Vision-Based Explainable Diagnosis Framework for Breast Cancer

Tshepiso Mahoko

University of Johannesburg, South Africa

220015607@student.uj.ac.za

Abstract

Medical image analysis has greatly advanced the detection of cancer abnormalities, primarily through the success of deep learning. Modern computer-aided detection and diagnosis (CAD) systems employ multiple medical imaging modalities to support clinicians in identifying and interpreting anomalies. These systems rely on image processing technologies to acquire, analyse, and store multi-modal data, forming a critical foundation for medical research and diagnosis. Cancer continues to be a major cause of morbidity and mortality worldwide. Specifically, breast cancer affects about 8% of women, with approximately 4,300 related deaths reported in the United States in 2022 [1].

This research focuses on the development of a vision-based CAD system, which differs from traditional approaches by leveraging visual deep learning models to minimize human error, fatigue, and diagnostic bias [2]. The growing integration of Explainable AI (XAI) in such systems has addressed long-standing concerns about the "black box" nature of deep learning models. XAI enhances interpretability by highlighting the visual features and regions influencing diagnostic predictions, fostering greater transparency and clinical trust. As deep learning continues to evolve in radiology, image segmentation has become essential for capturing fine-grained local features within medical images. This paper presents a vision-based breast cancer CAD system incorporating convolutional neural networks (CNNs), image segmentation, and Explainable AI, while also examining existing systems, their limitations, and the challenges that persist in current research.

Keywords: *Computer-Aided Detection (CADe), Computer-Aided Diagnosis (CADx), Convolutional Neural Network (CNN), Explainable AI (XAI)*

1. Introduction

Cancer remains one of the most critical global health challenges, with breast cancer representing the most common and deadliest form among women, causing approximately 670,000 deaths worldwide in 2022 [5]. Early and accurate detection is vital to improving survival rates and treatment outcomes. However, the diagnostic process often depends on human interpretation of complex medical images, which is prone to fatigue, workload stress, and diagnostic bias, leading to false negatives and delayed diagnoses [6]. Computer-aided detection and diagnosis (CADx) systems have therefore become increasingly valuable in assisting clinicians to detect and analyse abnormalities in medical images [3]. These systems use imaging modalities such as mammograms, ultrasounds, magnetic resonance imaging (MRI), and microscopic imaging (MI) to enhance diagnostic accuracy [4]. Yet, traditional CAD systems still rely heavily on manual interpretation and lack the visual intelligence and explainability required for consistent performance across diverse imaging conditions.

This research addresses these limitations by developing a vision-based CADx system for breast cancer detection that leverages deep learning and explainable artificial intelligence (XAI) to improve accuracy and interpretability, following established practices in medical machine learning outlined in Sidey-Gibbons & Sidey-Gibbon [10]. The proposed system seeks to overcome human-related diagnostic limitations by integrating supervised machine learning techniques, particularly Convolutional Neural Networks (CNNs), which have shown superior performance in image classification tasks [11]. CNNs automatically learn hierarchical spatial features from medical images, enabling the identification of subtle patterns in breast tissue, such as calcifications and asymmetries, which may not be immediately visible to radiologists [1], [12]. Multiple studies demonstrate that CNNs can match or even exceed expert-level accuracy in mammogram and histopathological classification, significantly reducing false positives and negatives [8].

To increase transparency and trust among healthcare professionals, this study incorporates U-Net segmentation for feature extraction and data augmentation before training, allowing the model to focus on clinically relevant regions. Additionally, Gradient-weighted Class Activation Mapping (Grad-CAM) is employed to visualize areas of the image that most influence the CNN's decision-making process, providing interpretability and clinical validation. Together, CNNs, U-Net, and Grad-CAM form an integrated and explainable deep learning pipeline that supports reliable, transparent, and automated cancer detection.

2. Societal Relevance

Breast cancer remains one of the leading causes of cancer-related deaths among women globally, with the World Health Organization reporting it as the most common cancer among women in 157 out of 185 countries in 2022 [5]. In South Africa, approximately 1 in 27 women [13] will be diagnosed with breast cancer during their lifetime, underscoring its significant contribution to female mortality rates. Several risk factors, as highlighted in

Fig. 1, contribute to the development of breast cancer, including advancing age, obesity, alcohol consumption, genetic predisposition, prior radiation exposure, reproductive history (such as age at menstruation and first pregnancy), tobacco use, and hormone therapy after menopause [5].

Fig 1. Risk factors contributing to Breast Cancer

This research seeks to bridge the gap in early breast cancer detection by applying machine learning techniques particularly convolutional neural networks (CNNs) to analyse medical images more efficiently and accurately. By enabling the identification of early-stage signs of breast cancer, this approach can help reduce diagnostic delays, improve treatment outcomes, and alleviate the workload of radiologists, especially in regions with limited access to specialized healthcare professionals. A reliable computer-aided detection and diagnosis (CADx) system, enhanced with explainable AI methods like Gradient-weighted Class Activation Mapping (Grad-CAM), can further ensure transparency and trust in automated medical decision-making, thereby supporting clinical adoption, improving diagnostic accuracy, and ultimately saving lives through more accessible and efficient screening processes.

3. Literature Review

3.1 Supervised Learning in Radiology

Machine learning and image processing form the foundation of Computer-Aided Diagnosis (CAD) systems in radiology. The choice of model significantly affects diagnostic accuracy and reliability [14]. Deep learning methods have shown superior performance, reducing false positives and negatives in manual screening [8].

Different machine learning paradigms suit different diagnostic tasks depending on dataset size, quality, and computational capacity [14]. Traditional models perform well with engineered features and smaller datasets, whereas Convolutional Neural Networks (CNNs) excel in learning hierarchical features directly from raw data. This research focuses on developing a CNN-based CADx system to assist radiologists in accurately detecting and diagnosing abnormalities in mammograms [15][16].

3.2 Mammographs and Training Data

Publicly available datasets have greatly advanced medical imaging research, improving detection accuracy for diseases such as breast cancer and pneumonia [14]. Annotated datasets allow models to learn reliable diagnostic patterns, particularly for tasks like tumour localization and organ segmentation [14].

This study uses the Cancer Imaging Archive (TCIA) dataset, which provides a large, annotated collection of medical images (MRI, CT, and mammography). High-quality annotations are essential for model training but are costly and time-consuming, requiring expert radiologist input [14]. Although this introduces variability, TCIA remains invaluable for developing robust, generalizable CAD systems. The dataset's diversity supports deep learning models capable of improving diagnostic reliability across various imaging modalities.

3.3 Convolutional Neural Networks (CNNs) in Medical Diagnosis

CNNs have demonstrated superior performance to radiologists in detecting malignant nodules, significantly reducing false positives [8]. Their ability to learn spatial hierarchies through convolutional, pooling, and dense layers enables automatic feature extraction without manual engineering [14][15][17].

A CNN typically consists of convolutional, normalization, pooling, dropout, and dense layers, each extracting increasingly abstract representations of the input. This hierarchy allows the network to detect subtle variations in imaging data that often signal pathology [15]. CNNs have also proven capable of generalizing across diverse datasets, achieving strong diagnostic accuracy without relying on manual segmentation [17].

3.4 Image Preprocessing and Feature Extraction

Preprocessing is crucial for improving image quality and ensuring model consistency. In this study, a UNet-based encoder pipeline was implemented to enhance mammography images before classification. The process includes artifact suppression, breast segmentation, and pectoral muscle removal, isolating diagnostically relevant regions.

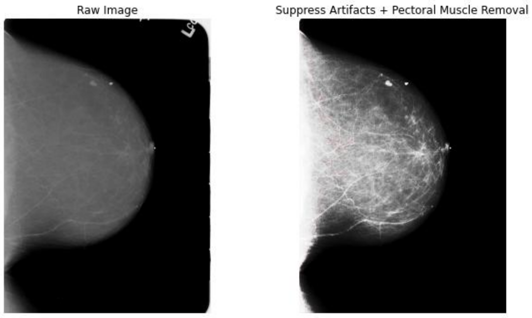

Fig 2. Data cleaning and feature extraction

Fig. 2 shows how raw mammograms are transformed into focused breast regions, demonstrating enhanced contrast and feature clarity. This step ensures that only high-quality, standardized images are used in model training.

3.5 Image Segmentation in Breast Cancer Analysis

Segmentation isolates regions of interest (ROIs), such as tumours, enabling the CNN to focus on relevant structures [1][14]. The integration of segmentation with CNNs enhances both accuracy and interpretability.

The U-Net architecture is widely used for biomedical image segmentation due to its encoder–decoder symmetry and skip connections that retain spatial detail [11][12]. This structure captures both local and global context, improving tumour delineation and diagnostic precision.

3.6 Explainable AI (XAI) and Grad-CAM in Medical Imaging

A major challenge in CADx systems is the “black-box” nature of AI models. Explainable AI (XAI) techniques, such as Grad-CAM and LIME, enhance interpretability by visually highlighting regions that influence model predictions [11][12]. These visual explanations, shown in Fig. 3, help radiologists confirm whether the model focuses on clinically relevant areas such as tumour boundaries [1].

This interpretability improves trust, aligns with healthcare regulations, and aids clinical training by providing insight into model decision-making [18].As AI tools become increasingly integrated into diagnostics, the inclusion of XAI ensures transparency, accountability, and clinician confidence essential for the ethical adoption of AI in healthcare.

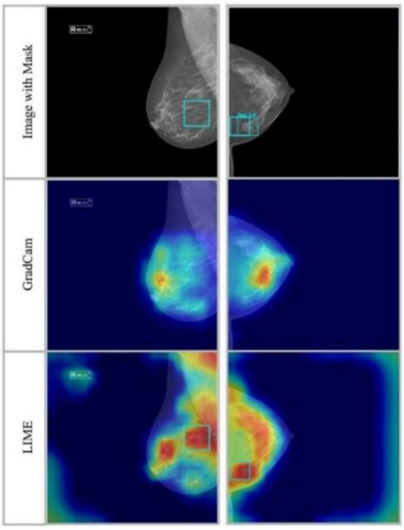

Fig. 3. Example Activation heatmaps through Explainable AI methods

4. Methodology

4.1 Deep Learning Framework

This research proposes a computer-aided diagnosis (CADx) system for breast cancer detection using a deep learning approach based on Convolutional Neural Networks (CNNs). Publicly available data from The Cancer Imaging Archive (TCIA) provides diverse, high-quality medical images for training, validation, and testing of the model.

4.2 U-Net Segmentation

The pipeline begins with image preprocessing and segmentation using the U-Net architecture, which efficiently isolates tumour regions and relevant breast tissue. Segmented images are then fed into a custom CNN model to classify regions as benign or malignant, supporting accurate diagnosis even with limited annotated data. Fig. 4 illustrates multiple segmented variations of a single mammogram, showing the model's ability to highlight ROIs effectively

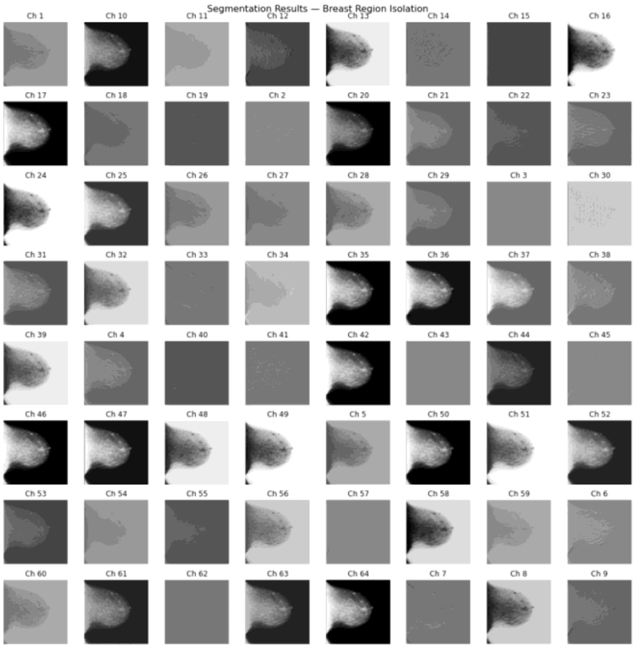

Fig. 4. segmented variations of a single mammogram scan using U-Net.

4.3 Explainable AI with Grad-CAM

To enhance interpretability and build clinical trust, Grad-CAM is applied to visualize the regions influencing model predictions. This allows radiologists to verify that the model focuses on clinically relevant areas, improving transparency and usability in diagnostic workflows.

4.4 Model Training and Validation

The dataset is split into training and test sets, and the model is trained iteratively in batches. Each batch updates network weights based on the loss function, while performance metrics such as accuracy, precision, recall, and F1-score are recorded after each epoch. This approach ensures efficient learning while continuously evaluating generalization on unseen data.

4.5 CNN Architecture

The CNN model comprises an input layer for pre-processed images, multiple convolutional layers with ReLU activations to extract spatial features, max-pooling layers to reduce dimensionality, batch normalization to stabilize training, and dropout layers to prevent overfitting. Fully connected layers integrate features for final classification, with a sigmoid output layer distinguishing between benign and malignant cases.

4.6 Proposed Comprehensive Framework

Fig. 5 illustrates the proposed CADx pipeline, which integrates preprocessing, U-Net segmentation, CNN-based classification, and Grad-CAM-based explainability. This structured workflow ensures the system captures critical image features while remaining interpretable and generalizable. By combining segmentation, deep learning, and explainable AI, the pipeline provides a robust and clinically relevant tool to support early breast cancer detection.

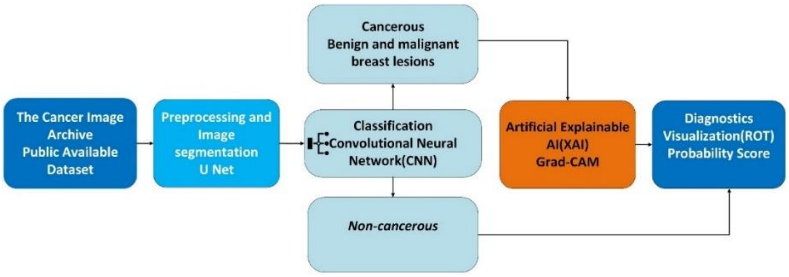

Fig. 5. Comprehensive Framework

5. Model Parameters and Performance Metrics

The performance of the developed Convolutional Neural Network (CNN) model was evaluated on a mammography dataset after applying the proposed preprocessing pipeline (artifact suppression, breast segmentation, and pectoral muscle removal) and image segmentation, with 64 channels (256×256 pixels) extracted from the segmented region.

5.1 Model Architecture

- Architecture: CNNModel
- Convolutional Layers: [[32, 3], [64, 3]]
- Hidden Units (Dense Layers): [256, 128]
- Dropout Rate: 0.1

These hyperparameters were selected after evaluating a series of different configurations, this specific setup yielded the highest performance in terms of validation accuracy, F1-score, and generalization on classification.

5.2 Dataset

- Total Samples: 245
- Classes: 2 (BENIGN, MALIGNANT)
- Train/Test Split: 220/25
- Input Shape: [64, 256, 256]

The dataset was balanced and preprocessed to standardize input size and intensity, ensuring compatibility with the CNN model.

5.3 Loss and Accuracy

- Epochs: 60
- Batch Size: 32
- Learning Rate: 0.001
- Device: CPU
- Training Time: 00:16:21

Loss vs Epoch: Fig. 6 shows the training loss steadily decreased over the epochs, indicating that the model was learning effectively from the preprocessed data. Initial fluctuations were observed during the early epochs, which gradually stabilized toward the later epochs.

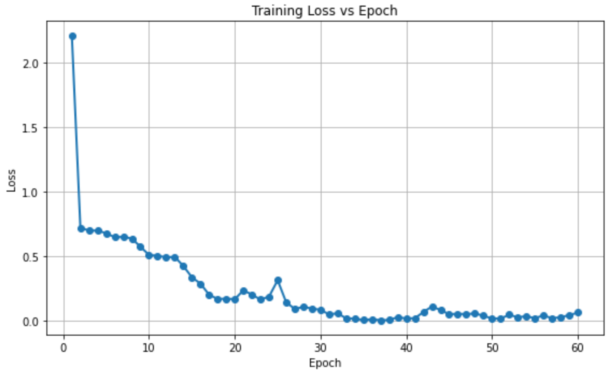

Fig. 6. Training loss

Validation Accuracy vs Epoch: The validation accuracy initially varied across the epochs but eventually reached a maximum at around 76% as illustrated on Fig. 7, reflecting good generalization and convergence of the model

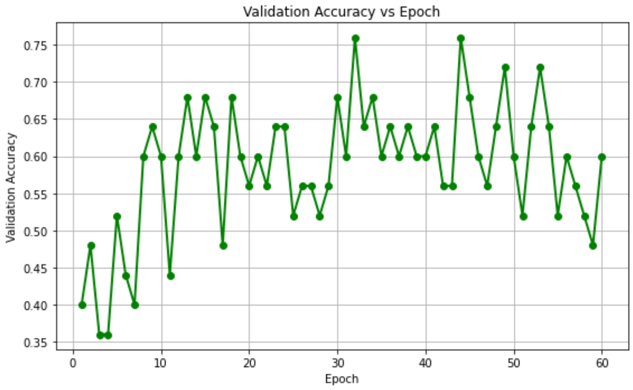

Fig. 7. Validation Accuracy

These plots highlight that the selected CNN architecture and preprocessing pipeline were effective in guiding the model toward robust feature extraction and classification performance.

5.4 Evaluation Metrics

The trained CNN achieved the following performance on the test set, summarised Table 1 and Table 2.

Table 1. Classification Metrics

Class	Precision	Recall	F1-score	Support
0 (BENIGN)	0.769	0.769	0.769	3
1 (MALIGNANT)	0.750	0.750	0.750	12
Weighted Avg	0.760	0.760	0.760	25

Table 2. Confusion Metrics

Actual \ Predicted	BENIGN	MALIGNANT
BENIGN	**10**	**3**
MALIGNANT	**3**	**9**

The model demonstrates balanced classification performance across both classes as shown in Table 1, with a high weighted average for precision, recall, and F1-score.

6. Analysis

The CNN-based CADx system achieved a weighted F1-score of 0.76, demonstrating balanced precision and recall for both benign and malignant classes. The confusion matrix revealed minimal misclassifications, indicating that the model effectively learned key features from pre-processed mammography images and could distinguish between the two categories with reasonable accuracy.

Model performance, however, was constrained by computational limitations. Training on full-resolution mammograms was not feasible due to memory constraints and the absence of GPU acceleration, forcing image downscaling to 256×256 pixels. This resizing likely reduced accuracy, as critical tumour details and fine textures were lost. Long CPU-based training cycles also limited experimentation with deeper architectures and slowed hyperparameter tuning, while early training was affected by vanishing and exploding gradients that later stabilized through batch normalization.

Despite these constraints, the preprocessing and segmentation pipeline successfully standardized inputs and emphasized relevant features, leading to steady loss reduction and stable validation accuracy around 76%.

The system achieved a balanced F1-score of 76%, which is comparable to the sensitivity of 76.5% reported for radiologists using AI-CAD assistance [7]. This aligns with reporting standards highlighted by Yusuf et al. [9] who emphasize the importance of transparent performance metrics and standardized evaluation in studies applying machine learning to medical diagnosis

Future work should employ GPU acceleration with higher-resolution imaging or cloud-based computing to handle full-resolution images and enable deeper models. Automated hyperparameter optimization and larger, more diverse datasets could further enhance generalization. Integrating advanced explainable AI (XAI) techniques would also improve interpretability and clinical trust, supporting real-world adoption of AI-driven diagnostic tools.

7. Conclusion

This research developed a vision-based hybrid deep learning CADx system for breast cancer detection that integrates Convolutional Neural Networks (CNNs), U-Net segmentation, and Explainable AI (XAI) to enhance diagnostic accuracy and interpretability. The system achieved a balanced F1-score of 76%, demonstrating its ability to learn key features from mammography images despite computational constraints and reduced image resolution. Incorporating Grad-CAM provides visual transparency, enhancing trust and usability while supporting reliable early breast cancer detection and aiding radiologists' decisions.

References

[1] F. Rajeena P.P and S. Tehsin, "A Framework for Breast Cancer Classification with Deep Features and Modified Grey Wolf Optimization," *Mathematics,* vol. 13, no. 8, 2025.

[2] E. M. F. El Houby, “Framework of Computer Aided Diagnosis Systems for Cancer Classification Based on Medical Images,” *National Library of medicine,* p. 1, 2018.

[3] N. Petrick and B. Sahiner, “Evaluation of computer-aided detection and diagnosis systems,” *Medical Physics,* pp. 4-20, 2013.

[4] T. M. Deserno (né Lehmann), H. Handels, K. H. Maier-Hein (né Fritzsche), S. Mersmann, C. Palm, . T. Tolxdorff, G. Wagenknecht and T. Wittenberg, “Viewpoints on Medical Image Processing: From Science to Application,” *Bentham Open Access,* pp. 3-8, 2013.

[5] World health organization, “World health organization,” 2022. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cancer. [Accessed 08 03 2024].

[6] M. Payal, K. S. Kumar and D. T. A. Kumar, “A Review of Computer Aided Diagnosis (CAD) Techniques in Healthcare,” *INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCHIN SCIENCE, ENGINEERING AND TECHNOLOGY,* vol. 5, no. 5, pp. 3-6, 2022.

[7] L. Si Eun, H. Hanpyo and K. Eun-Kyung, “Diagnostic performance with and without artificial intelligence assistance in real-world screening mammography,” *European Journal of Radiology Open,* vol. 12, pp. 1-5, 2024.

[8] S. Pawlusek, “AI: The Effectiveness of Early Cancer Detection Programs Using AI Algorithms,” *Advances in Breast Cancer Research,* vol. 14, no. 2, pp. 2168-1597, 2025.

[9] M. Yusuf, I. Atal, a. JLi, P. Smith, P. Ravaud, M. Fergie, M. Callaghan, R. Philippe, . F. Martin and M. Callaghan, “Reporting quality of studies using machine learning models for medical diagnosis: a systematic review, ” *bjm,* vol. 10, no. 3, p. 2024, 2019.

[10] J. A. M. Sidey-Gibbons and C. J. Sidey-Gibbons, “Machine learning in medicine: a practical introduction, ” *BMC Medical Research Methodology,* pp. 3-18, 2019.

[11] O. Kresse, Y. K. Hassan Rezk, A. I. Said, R. H. Mousa and M. A. A. Ibrahim, “Evaluating Machine Learning Techniques for Breast Cancer,” *Mansoura Engineering Journal,* vol. 50, no. 1, pp. 1-3, 2025.

[12] S. Paavankumar, R. Karthik, G. Idayachandiran and P. D. Penchala Sri, “Classification of benign and malignant breast lesions in mammograms using dense-unified multiscale attention network and data-efficient image transformers,” *THE EUROPEAN PHYSICAL JOURNAL S PECIAL T OPICS,* 2025.

[13] Medical Academic, “Medical Academic,” [Online]. Available:

https://www.medicalacademic.co.za/womens-health/1-in-27-sa-women-affected-by-breast-cancer/#:~:text=Breast%20cancer%20is%20the%20most%20prevalent%20cancer,women%20and

%2023%%20of%20all%20cancers%20diagnosed.. [Accessed 21 04 2025].

[14] E. C. Chianumba, N. Ikhalea, A. Y. Mustapha and A. . Y. Forkuo, “A Conceptual Model for Using Machine Learning to Enhance Radiology,” *International Journal of Applied and Advanced Multidisciplinary Research,* vol. 4, no. 6, p. 1626–1644, 2024.

[15] H. Eman, S. M. Adnan, W. Ahmad and I. Mahmood, “Early Detection and Classification of Lung Cancer using,” *International Journal of Innovations in Science & Technology,* vol. 7, no. 1, pp. 479-489, 2025.

[16] A. Elaraby, A. Saad, H. Elmannai, M. Alabdulhafith, M. Hadjouni, and M. Hamdi, “An approach for classification of breast cancer using lightweight deep convolution neural network,” Heliyon, vol. 10, no. 6, 2025.

[17] H. Sadr, M. Nazari, S. Yousefzade-Chabok, H. Emami, R. Rabiei and A. Ashraf, “Enhancing brain tumor classification in MRI images: A deep learning-based approach for accurate classification and diagnosis,” *Image and Vision Computing,* 2025.

[18] A. Y. Forkuo, N. Ikhalea, E. C. Chianumba and A. Y. Mustapha, “Reviewing the Impact of AI in Improving Patient Outcomes through Precision Medicine,” *International Journal of Advanced Multidisciplinary Research and Studies,* vol. 4, no. 6, pp. 1554-1572, 2024.