Dean&Francis

ISSN 2959-6157

Application of Convolutional Neural Networks in Breast Cancer Detection: Hybrid and Attention-based Models

Junshuang Wang

Electronic and Information Engineering, Xiamen University Tan Kah Kee College, Xiamen, China
wangjunshuang8@gmail.com

Abstract:

For breast cancer, a malignant tumor developing from breast epithelial tissue, the limitations of traditional diagnostic methods (e.g., diagnostic errors and invasiveness) create critical challenges, underscoring the urgent need for artificial intelligence-assisted technological development. This paper systematically reviews the applications of Convolutional Neural Networks (CNNs) across the entire workflow of breast cancer, including screening, diagnosis, and prognosis, with a focus on hybrid CNNs such as architectures combining Transformer/ Long Short-Term Memory (LSTM) and CNN-Support Vector Machine (SVM) models and attention-based CNNs such as contour-enhanced attention and cross-attention mechanisms. It analyzes how these models automate feature extraction from medical images to achieve breast lesion detection, benign-malignant differentiation, and prognosis prediction. Results show that hybrid models like Fusion of Hybrid Deep Features (FHDF) achieve over 98% classification accuracy on datasets such as MIAS by fusing features from multiple CNNs, while attention-based models like Convolutional Block Attention Module (CBAM)-Xception attain an Area Under Curve (AUC) of 0.97 and an accuracy of 89.1% in differentiating benign and malignant lesions. However, challenges remain, including insufficient interpretability, cross-institutional data heterogeneity, and privacy risks. The study proposes integrating medical expert systems and applying transfer learning and domain adaptation techniques to enhance model reliability and generalizability, promoting the translation of CNN technologies into clinical practice and constructing a precise and trustworthy AI-driven breast cancer diagnosis and treatment system.

Keywords: CNNs; breast cancer diagnosis; hybrid models and attention mechanisms.

1. Introduction

Breast cancer is defined as a malignant neoplasm that originates from the epithelial tissue of the breast. It occurs due to genetic mutations in breast cells, leading to abnormal and uncontrolled cell proliferation that forms a neoplastic mass. These mutated cells can locally invade adjacent tissues and organs, or metastasize distantly through the lymphatic or circulatory systems. Breast cancer causes extensive harm. Without treatment, it can progress to cachexia and death. Additionally, it inflicts heavy psychological stress, triggering emotional distress, anxiety, and depression. Traditional diagnostics (clinical exams, imaging like mammography/ultrasound/MRI, and biopsies) have key disadvantages: clinical palpation misses small/deep lesions due to subjectivity; imaging faces issues like mammography's poor dense-breast resolution and radiation risks, ultrasound's microcalcification insensitivity, and MRI's high cost/false positives; biopsies risk under-sampling (fine-needle) or invasiveness/complications (core/surgical). These highlight the urgent need for advanced assisted diagnostic technologies such as artificial intelligence algorithms.

Artificial Intelligence (AI), particularly Convolutional Neural Networks (CNNs), has brought breakthroughs to breast cancer diagnosis and treatment. With their ability to automatically extract features, CNNs have been applied to medical image classification and segmentation tasks, such as lung lesion identification and fundus disease diagnosis. In breast cancer research, technologies like Deep Belief Networks (DBNs) improve the efficiency of histopathological image classification and reduce manual misdiagnosis rates [1-3]. For example, deep CNNs have achieved 88% accuracy in breast abnormality classification, and CNN-based analysis of high-resolution images outperforms traditional machine learning models in precision. Standard CNN-based breast cancer classification methods rely on prior knowledge. Although they excel in the local feature extraction, they lack global perception. Transformer, while capable of capturing global dependency relationships, requires massive data and has insufficient low-level feature extraction. Compared with models such as VGG-19, ResNet-50, and ViT, hybrid architectures are significantly superior in terms of indicators, verifying the effectiveness of fusing local and global features [4]. Given the computational intensity of Transformers and the difficulty of optimizing efficiency and feature representation in 3D segmentation, HCMA-UNet—a hybrid architecture—is proposed. It employs an encoder-decoder framework to integrate CNN's local feature extraction and Mamba's global modeling capabilities. It is a lightweight and high-precision automatic segmentation algorithm that reduces clinical dependence on manual segmentation [5]. Current research still needs to address: significant variations in datasets, architectures, and training methods across studies, leading to poor comparability of results; and the need to enhance the reliability and interpretability of models for clinical use. Therefore, a systematic review of CNN applications in the full workflow of breast cancer (screening, diagnosis, and prognosis) is crucial to unify research foundations and accelerate technological translation into clinical practice.

Given the rapid evolution of convolutional neural networks (CNNs) in breast cancer care, this paper systematically reviews their applications across the full breast cancer workflow (screening, diagnosis, prognosis). The study first classifies CNN-based technical strategies, including hybrid CNN and attention-based CNN architectures, and evaluates their efficacy in breast lesion detection, benign-malignant differentiation, and prognosis prediction. It then deeply analyzes the core bottlenecks restricting clinical translation: inconsistent cross-institutional data annotation standards, insufficient generalization of models to rare subtypes (such as triple-negative breast cancer), and the „black-box" nature of deep models that renders the extraction logic of pathological features (e.g., mass margins, calcification patterns) uninterpretable. In response to these challenges, the paper proposes future development pathways: establishing multi-center standardized datasets to bridge data heterogeneity, developing interpretable CNN architectures embedded with medical rules (such as hybrid models integrating expert systems), and enhancing model reliability through large-scale clinical cohort validation. These explorations will propel CNN technologies from the laboratory to clinical practice, constructing a precise and trustworthy AI-driven breast cancer diagnosis and treatment system, and ultimately improving patient survival outcomes.

2. Introduction of CNNs and Related Variants

2.1 Preliminaries of CNN

CNNs are built with a series of specialized layers including convolutional layers, pooling layers, activation layers, normalization layers and fully connected layers. Convolutional layers utilize trainable filters to slide over input data, detecting local features like edges, textures, and shapes by computing dot products. Pooling layers (max or average) then downsample these feature maps, reducing spatial dimensions to cut computation and enhance translation invariance. Activation layers, such as Rectified

ISSN 2959-6157

Linear Unit (ReLU), introduce nonlinearity, allowing the network to learn complex, hierarchical representations—critical for capturing real - world patterns. Normalization layers (e.g., Batch Normalization) stabilize training by normalizing inputs, mitigating internal covariate shift. Fully - connected layers at the end integrate global information from across the network, mapping features to final outputs for tasks like classification or segmentation.

2.2 Attention-Based CNN

2.2.1 Contour-enhanced attention

Karthik et al. proposes a Contour-Enhanced Attention CNN [6]. The framework adopts an encoder-decoder architecture, where the encoder uses a Multi-Kernel Encoding (MKE) module with multi-branch convolutions to extract multi-scale features. The Cross-Context Attention Fusion (CCAF) upsampler employs a query-key-value mechanism to extract structural details from auxiliary encoder features, enabling lossless reconstruction of high-resolution features. The decoder first extracts CT contour regions via edge detection and threshold segmentation, generates convolutional features through a Contour Feature Extraction (CFE) network, and then fuses boundary/shape cues from contours with deep semantic features using pixel-level attention within 3×3 windows to refine segmentation. Experiments on MosMedData and Jun Ma datasets, followed by preprocessing and data augmentation, are trained with the Adam optimizer on dual GPUs, achieving a Dice coefficient of 85.43%. This approach addresses the challenges of blurred boundaries and complex morphologies in infected areas, providing a reliable solution for clinical auxiliary diagnosis.

2.2.2 Cross-attention

A total of 160 samples were collected and grouped by feature dimensions. After preprocessing including region annotation and standardization, 5-fold cross-validation (80% training / 20% testing) was performed, with the model trained for 200 epochs using the RMSprop optimizer. Performance was evaluated using metrics such as accuracy and AUC.

The model framework consists of: a three-branch architecture (SFEpath for shallow feature extraction, LTTpath1/2 for multi-scale deep feature extraction via asymmetric convolutions with residual connections); a cross-attention module fusing spatial/channel attention to enhance critical feature extraction; and a deep feature fusion module that integrates multi-branch outputs through concatenation and dimensionality reduction. Key innovations include: collaborative enhancement of feature discrimination through multi-branch and cross-attention mechanisms; optimization of classification performance by integrating external feature parameters; and a lightweight architecture that achieves comparable accuracy to traditional CNNs with significantly improved computational efficiency [7].

2.2.3 A deep spatial attention

Lu et al. introduced SAFNet, an innovative breast cancer detection framework utilizing ultrasound imaging, which employs a ResNet-18 architecture with spatial attention modules as its foundational network through transfer learning. When extracting image features, ResNet-18 is first pre-trained on ImageNet and fine-tuned on the ultrasound dataset. It is between the ReLU and multiplication layers that the spatial attention module is placed, leveraging pooling and fusion techniques to refine feature extraction. In the classification stage, three randomized neural networks, including Extreme Learning Machine (ELM), Random Vector Functional-Link Network (RVFL), and Schmidt Neural Network (SNN), are trained as classifiers. A late fusion mechanism with majority voting is used to integrate the prediction results of the three, and 5-fold cross-validation is adopted to evaluate the model performance. Its innovations lie in combining the spatial attention module with ResNet-18 to enhance feature extraction ability; using three randomized neural networks as classifiers to avoid overfitting and enable efficient training; and introducing a majority voting fusion strategy to enhance classification stability. Experiments show that SAFNet achieves excellent results on a public ultrasound dataset, with an average accuracy of 94.10%, outperforming four existing methods. Grad-CAM visualization demonstrates that it can accurately locate lesion areas, providing an effective tool for clinical breast cancer diagnosis [8].

2.2.4 Large separable kernel attention

The LSKA module makes Vision Attention Networks function faster by changing the LKA process. It replaces the 2D depth-wise convolution with successive 1D kernels along orthogonal spatial dimensions, which are then combined by hybrid pooling-convolution operations. The primary innovation of this work is the deconstruction of LKA's 2D depth-wise convolution into a sequence of 1D horizontal and vertical kernels. These kernels are then combined through convolution fusion and average/max pooling. This direct replacement of standard depth-wise convolution reduces computational complexity and memory overhead. The architecture uses a pre-trained ResNet-18 backbone with embedded LSKA modules for feature extraction, generates attention maps via 1×1 convolutions, and integrates three randomized neural network classifiers through majority voting. Performance is evalu-

ated using 5-fold cross-validation across image classification, object detection, and semantic segmentation tasks.
Key innovations include resolving LKA‘s quadratic computational growth with kernel decomposition, enabling the model to focus on object shapes over textures as kernel size increases, thereby enhancing robustness to image corruptions. LSKA maintains comparable performance to LKA while significantly reducing parameters and computations, offering an efficient solution for large-kernel convolutions in vision tasks [9].

2.2.5 Cascaded dual attention

The DA-CNN+Bi-GRU framework developed by Ullah et al. combines the attention CNN and the recurrent network, and uses a 7:3 ratio for training and validation on five public datasets.The Adam optimizer is applied for 300 epochs with a batch size of 16 and a sequence length of 16 frames, evaluating performance via accuracy and FPS.
The overall framework consists of three parts: a lightweight CNN architecture (8 convolutional layers with up to 64 3×3 kernels) for spatial feature extraction, a dual attention module (fusing channel and spatial attention) to enhance focus on action regions, and a bi-directional GRU network (3-layer bidirectional structure) to capture long-term temporal patterns. Innovations lie in designing a lightweight dual attention CNN to focus on human key areas through channel-spatial attention mechanisms, combining with bi-directional GRU to enable forward-backward gradient learning for enhanced temporal modeling. The framework requires only 5.4MB storage, achieves 300FPS inference speed on GPU, outperforms existing methods on multiple datasets, and balances accuracy and efficiency [10].

2.3 Hybrid CNN

2.3.1 Hybrid CNN-MLP model

Linear and angular acceleration signals are acquired via shaft-mounted wireless sensors. The linear signals undergo Hilbert-Huang Transform (HHT) to generate 32×32 time-frequency images as inputs for a CNN, while the angular acceleration signals are processed using Fast Fourier Transform (FFT) to compute signal power at the first two torsional mode frequencies (N1, N2) as inputs for an MLP, forming a „CNN+MLP“ hybrid architecture. The CNN branch employs three convolution-pooling layers for feature extraction, and the MLP branch performs two fully connected layers on N1 and N2, with final classification achieved through feature fusion. Key innovations include the first use of shaft-mounted sensors for high-sensitivity data acquisition, the design of a hybrid model architecture for heterogeneous input processing, and the integration of HHT time-frequency images with signal power features to enhance feature representation capabilities [11].

2.3.2 CNN-LSTM-Transformer model

The self-organizing map (SOM) algorithm’s topological mapping capability is exploited for seasonal clustering of input data, enabling the retention of data distribution patterns in a lower-dimensional space.The preprocessing pipeline includes z-score normalization for feature scale standardization, followed by temporal segmentation to align with seasonal periodicity. The CNN component, consisting of 4 layers with 3×3 kernels and ReLU activations, extracts hierarchical spatial features through alternating convolution-max pooling operations, while a 2-layer LSTM with 256 units captures long-range temporal dependencies via memory cell structures.
The Transformer module integrates 8-head multi-head self-attention mechanisms with probe-sparse operations, dynamically pruning irrelevant attention connections to reduce computational complexity by 40% compared to standard architectures. A generative decoder employs autoregressive prediction for multi-step output optimization. The framework comprises three functional modules: data collection with adaptive sampling and noise filtering;cross-modal feature fusion of CNN-extracted spatial maps and LSTM-encoded temporal sequences; model evaluation using point-wise and distribution-based metrics [12].

2.3.3 CNN-SVM model

The hybrid CNN-SVM model adheres to a systematic workflow of „preprocessing-segmentation-feature extraction-classification“. It commences by downsampling original images to 32×32 resolution, followed by median filtering to suppress noise artifacts and histogram equalization for contrast enhancement. Otsu‘s thresholding algorithm is applied to binarize grayscale images, dynamically determining optimal cutoff values to isolate regions of interest. This CNN model consists of four layers, each layer is alternately composed of 3×3 convolutional operations and 2×2 stride max - pooling operations. ReLU is used as the activation function to automatically learn multi - level spatial features. The penultimate fully connected layer projects these features into a compact vector space, which serves as input to a multi-class SVM classifier. Leveraging kernel functions (e.g., RBF or polynomial), the SVM optimizes hyperplane margins through structural risk minimization, enabling robust classification even with limited labeled data. This integration of CNN‘s feature engineering capabilities and SVM‘s generalization prowess yields efficient handling of non-linear patterns while mitigating overfitting [13].

ISSN 2959-6157

2.3.4 CNN-BiLSBVM model

A hybrid CNN-BiLSTM model is proposed for long-term sequence forecasting. The model preprocesses nine predictor variables, including various time-series and categorical features, normalizing and reshaping them into a 9×4 matrix. A CNN layer with 256 2×2 kernels and max pooling extracts spatial features, which are flattened and fed into a 500-unit BiLSTM layer to capture bidirectional temporal dependencies. A dropout layer (rate 0.5) prevents overfitting, and a fully connected layer generates multi-step predictions. The innovation lies in fusing CNN's noise filtering/feature extraction with BiLSTM's bidirectional temporal context learning, forming an integrated „feature extraction-temporal modeling-prediction" framework for complex sequence forecasting tasks [14].

3. The Application and Discussion of CNNs in Breast Cancer Detection

3.1 CNN-Based Approaches in Breast Cancer

3.1.1 Hybrid CNN model

For breast cancer detection, the HCPELM model innovatively integrates a CNN feature extractor and a pruned integrated extreme learning machine classifier. By applying the ReLU activation function for data analytics enhancement, the model further includes preprocessing to eliminate artifacts and pectoral muscles from mammogram images. Combine the spatial feature extraction ability of the convolutional network and the non - linear classification characteristics of the fully - connected layer, and cooperate with the transfer learning mechanism of layer freezing to achieve parameter optimization. HCPELM, trained on the MIAS database, achieves an 86% accuracy rate in breast image recognition. This performance exceeds benchmark deep learning models and highlights its utility for early breast cancer diagnosis [15].

For the multi - classification problem of breast cancer, researchers have developed the FHDF framework. First, pre-processing such as pectoralis muscle removal and image enhancement is carried out on mammogram images. Mammogram images undergo pre-processing, which includes the elimination of the pectoralis muscle and the enhancement of the image. The approach outperforms single CNN models and traditional late fusion methods, demonstrating the effectiveness of multi-feature collaboration in identifying complex lesions [16].

3.1.2 Attention-Based CNN

To improve the accuracy of breast cancer diagnosis, researchers skillfully converted the CBAM attention module into three CNN architectures (DenseNet121/Xception/ResNet50) for development. The study involved 1, 239 patients across multiple centres, with region-of-interest (ROI) segmentation performed by experienced radiologists. The CBAM-Xception model demonstrated superior performance, achieving an area under the ROC curve (AUC) of 0.970, 84.8% sensitivity, 100% specificity, and 89.1% accuracy on the external test set. This outperformed conventional radiomics models and two radiologists, while also improving inter-rater agreement when assisting radiologists in breast cancer assessment. The clinical application value of this model in non-invasive thoracic. Confirm the classification of lesions through visual inspection. Introduce the main diagnostic areas of breast cancer for use [17].

An interpretable attention mechanism deep learning model has been created by the Anari research team. By combining the UNet architecture with ResNet-18, DenseNet-121, and EfficientNet-B0 encoders, it achieves accurate segmentation of breast cancer tumors. The model incorporates CBAM and Non-Local Attention mechanisms to enhance focus on tumor regions, using Depthwise Separable Convolutions to reduce computational complexity. Trained on the BUSI dataset, it achieved a Dice score of 0.6140 and AUC of 0.97, outperforming state-of-the-art models like Swin-UNet and ADU-NET. Grad-CAM visualizations validated its ability to accurately highlight breast cancer tumor areas, demonstrating clinical potential for improving segmentation accuracy in breast cancer diagnosis [18].

3.2 Discussion

Although many progresses have been achieved for CNN-based classification models such as hybrid CNN and attention-based CNN, several challenges including insufficient interpretability, limited generalizability and privacy risks should be considered.

Insufficient Interpretability: Both models struggle to explain to clinicians how they extract key pathological features (such as mass margins and calcification distributions) from breast imaging. Their „black-box" nature obscures the logic behind diagnostic decisions, making it difficult for medical professionals to trust or validate the models' conclusions.

Limited Generalizability: Breast imaging data vary significantly across hospitals due to differences in equipment models, imaging parameters, and annotation standards. This heterogeneity causes models to underperform in cross-institutional applications, particularly for rare breast cancer subtypes (e.g., triple-negative breast cancer) with small sample sizes. Additionally, as pathological diagnostic standards evolve with new research, models require frequent retraining, limiting their adaptability and practicality.

Privacy Risks: Breast imaging contains sensitive patient information, and traditional model training relies on centralized data storage, posing significant leakage risks. While technologies like federated learning enable decentralized training, gradient data in cross-institutional collaborations remain vulnerable to adversarial reconstruction attacks. Adding privacy-preserving techniques (e.g., differential privacy) often degrades model performance, creating a trade-off between privacy protection and diagnostic accuracy.

To overcome the bottlenecks in the application of Hybrid CNN and attention-based CNN in breast cancer, there are two key development directions. On one hand is the integration of expert systems and domain knowledge. An expert system is like translating the diagnostic experience and rules accumulated by medical experts over the years—such as „irregular mass margins with clustered calcifications suggest malignancy“—into program logic that computers can understand, which is then combined with CNN. This makes the model no longer a black box, allowing doctors to understand how it makes judgments. The attention mechanism in the model is guided by this knowledge, which helps the model focus on image regions with true diagnostic value and elevates the accuracy and credibility of diagnoses. On the other hand, there is the application of transfer learning and domain adaptation techniques. Transfer learning is akin to enabling the model to „stand on the shoulders of giants“: first, pre-training the model on large amounts of other medical imaging data to learn general image feature extraction capabilities, and then fine-tuning it on breast cancer data from specific institutions to reduce reliance on small amounts of data from a single institution. Domain adaptation techniques aim to address data differences caused by varying equipment and imaging parameters across hospitals. By adjusting the model's feature representation or training methods, these techniques ensure the model performs stably on data from different institutions, particularly for accurately diagnosing rare subtypes like triple-negative breast cancer. Additionally, incremental learning and other methods allow the model to automatically optimize as medical diagnostic standards evolve. The synergistic effort of these two types of technologies can balance interpretability, generalizability, and privacy protection, propelling AI diagnostic technologies for breast cancer toward more practical and reliable clinical applications.

4. Conclusion

This article systematically investigates hybrid CNN and attention-based CNN models to tackle diagnostic challenges in breast cancer. These frameworks enable automated feature extraction from medical images, demonstrating superior performance over traditional approaches. However, limitations like black-box interpretability, cross-hospital data variability, and privacy risks in centralized training were identified. To overcome these, this article advocates integrating expert medical knowledge into model architectures for transparency and applying transfer learning with domain adaptation to enhance generalizability across institutions. These strategies aim to develop robust, interpretable AI tools that balance diagnostic accuracy with privacy protection, fostering clinical adoption for improved breast cancer care.

References

[1] Li S, Kou P, Ma M, Yang H, Huang S, Yang Z. Application of semi-supervised learning in image classification: Research on fusion of labeled and unlabeled data. IEEE Access, 2024.

[2] Liu Y, Yang H, Wu C. Unveiling patterns: A study on semi-supervised classification of strip surface defects. IEEE Access, 2023, 11: 119933–119946.

[3] Hirra I, et al. Breast cancer classification from histopathological images using patch-based deep learning modeling. IEEE Access, 2021, 9: 24273–24287.

[4] Zeynali A, Tinati MA, Mozaffari Tazehkand B. Hybrid CNN-Transformer architecture with Xception-based feature enhancement for accurate breast cancer classification. IEEE Access, 2024.

[5] Li H, et al. HCMA-UNet: A hybrid CNN-Mamba UNet with inter-slice self-attention for efficient breast cancer segmentation. arXiv preprint arXiv:2501.00751, 2025.

[6] Karthik R, Menaka R, Won D. Contour-enhanced attention CNN for CT-based COVID-19 segmentation. Pattern Recognition, 2022, 125: 108538.

[7] Sun L, et al. Cross-attention multi-branch CNN using DCE-MRI to classify breast cancer molecular subtypes. Frontiers in Oncology, 2023, 13: 1107850.

[8] Lu SY, Wang SH, Zhang YD. SAFNet: A deep spatial attention network with classifier fusion for breast cancer detection. Computers in Biology and Medicine, 2022, 148: 105812.

[9] Lau KW, Po LM, Rehman YAU. Large separable kernel attention: Rethinking the large kernel attention design in CNN. Expert Systems with Applications, 2024, 236: 121352.

[10] Ullah H, Munir A. Human activity recognition using cascaded dual attention CNN and bi-directional GRU framework. Journal of Imaging, 2023, 9(7): 130.

ISSN 2959-6157

[11] Sinitsin V, et al. Intelligent bearing fault diagnosis method combining mixed input and hybrid CNN-MLP model. Mechanical Systems and Signal Processing, 2022, 180: 109454.
[12] Al-Ali EM, et al. Solar energy production forecasting based on a hybrid CNN-LSTM-transformer model. Mathematics, 2023, 11(3): 676.
[13] Khairandish MO, et al. A hybrid CNN-SVM threshold segmentation approach for tumor detection and classification of MRI brain images. IRBM, 2022, 43(4): 290–299.
[14] Méndez M, Merayo MG, Núñez M. Long-term traffic flow forecasting using a hybrid CNN-BiLSTM model. Engineering Applications of Artificial Intelligence, 2023, 121: 106041.
[15] Sureshkumar V, Prasad RSN, Balasubramaniam S, et al. Breast cancer detection and analytics using hybrid CNN and extreme learning machine. Journal of Personalized Medicine, 2024, 14(8): 792.
[16] Chakravarthy S, Bharanidharan N, Khan SB, et al. Multi-class breast cancer classification using CNN features hybridization. International Journal of Computational Intelligence Systems, 2024, 17(1): 191.
[17] Mao N, Zhang H, Dai Y, et al. Attention-based deep learning for breast lesions classification on contrast enhanced spectral mammography: A multicentre study. British Journal of Cancer, 2023, 128(5): 793–804.
[18] Anari S, Sadeghi S, Sheikhi G, et al. Explainable attention based breast tumor segmentation using a combination of UNet, ResNet, DenseNet, and EfficientNet models. Scientific Reports, 2025, 15(1): 1027.