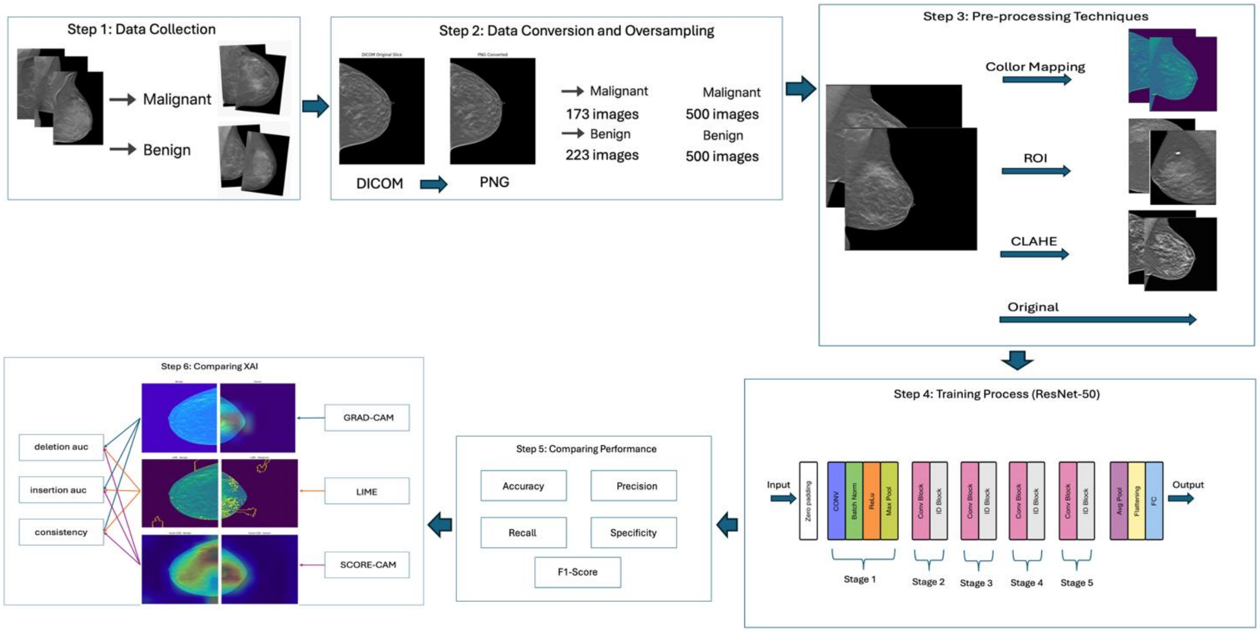

Fig. 1. Research flow.

The following sub-sections will provide a detailed explanation of the flow stages depicted in Fig. 1.

B. Pre-Processing

- Data Collection: This study uses data from patients diagnosed with breast cancer, consisting of two classes, namely 223 benign images and 173 malignant images, in DICOM format. Data were obtained from the National Cancer Institute through the Breast Cancer Screen – Digital Breast Tomosynthesis (BCS-DBT) dataset [13]. To visualise, the two types of classes, benign in Fig. 2 and malignant in Fig. 3, are shown below.

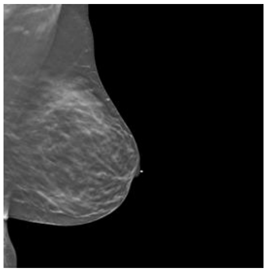

Fig. 2. Benign.

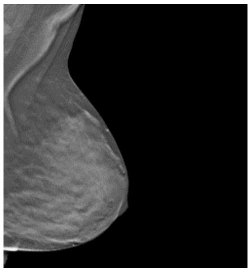

Fig. 3. Malignant.

- Data Conversion and Oversampling: The collected images are processed to facilitate the next stage of analysis. Images in DICOM format are converted into PNG format, called super images [14], with the aim of reducing the computational load without sacrificing image quality. Next, a random overlapping process is carried out to overcome the imbalance in the number of images between benign and malignant classes so that each class has 500 images [15].
- Pre-processing Techniques Image: This stage is carried out to obtain optimal CNN training results by applying the colour mapping technique. The colour mapping technique showed the highest accuracy compared to the other three techniques on DBT images, namely 94%.

C. Image Classification

Researchers used the ResNet-50 architecture to sort pictures. Post-training evaluation was performed using a confusion matrix with metrics such as accuracy, F1-score, recall, and precision. Subsequently, XAI methods were employed to visualize and interpret the underlying rationale behind the model's predictions.

1) *ResNet-50*: This model is used in this study to sort DBT images is a deep convolutional neural network with 50 layers that learns by connecting shortcuts. This mechanism effectively mitigates the accuracy degradation in exceedingly deep networks by acquiring residual functions, thereby resolving the issue of vanishing gradients. ResNet-50 has been shown to work better than other methods for a wide range of medical image classification tasks, such as finding breast and prostate cancer [16]. Fig. 4 presents the ResNet-50 architecture.

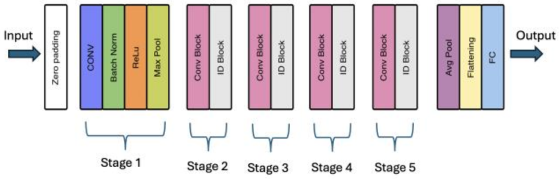

Fig. 4. ResNet-50 architecture.

The last layer of the ResNet-50 design was modified to correspond with the number of classes in the DBT dataset [17]. Training was conducted for 25 epochs using the Adam optimizer, with a learning rate of 0.0001 and a batch size of 10. A confusion matrix was utilized to evaluate the model's performance by determining the accuracy, precision, recall, specificity, and F1-score for each class.

2) *Confusion matrix*: Researchers performed evaluations of model classification using a confusion matrix, the value of which is measured based on four metrics: accuracy, F1-score, recall, and precision. The results of these calculations were used to assess whether the previous process successfully improved classification performance, achieved high accuracy, and optimally differentiated between benign and malignant classes [18]. The following section presents the equations for these four metrics.

a) Accuracy: It is a metric that indicates the overall accuracy of a model's predictions. This indicator is often used as an initial benchmark for model performance because it considers the accuracy of predictions for both the positive and negative classes. The accuracy calculation is shown in Eq. (1):

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100\% \qquad (1)$$

b) Recall or sensitivity: It is a metric that assesses a model's ability to identify positive data from a pool of truly positive data. This metric is crucial in cases such as disease detection or fraud, where missing a positive case can have serious consequences. The recall is defined in Eq. (2):

$$Recall = \frac{TP}{TP+FN} \times 100\% \qquad (2)$$

c) Precision: It describes the level of accuracy of positive predictions produced by the model, namely the proportion of correct positive predictions out of all positive predictions made. A high precision value indicates that the model rarely produces false positive predictions. The precision calculation is presented in Eq. (3):

$$Precision = \frac{TP}{TP+FP} \times 100 \quad (3)$$

d) F1-Score: It is an evaluation metric that combines precision and recall into a single value using the harmonic mean of both. This metric provides a balance between the model's ability to detect positive data (recall) and produce accurate positive predictions (precision). The F1-Score is determined using Eq. (4):

$$F1 - Score = 2 \times \frac{Sensivity \times Precision}{Sensitivity + Precision} \times 100\% \quad (4)$$

D. *Explainable Artificial Intelligence (XAI)*

Next, the researchers visualized the results of DBT-based breast cancer image classification using XAI to identify areas in the image that most influenced the model's decision. XAI was used to provide a transparent interpretation of the classification process by highlighting points or areas that serve as the basis for distinguishing between two cancer classes. In this study, three XAI methods were used, and their performance was evaluated using two metrics: Deletion AUC and Insertion AUC. The following is an explanation of each XAI model and an evaluation of its performance:

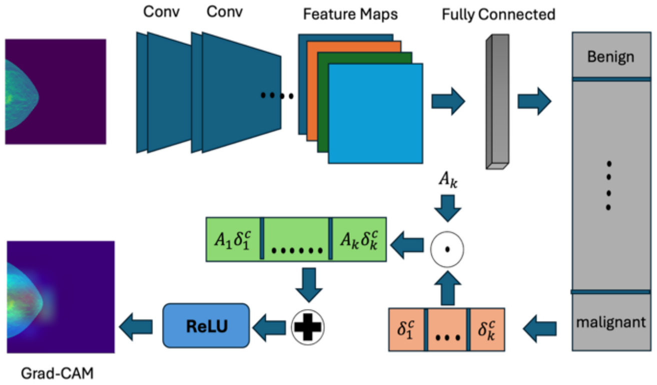

Fig. 5. Grad-CAM architecture.

1) Grad-Cam: Gradient-weighted Class Activation Mapping (Grad-CAM) is a method for showing which parts of DBT images have the most impact on the model's choice between benign and malignant breast cancer. Grad-CAM makes a heatmap that shows how much each area helped the model make a prediction by using the gradients from the last convolutional layer. This makes it easier to understand how to classify things visually [19]. The Grad-CAM model architecture is shown in Fig. 5.

Fig. 5 shows how the Grad-CAM method was used in this study with a CNN architecture based on ResNet-50.The DBT image passes through a number of convolutional layers, and the last one makes feature maps. After that, it goes to a fully connected layer that classifies tumours as either benign or malignant. To find the importance weights for each feature map, Grad-CAM takes the gradient of the target class output and multiplies it by the feature map that goes with it. The results go through a ReLU activation function, which only keeps the positive contributions. This creates a heatmap that shows which parts of the DBT image have the most effect on the model's prediction.

2) SCORE-CAM: The Score-Weighted Class Activation Mapping (Score-CAM) method is used to visualise important areas in DBT images that influence the model's decision to differentiate between benign and malignant breast cancer. This method produces a heatmap of the colour-mapped image, where colours with higher intensity indicate areas that contribute most to the model's prediction. This approach allows for a more transparent interpretation of the classification results, thus improving understanding of the model's decision-making process [11]. Fig. 6 displays the architecture of the Score-CAM model.

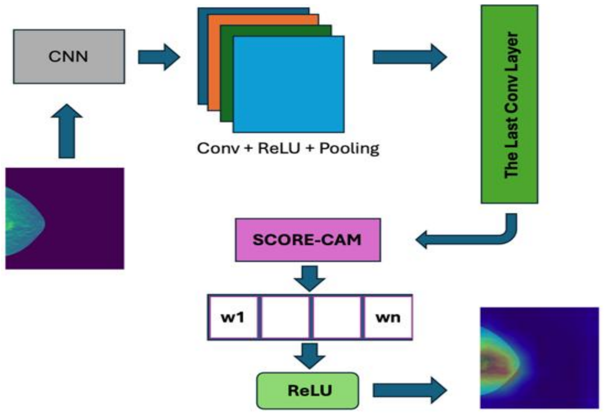

Fig. 6. Score-CAM architecture.

Fig. 6 shows the implementation flow of the Score-CAM method in this study using a CNN architecture using ResNet-50. The DBT is first processed by the ResNet-50 model through a series of convolutional layers, ReLU activation, and pooling to produce output in the final convolutional layer. Next, Score-CAM utilises the feature maps from the final layer by assigning weights (w_1 … w_n) based on the contribution of each feature map to the prediction. These weights are then combined and passed through a ReLU activation function to produce a heatmap that highlights the areas of the DBT image that are most influential in the classification of both benign and malignant tumour types.

3) LIME: Local Interpretable Model-Agnostic Explanations is an XAI method that locally explains black-box model predictions by generating synthetic data around an instance through random perturbation, predicting it again with the original model, and then training a simple model, such as linear regression, to determine the contribution of each feature. This method is model-agnostic and generates explanations that are appropriate for specific instances; however, it may yield different explanations for the same instance because of the randomness inherent in the process [20]. The results can be seen in Fig. 7.

Fig. 7 shows the flow of the LIME method implementation in this study with a CNN architecture using ResNet-50. The DBT image is processed by the CNN model to obtain an initial prediction, then LIME generates several perturbed samples around the original image. Each sample is re-predicted using the original model, and the results are then weighted based on their level of similarity to the original image. This weighted data is used to train a simple model that can be interpreted locally, resulting in a visualisation of the areas in the DBT image that most influence the decision to classify benign or malignant tumours.

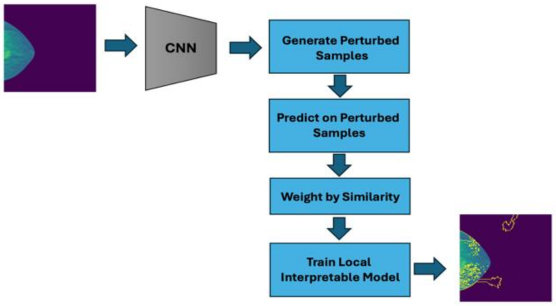
CNN
Generate Perturbed Samples
Predict on Perturbed Samples
Weight by Similarity
Train Local Interpretable Model

Fig. 7. LIME architecture.

E. XAI Evaluation Metrics

The evaluation of heatmap quality in this study was conducted using two quantitative metrics, namely deletion AUC and insertion AUC, which are commonly used to assess the performance of the XAI method [21]. The following section provides an explanation of the evaluation metrics.

- Deletion AUC: An evaluation metric used to measure the rate of decline in the model's prediction score when pixels deemed most important based on the heatmap are gradually removed from the original image. A lower Deletion AUC value indicates that the resulting importance map has better accuracy in identifying relevant areas. The deletion AUC calculation is shown in Eq. (5):

$$AUC_{del} = \frac{1}{T}\sum_{t=1}^{T} \frac{s_{t-1}^{del} + s_t^{del}}{2} \qquad (5)$$

- Insertion AUC: An evaluation metric used to measure the rate of improvement in the model's prediction score when important pixels are gradually added to the baseline image (blank or blurred). A higher Insertion AUC value indicates that the heatmap is able to identify important areas more effectively, increasing the confidence of the model's predictions. The insertion AUC calculation is shown in Eq. (6):

$$AUC_{ins} = \frac{1}{T}\sum_{t=1}^{T} \frac{s_{t-1}^{ins} + s_t^{ins}}{2} \qquad (6)$$

The methodology outlined includes every step, from getting and processing data to using the XAI method and checking how well it works. These steps are meant to make sure that the analysis is done in a planned way and to help talk about the results in the next sections.

IV. Results and Discussion

In this section, the researcher shows what happened when they used and tested the methods from Section III. The results include the use of three XAI methods—Grad-CAM, Score-CAM, and LIME—to show how easy it is to understand the data. They also include quantitative evaluation results using the Deletion AUC and Insertion AUC metrics. The analysis was performed to evaluate the visualisation quality and quantitative efficacy of each method for interpreting the outcomes of DBT image classification using the ResNet-50 architecture.

A. Image Classification Result

Researchers present the training results and confusion matrix of the ResNet-50 classification model to assess the model's performance in distinguishing benign and malignant breast cancer DBT images. The confusion matrix is visualised in Fig. 8, and the model training results are visualised in Fig. 9.

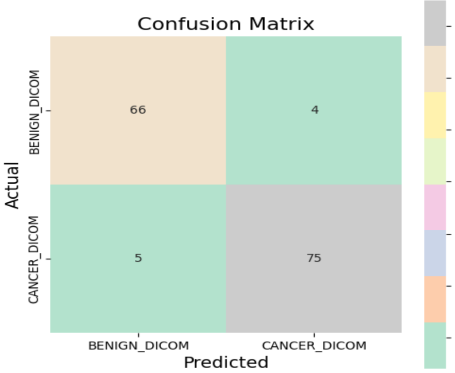

Fig. 8. Confusion matrix ResNet-50.

The graphs in Fig. 8, for accuracy and loss, show that the training is going well, with validation accuracy reaching about 94%. The loss values in the validation data change a bit, but the loss values in the training data keep going down. The classification results' confusion matrix is shown in Fig. 9. It shows that 66 benign images and 75 cancerous images were correctly sorted. The colour mapping method gives more accurate results than other ways of pre-processing.

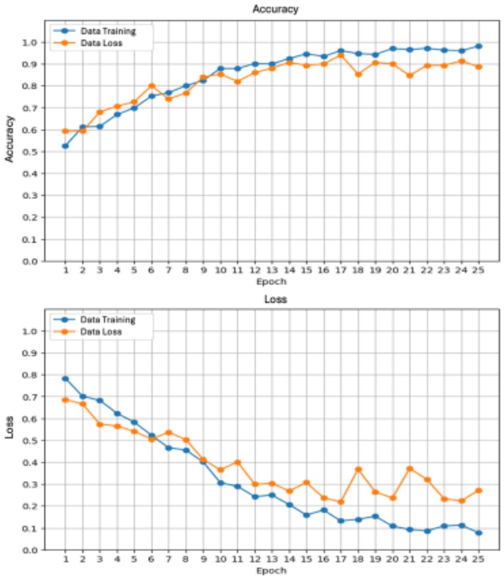

Fig. 9. ResNet-50 training results.

Additionally, performance metrics were evaluated to measure the model's accuracy, recall, precision, and F1-score in classifying. In this test, class 0 represents the benign category, while class 1 represents the malignant category. The performance evaluation results are shown in Table I.

TABLE I. PERFORMANCE EVALUATION OF RESNET-50

Accuracy	Recall		Precision		F1-Score	
	0	1	0	1	0	1
94%	94%	94%	93%	95%	94%	94%

Table I says that the ResNet-50 model, which used colour mapping pre-processing, got 94% accuracy. The values for recall, precision, and F1-score were about the same for both classes. This indicates that the model can consistently and accurately distinguish between benign and malignant breast cancer images.

B. *Explainable Artificial Intelligence Results*

This section displays the visualization results of Grad-CAM, Score-CAM, and LIME on DBT images, indicating the model's focus areas during classification. Visualizations are presented for both benign and malignant classes, allowing for comparison of the model's attention patterns in each category. The results and explanations are shown in the following figure.

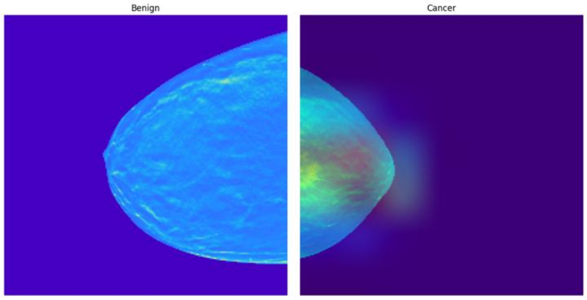
Benign
Cancer

Fig. 10. Grad-CAM result.

Fig. 10 shows the results of Grad-CAM on DBT images for two classes: benign and malignant. In benign images, the colour distribution appears predominantly blue, indicating a low level of model activation in certain areas, resulting in no significant focus on the suspected tissue. Meanwhile, in malignant images, a yellow to red area is visible in the centre, indicating a high level of activation. This indicates that the model is focusing more attention on these areas as indicators of the presence of cancerous lesions.

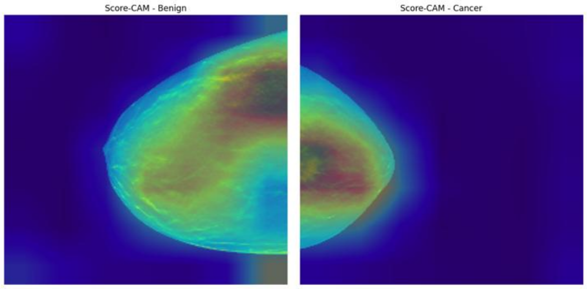
Score-CAM - Benign
Score-CAM - Cancer

Fig. 11. Score-CAM result.

Fig. 11 presents Score-CAM visualisation results. In benign images, the model's activation areas are fairly evenly distributed with green to yellow colour intensities, indicating a moderate level of focus in some parts of the breast tissue. Meanwhile, in malignant images, the model's focus appears more concentrated in the central area, with yellow-to-red colour intensities, but the activation distribution appears less clear than the Grad-CAM results. This aligns with the metric evaluation results, where Score-CAM performed slightly lower than Grad-CAM.

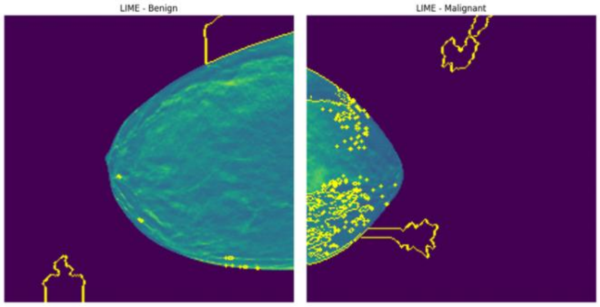
LIME - Benign
LIME - Malignant

Fig. 12. LIME result.

In Fig. 12, the LIME model highlights the image areas deemed most relevant by the model using yellow borders. In benign images, the highlighted areas are relatively small and scattered across a few points, indicating a lack of indications deemed significant by the model. In contrast, in malignant images, the highlighted areas are much larger and concentrated in specific areas of the breast, indicating that the model identified these areas as strong indicators of abnormalities.

C. *XAI Evaluation Metrics Result*

The performance of the XAI method was measured quantitatively using the Deletion AUC and Insertion AUC metrics. These two metrics assess the quality of the saliency map generated by the XAI method based on its impact on model predictions. The resulting metrics can be seen in Table II.

TABLE II. XAI MODEL EVALUATION

Model XAI	Benign		Malignant		Mean	
	Deletion AUC	Insertion AUC	Deletion AUC	Insertion AUC	Deletion AUC	Insertion AUC
Grad CAM	0.8241	0.9092	0.2248	0.9064	0.5245	0.9078
Score CAM	0.9089	0.9218	0.1570	0.8152	0.5330	0.8685
LIME	0.9145	0.9559	0.1323	0.0894	0.5234	0.5226

Insertion AUC value (0.9078) and a competitive Deletion AUC value, demonstrating its ability to highlight areas that are truly relevant to the model's decision. Score-CAM provides results close to Grad-CAM on the Deletion AUC metric, but lower on the Insertion AUC. Meanwhile, LIME produces relatively stable values on both metrics but does not exceed the performance of Grad-CAM. These results indicate that Grad-CAM is the most effective XAI method for interpreting DBT image classification models in this study.

D. *Discussion*

The results of this study provide significant insights into the interpretability of DBT. Even though the images were processed as 2D slices for model training, they still have the complicated quasi-3D features and overlapping tissue structures that are unique to the DBT modality. This modality is very different from regular mammography. The excellent localization of Grad-CAM, which has an Insertion AUC of

0.9078, suggests that it could be used as a "second opinion" tool to help radiologists understand these complicated features. The results also show that gradient-based methods like Grad-CAM are more reliable than perturbation-based methods like LIME, which gave less consistent results on these particular textures. The researcher believes that these results show that high classification accuracy, 94%, in this study needs to be combined with precise feature localization to be useful in future clinical decision support.

V. Conclusion

This study assesses three XAI methodologies—Grad-CAM, Score-CAM, and LIME—for elucidating breast cancer classification in Digital Breast Tomosynthesis (DBT) images, employing a meticulously calibrated ResNet-50 architecture. The model was 94% accurate, and the precision, recall, and F1-score measures were all equal. Visual analysis indicates that Grad-CAM generates the most effective attention maps for lesion regions, while Score-CAM and LIME exhibit greater inconsistency, particularly in cases of cancer. Quantitative analysis corroborates these results, with Grad-CAM attaining the highest Insertion AUC (0.9078). This is why Grad-CAM is the best XAI framework for DBT categorization: it makes it easy and accurate to use AI in medicine.

Limitations and Prospective Research: The study's results are promising, but they are limited by a small dataset of 396 photos and the use of only one model architecture. The research emphasizes 2D-converted slices instead of comprehensive 3D volumetric analysis. Subsequent research ought to employ larger, multi-centre datasets and explore transformer-based architectures or 3D-CNNs to improve the generalizability of XAI performance in breast cancer diagnosis.

Acknowledgment

The authors would like to acknowledge Universitas Muhammadiyah Yogyakarta for the financial support of this research and publication.

References

[1] E. Nolan, G. J. Lindeman, and J. E. Visvader, "Deciphering breast cancer: from biology to the clinic," *Cell*, vol. 186, no. 8, pp. 1708–1728, Apr. 2023, doi: 10.1016/j.cell.2023.01.040.

[2] Ferlay *et al.*, "Global Cancer Observatory: 360 Indonesia Fact Sheet. International Agency for Research on Cancer," 2024. Accessed: May 29, 2025. [Online]. Available: https://gco.iarc.who.int/today, accessed.

[3] A. M. Sharafaddini, K. K. Esfahani, and N. Mansouri, "Deep learning approaches to detect breast cancer: a comprehensive review," *Multimed Tools Appl*, Aug. 2024, doi: 10.1007/s11042-024-20011-6.

[4] H. Zhang and Y. Qie, "Applying Deep Learning to Medical Imaging: A Review," *Applied Sciences*, vol. 13, no. 18, p. 10521, Sep. 2023, doi: 10.3390/app131810521.

[5] Q. Teng, Z. Liu, Y. Song, K. Han, and Y. Lu, "A survey on the interpretability of deep learning in medical diagnosis," *Multimed Syst*, vol. 28, no. 6, pp. 2335–2355, Dec. 2022, doi: 10.1007/s00530-022-00960-4.

[6] N. Rane, S. Choudhary, and J. Rane, "Explainable Artificial Intelligence (XAI) in healthcare: Interpretable Models for Clinical Decision Support," *SSRN Electronic Journal*, 2023, doi: 10.2139/ssrn.4637897.

[7] R. Karthiga, K. Narasimhan, T. V, H. M, and R. Amirtharajan, "Review of AI &amp; XAI-based breast cancer diagnosis methods using various imaging modalities," *Multimed Tools Appl*, vol. 84, no. 5, pp. 2209–2260, Oct. 2024, doi: 10.1007/s11042-024-20271-2.

[8] M. Buda *et al.*, "A Data Set and Deep Learning Algorithm for the Detection of Masses and Architectural Distortions in Digital Breast Tomosynthesis Images," *JAMA Netw Open*, vol. 4, no. 8, p. e2119100, Aug. 2021, doi: 10.1001/jamanetworkopen.2021.19100.

[9] M. Madani, M. M. Behzadi, and S. Nabavi, "The Role of Deep Learning in Advancing Breast Cancer Detection Using Different Imaging Modalities: A Systematic Review," *Cancers (Basel)*, vol. 14, no. 21, p. 5334, Oct. 2022, doi: 10.3390/cancers14215334.

[10] Y. S. Kim *et al.*, "Use of Artificial Intelligence for Reducing Unnecessary Recalls at Screening Mammography: A Simulation Study," *Korean J Radiol*, vol. 23, no. 12, p. 1241, 2022, doi: 10.3348/kjr.2022.0263.

[11] R. Kursun and M. Koklu, "Enhancing Explainability in Plant Disease Classification using Score-CAM: Improving Early Diagnosis for Agricultural Productivity," in *2023 IEEE 12th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, IEEE, Sep. 2023, pp. 759–764. doi: 10.1109/IDAACS58523.2023.10348713.

[12] Y. Jusman, A. Z. U. Haq, M. A. Nur'Aini, and N. Hadiansyah, "Classification of Covid Image Based on Deep Learning Model GoogLeNet and ResNet-50," in *ICECOS 2024 - 4th International Conference on Electrical Engineering and Computer Science, Proceeding*, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 199–204. doi: 10.1109/ICECOS63900.2024.10791165.

[13] M. Buda *et al.*, "Breast Cancer Screening – Digital Breast Tomosynthesis (BCS-DBT) (Version 5)," 2020.

[14] I. Sobirov, N. Saeed, and M. Yaqub, "Super Images -- A New 2D Perspective on 3D Medical Imaging Analysis," May 2023, [Online]. Available: http://arxiv.org/abs/2205.02847

[15] S. Riyadi, A. D. Andriyani, and A. M. Masyhur, "Improving Hate Speech Detection Accuracy Using Hybrid CNN-RNN and Random Oversampling Techniques," in *2024 IEEE Symposium on Industrial Electronics and Applications, ISIEA 2024*, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/ISIEA61920.2024.10607232.

[16] Y. Jusman, M. A. F. Nurkholid, and F. Utomo, "Prostate Image Classification Using Pretrained Models: GoogLeNet and ResNet-50," in *2021 15th International Conference on Signal Processing and Communication Systems, ICSPCS 2021 - Proceedings*, T. A. Wysocki and B. J. Wysocki, Eds., Institute of Electrical and Electronics Engineers Inc., 2021. doi: 10.1109/ICSPCS53099.2021.9660334.

[17] N. Dif and Z. Elberrichi, "A New Intra Fine-Tuning Method Between Histopathological Datasets in Deep Learning," *International Journal of Service Science, Management, Engineering, and Technology*, vol. 11, no. 2, pp. 16–40, Apr. 2020, doi: 10.4018/IJSSMET.2020040102.

[18] D. R. Nayak, N. Padhy, P. K. Mallick, M. Zymbler, and S. Kumar, "Brain Tumor Classification Using Dense Efficient-Net," *Axioms*, vol. 11, no. 1, p. 34, Jan. 2022, doi: 10.3390/axioms11010034.

[19] S. Riyadi, E. N. Pramudya, C. Damarjati, J. M. Molina Lopez, and J. G. Herrero, "Explainable optimization of deep learning model for COVID-19 detection using chest images," *Inform Med Unlocked*, vol. 49, p. 101559, 2024, doi: 10.1016/j.imu.2024.101559.

[20] M. R. Zafar and N. Khan, "Deterministic Local Interpretable Model-Agnostic Explanations for Stable Explainability," *Mach Learn Knowl Extr*, vol. 3, no. 3, pp. 525–541, Jun. 2021, doi: 10.3390/make3030027.

[21] N. Hama, M. Mase, and A. B. Owen, "Deletion and Insertion Tests in Regression Models," 2023. [Online]. Available: http://jmlr.org/papers/v24/22-0560.html.