IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 30 May 2025, accepted 30 June 2025, date of publication 10 July 2025, date of current version 16 July 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3586139

RESEARCH ARTICLE

Token Mixing for Breast Cancer Diagnosis: Pre-Trained MLP-Mixer Models on Mammograms

HOSAMELDIN O. A. AHMED AND ASOKE K. NANDI, (Life Fellow, IEEE)
Department of Electronic and Electrical Engineering, Brunel University of London, UB8 3PH Uxbridge, U.K.

Corresponding author: Asoke K. Nandi (asoke.nandi@brunel.ac.uk)

This work was supported in part by the Brunel University of London Research Funding Scheme.

ABSTRACT Breast cancer remains a leading cause of mortality among women, necessitating accurate and computationally efficient diagnostic solutions. Deep learning, particularly convolutional neural networks (CNNs), has significantly advanced mammographic analysis by automating feature extraction and improving early detection. However, CNNs rely on localised feature extraction, limiting their ability to capture long-range dependencies essential for robust classification. This study introduces and evaluates the effectiveness of pre-trained MLP-Mixer models using transfer learning as an alternative to CNN-based approaches, utilising their token-mixing and channel-mixing mechanisms to integrate local and global spatial features in mammograms. Four MLP-Mixer variants (B/16, L/16, B/32, and L/32) were systematically assessed on three benchmark datasets: CBIS-DDSM, INbreast, and MIAS. The results demonstrate that MLP-Mixer models, particularly those with smaller patch sizes (L/16 and B/16), consistently achieve state-of-the-art accuracy and sensitivity, while also offering 30 – 50% faster inference times compared to leading CNNs such as ResNet and DenseNet. These models demonstrate strong generalisation across multiple benchmark datasets and strike an effective balance between diagnostic accuracy and computational efficiency, which are essential requirements for clinical deployment. Their performance underscores the importance of fine-grained feature extraction in mammographic analysis. Comparative results indicate that MLP-Mixer models offer a compelling alternative to conventional CNNs by efficiently capturing both local and global dependencies without the high computational demands of deep convolutional network architectures. These findings highlight the promise of token-based models for AI-assisted breast cancer diagnosis and suggest that MLP-Mixer architectures are well-suited for real-time medical imaging applications. By enabling direct global spatial interaction, reducing architectural complexity, and improving diagnostic precision across varied imaging conditions, MLP-Mixers offer a computationally efficient alternative to traditional CNNs without compromising accuracy.

INDEX TERMS Breast cancer diagnosis, computer-aided diagnosis, deep learning, pretrained convolution neural network models, pre-trained multi-layer perceptron (MLP)-mixer models, mammography.

I. INTRODUCTION

Breast cancer is a major global health challenge, with around 2.3 million new cases diagnosed and 685,000 deaths in 2020, followed by approximately 670,000 deaths in 2022 [1], [2]. Being the most common cancer in women, it requires ongoing awareness, early detection, and improved treatment approaches. Breast cancer staging ranges from Stage 0 (in situ, confined to ducts/lobules) to Stage IV (metastatic, spread to distant organs). Lower stages (I-III) involve increasing tumour size and lymph node spread, with a better prognosis at earlier stages. Treatment depends on stage and tumour biology [3], [4]. In Canada, 5-year survival for stage I breast cancer is 99.8%, falling to 23.2% for stage IV [5]. U.S. data (2012-2018) shows similar trends, with stage I survival at over 99% and stage IV at 29% [6]. One Netherlands study (1989-2017) confirms improved survival and lower mortality due to early detection and advanced treatments [7]. The significant decline in survival rates from early to advanced stages highlights the critical importance

The associate editor coordinating the review of this manuscript and approving it for publication was Vishal Srivastava.

of early detection and timely intervention to improve patient outcomes.

Diagnostic medical imaging has significantly expanded the tools for breast cancer detection, including mammography, breast thermography, magnetic resonance imaging (MRI), ultrasound, positron emission tomography (PET), histopathology, and computed tomography (CT) [8]. Of these, mammography remains a basis of early detection, particularly improving treatment success and survival rates by identifying abnormalities before symptoms become apparent, even in low and middle-income countries with limited healthcare access [9], [10], [11], [12]. Digital mammography offers high diagnostic accuracy, matching traditional screen-film mammography (SFM) while excelling in detecting tumours in dense breast tissue [13]. However, the screening process has limitations, including high costs, procedural complexity, false positives, and human error. Additionally, existing imaging techniques have inherent drawbacks; for instance, up to 35% of breast cancer may be missed during screening, often due to dense or overlapping breast tissue, leading to interval cancers detected between screenings [14], [15].

Integrating advanced technologies is crucial for overcoming limitations in breast cancer screening by improving accuracy, efficiency, and patient outcomes. AI-powered diagnostic systems are at the forefront of this transformation, providing invaluable support to physicians. Utilising advanced algorithms, these tools detect subtle patterns in medical data that are often missed by human analysis. This improved analytical precision reduces false positives, providing clinicians with more reliable initial evaluations. Moreover, accuracy is essential in diagnostics, enabling informed decisions about patient treatment and management. By enhancing decision-making, these technologies not only support clinicians but also improve patient care and health outcomes, advancing the efficiency of healthcare systems [16].

Computer-aided diagnosis (CAD) systems have transformed mammogram-based breast cancer detection by overcoming limitations in manual interpretation, including high error rates, inconsistencies, and challenges posed by dense breast tissue. By enhancing image quality, automating feature detection, and utilising advanced technologies such as deep learning, CAD improves sensitivity and specificity in identifying abnormalities such as microcalcifications and masses. As a ''second opinion'' for radiologists, CAD reduces diagnostic errors, alleviates workload, and supports early cancer detection, which is necessary for effective treatment and improved survival rates. Research demonstrates that CAD-integrated mammography enhances diagnostic accuracy, lowers costs, and reduces the need for double readings, solidifying its role as an essential tool in contemporary breast cancer screening [17], [18], [19], [20].

Traditional Machine Learning (ML) techniques have significantly advanced CAD systems for breast cancer detection using mammography. Support Vector Machines (SVM) are widely utilised for binary classification tasks, often combined with feature extraction methods such as wavelets and texture analysis. Artificial Neural Networks (ANNs), including Multi-Layer Perceptron (MLP), are also prominent but require careful parameter tuning. Decision Trees (DT) and Random Forests (RF) offer interpretable and robust classification, while ensemble methods such as AdaBoost enhance performance through improved generalisation. Feature extraction techniques, such as texture analysis (e.g., Gray-Level Co-occurrence Matrix) and morphological descriptors, are critical, with dimensionality reduction methods such as Principal Component Analysis (PCA) aiding efficiency. Despite their strengths with small datasets, traditional ML techniques face challenges, including reliance on handcrafted features and limited scalability [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31].

Although traditional techniques were effective, principally with limited annotated data, the field has gradually shifted toward deep learning (DL). Unlike conventional techniques, DL automates feature extraction directly from raw mammogram images, eliminating the need for manual feature engineering. This allows models to learn discriminative features independently. For example, several studies have demonstrated the effectiveness of convolutional neural networks (CNNs) for feature extraction and classification, achieving high accuracy across multiple datasets. They are widely applied in four key areas: breast density classification, asymmetry detection, calcification detection, and mass detection. For breast density classification, CNNs utilise pre-trained networks such as ResNet and VGGNet, with attention-based mechanisms enhancing feature extraction. In asymmetry detection, CNNs analyse bilateral features from Craniocaudal (CC) and Mediolateral-Oblique (MLO) views, using deep residual networks and attention mechanisms to improve performance. Calcification detection benefits from CNNs' ability to segment microcalcifications in high-resolution images, with architectures such as U-Nets and dilated CNNs capturing fine details. Mass detection employs two-stage CNN architectures, combining region-based CNNs for initial detection and classification CNNs for malignancy assessment, while multi-view CNNs integrate CC and MLO views to enhance accuracy. The success of CNNs in mammography is supported by advanced DL architecture such as AlexNet, VGGNet, ResNet, DenseNet, EfficientNet and U-Net, each fitted to specific tasks. Training strategies, including data preprocessing, augmentation, and transfer learning, further optimise performance. Multi-view learning and hybrid approaches, which combine hand-crafted radiomics features with deep learning, have also proven effective. Evaluation metrics such as accuracy, sensitivity, specificity, and AUC-ROC demonstrate CNNs' superior performance over traditional machine learning models, with state-of-the-art models achieving AUC values exceeding 0.95 in mass detection tasks [32], [33], [34].

Despite advancements in CNN-based techniques, several challenges continue to impact performance. Issues such as limited labelled data and class imbalance can hinder effectiveness; however, strategies like data augmentation, synthetic

data generation, and transfer learning can help mitigate these problems. Additionally, the interpretability of CNNs remains a significant concern. Attention mechanisms such as Gradient-weighted Class Activation Mapping (Grad-CAM) and the Shapley additive explanations (SHAP) can partially address this issue [35], [36].

Transfer learning enables models pre-trained on large-scale datasets such as ImageNet to be fine-tuned on domain-specific medical imaging datasets. Using pre-trained CNN models, transfer learning improves model generalisation, accelerates training, and lowers computational resource needs, making it an effective method for mammographic analysis [37]. Various studies have investigated the effectiveness of transfer learning in mammographic analysis. Early studies showed that VGG16, ResNet50, and InceptionV3 were effective for tumour classification, achieving high accuracy and sensitivity in differentiating between benign and malignant cases [38]. For instance, Guan et al. utilised VGG16 for mammographic analysis, achieving 90.5% accuracy on the MIAS and DDSM datasets, while Falconi et al. stated an AUC of 0.844 on CBIS-DDSM using fine-tuned VGG16. Additional studies focused on improving CNN architecture for mammographic classification. Alruwaili et al. examined ResNet50 and NasNet-Mobile, finding that ResNet50 achieved an accuracy of 89.5% on the MIAS dataset. Furthermore, a recent study evaluated five popular CNN pre-trained architectures: VGG19, ResNet50, EfficientNetB4, MobileNetV2, and InceptionV3. It demonstrated differences in performance related to dataset size, computational efficiency, and feature extraction capabilities [36].

While CNNs have advanced mammographic analysis by automating feature extraction and achieving high diagnostic accuracy, their dependence on local spatial features presents a significant limitation. CNNs stand out at capturing localised patterns but may fail to effectively model global dependencies, which are equally critical for accurate breast cancer detection. Mammograms present intricate tissue structures where local and global contextual information are essential for distinguishing between benign and malignant lesions. This highlights the need for approaches that integrate both local and global features to enhance diagnostic precision.

In contrast to CNNs, multi-layer perceptron (MLP) mixers present a novel approach by capturing both spatial and channel-based features through token-mixing and channel-mixing mechanisms. The MLP-Mixer model is architecturally distinct from convolutional and attention-based models. Unlike CNNs, which extract features through local receptive fields and require multiple stacked layers to capture global context, MLP mixers perform token-mixing operations that allow global spatial feature interaction at each layer. This enables more direct modelling of long-range dependencies in mammographic images, which is particularly beneficial in detecting distributed subtle anomalies. This capability makes them particularly promising for addressing the limitations of CNNs, such as identifying understated textural variations and architectural distortions that are critical for accurate diagnosis. While CNN-based pre-trained architectures such as DenseNet, ResNet50, and EfficientNet have demonstrated strong performance in analysing mammograms, their dependence on local feature hierarchies and deep stacking can lead to significant computational complexity and latency. Moreover, in contrast to hybrid models such as MoEffNet or deep ensembles, MLP-mixer models allow for the modelling of both local and global contexts from the early layers through token-mixing MLPs, thereby reducing the need for deep hierarchies and offering a simpler and more unified structure with fewer parameters, low latency, and greater interpretability. These characteristics make it a novel and fundamentally different paradigm for breast cancer detection in medical imaging.

This study represents, to the best of our knowledge, the first comprehensive evaluation of pre-trained MLP-Mixer models for breast cancer detection using mammograms. We propose a novel framework that:

1) Matches or exceeds the diagnostic performance of state-of-the-art CNNs,
2) Reduces inference time by up to 50%.
3) Demonstrates consistent generalisation across multiple mammographic datasets.

Through systematic evaluation of four pre-trained MLP-Mixer variants (B/16, L/16, B/32, and L/32) on three benchmark mammography datasets, CBIS-DDSM, INbreast, and MIAS, this study explores the effectiveness of token-mixing architectures as an alternative to CNN-based models for breast cancer detection. We compare their performance against state-of-the-art CNN baselines using key clinical metrics such as accuracy, sensitivity, specificity, and AUC. By utilising the unique token-mixing and channel-mixing mechanisms of MLP-Mixers, we aim to assess whether these models can enhance diagnostic accuracy, reduce false positives and negatives, and offer improved interpretability. Ultimately, this research contributes to the advancement of scalable, reliable, and real-time AI-driven decision support systems in radiology.

The main contributions of this article are summarised as follows:

- Introduction of MLP-Mixer Architectures for Mammographic Image Classification**:** This study is the first to systematically evaluate pre-trained MLP-Mixer architectures (Mixer-B/16, Mixer-L/16, Mixer-B/32, Mixer-L/32) for breast cancer detection using mammograms across multiple benchmark datasets, including CBIS-DDSM, INbreast, and MIAS. By introducing this token-based MLP approach to breast cancer detection, our work addresses a critical gap in the literature and demonstrates the potential of MLP-Mixer models as a novel alternative to CNNs and Transformers in the analysis of mammographic images.
- Architectural novelty and simplicity: Unlike traditional CNNs and hybrid attention-based models, MLP-Mixers eliminate convolutional and attention layers, relying only on token- and channel-mixing MLPs. This allows

early and direct global spatial interaction across the image, enabling more efficient feature representation and reducing model complexity. This represents a novel architectural paradigm for medical imaging and directly overcomes the limitations of CNNs in modelling long-range patterns critical to detecting subtle and distributed mammographic abnormalities.

- Superior diagnostic performance and generalisation: Extensive experiments on three benchmark mammographic datasets (CBIS-DDSM, INbreast, and MIAS) demonstrate that MLP-Mixer models achieve classification performance (Accuracy, Sensitivity, AUC) that is comparable to or exceeds state-of-the-art CNN and hybrid models, with notably better generalisation across dataset variations.
- Computational efficiency and clinical scalability: MLP-Mixer models achieve faster inference times of 30–50% lower than deep CNNs such as DenseNet and ResNet, while maintaining high diagnostic accuracy. This makes them highly suitable for deployment in real-time and resource-constrained clinical environments, including mobile diagnostics and rural screening settings.
- Comprehensive comparative analysis: The study provides a detailed comparison with existing state-of-the-art methods [34], [50], [51], [52], [53], [54], [55], [56], [57], [58], [59], [60], including traditional, CNN-based, hybrid, segmentation-guided, and ensemble approaches. Our results demonstrate that MLP-Mixer models offer a simpler yet more effective alternative for AI-assisted breast cancer detection, both architecturally and diagnostically.

The remainder of this paper is structured as follows: Section II details the materials and methods, including dataset descriptions, preprocessing techniques, and model architecture. Section III presents experimental results, comparing the performance of MLP-Mixer models against CNN-based models. Finally, Section IV draws some conclusions from this study and outlines potential directions for future research.

II. MATERIALS AND METHODS

A. DESCRIPTION OF THE MLP-MIXER MODEL

The MLP-Mixer represents an innovative neural network architecture designed for image classification, distinguished by its reliance solely on multi-layer perceptron (MLP) rather than convolutions or attention mechanisms. As depicted in Figure 1, this architecture is structured around three key components: patch embedding, mixer layers, and a classification head. MLP-Mixer processes image features along two axes: (1) Channel Mixing, which enables communication between feature channels within individual patches, and (2) Token Mixing, which facilitates communication across spatial locations between different patches. This separation of operations, achieved through dense matrix multiplications, reshaping, and nonlinearities, sets the MLP-Mixer apart from traditional convolutional or attention-based models [39], [40], [41].

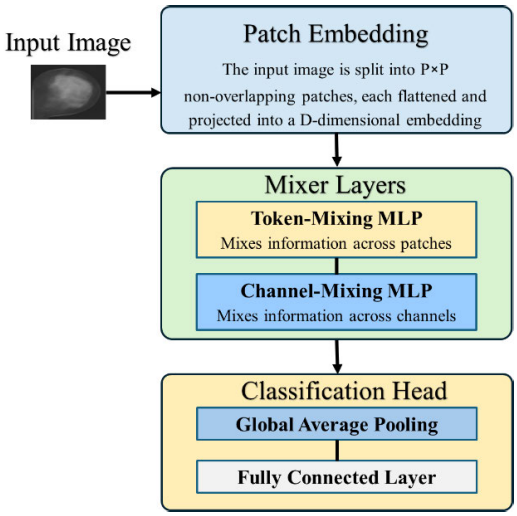

FIGURE 1. Overview of the MLP-mixer architecture for image classification.

1) PATCH EMBEDDING

The architecture begins with patch embedding, where an input image of the size $H \times W \times C$ (height, width, and channels) is divided into non-overlapping patches of size $P \times P$. Each patch is flattened and projected into a fixed-dimensional embedding vector of size D. If H and W are divisible by P, the number of patches N is computed as follows:

$$N = \frac{H.W}{P^2} \tag{1}$$

These patches are then linearly projected into a D-dimensional embedding space, forming a patch embedding table Z_0 that can be represented as follows:

$$Z_0 = [z_1, z_2, \ldots, z_N] \in \mathbb{R}^{NxD} \tag{2}$$

Here each $z_i \in \mathbb{R}^D$ corresponds to the embedding of the *ith* patch extracted from an image. This patch embedding, $\{z_1, z_2, \ldots, z_N\}$, forms the input sequence used for processing in vision models, with each z_i encoding spatial and structural information from its respective region.

2) MIXER LAYERS

The core of the MLP-Mixer lies in its mixer layers, which consist of two primary components: the *Token-Mixing MLP* and the *Channel-Mixing MLP*. These components operate sequentially, with skip connections to preserve input signals and ensure stable gradient flow during training. The *Token-Mixing MLP* operates on the columns of the patch embedding table, mixing information spatially across patches while keeping channels independent. Mathematically, this is expressed as:

$$Z' = Z_0 + W_2.GELU\left(W_1.Norm\,(Z_0)^T\right)^T \tag{3}$$

Here $Norm\,(Z_0)$ denotes layer normalisation, W_1 and W_2 are learnable weight matrices of the MLP, and GELU (Gaussian Error Linear Unit) serves as the activation function.

The GELU activation function can be defined as follows:

$$GELU\ (x) = x.\Phi\ (x) \tag{4}$$

Here $\Phi\ (x)$ is the cumulative distribution function (CDF) of the standard normal distribution that provides smooth and continuous activation and can be represented as follows:

$$\Phi\ (x) = \frac{1}{2}\left(1 + erf\left(\frac{x}{\sqrt{2}}\right)\right) \tag{5}$$

Here $erf\left(x/\sqrt{2}\right)$ is the error function. $GELU$ often approximated for computational efficiency as:

$$GELU\ (x) \approx 0.5\ x(1 + \tanh(\sqrt{\frac{2}{\pi}}(x + 0.044715x^3) \tag{6}$$

The *Channel-Mixing MLP* operates across channels, mixing information within each patch while keeping patches independent, such that

$$Z'' = Z' + W_4.GELU\left(W_3.Norm\left(Z'\right)\right) \tag{7}$$

Here, W_3 and W_4 are learnable weight matrices for channel mixing. Each mixer layer applies these two MLP blocks sequentially, with skip connections ensuring robust training dynamics.

3) CLASSIFICATION HEAD

After processing through N mixer layers, the final output Z_{final} undergoes global average pooling to aggregate information across patches, generating a pooled representation h:

$$h = \frac{1}{N}\sum_{i=1}^{N} Z_{final,i} \tag{8}$$

This representation is then mapped to class probabilities using a fully connected layer:

$$y = Softmax\ (W_c.h) \tag{9}$$

Here W_c is the weight matrix of the classification layer.

The MLP-Mixer offers several strengths, including simplicity, as it avoids the complexity of convolutions or self-attention mechanisms, relying instead on lightweight operations like matrix multiplication and normalisation. Its positional invariance is achieved through token mixing, which essentially encodes spatial relationships without requiring explicit positional embedding. The architecture is also scalable, with computational complexity scaling linearly with the number of patches, $O\ (N)$, unlike vision transformers, which scale quadratically, $O\left(N^2\right)$. Finally, the MLP mixer demonstrates flexibility, performing competitively on large and small datasets when combined with regularisation and data augmentation techniques. These features make the MLP mixer a potentially powerful alternative to traditional deep-learning models for image classification tasks.

B. DETAILED DESCRIPTION OF PRE-TRAINED MLP MIXER ARCHITECTURES: B/16, B/32, L/16, AND L/32

Four pre-trained MLP-Mixer variants were explored in this study, including B/16, B/32, L/16, and L/32. These models provide varying levels of flexibility in model size, depth, patch resolution, and computational efficiency, allowing a balance between performance and resource requirements. These configurations enable the architecture to scale effectively for different applications, balancing computational cost with predictive accuracy. The following sections provide a detailed overview of these configurations and their implications for breast cancer detection tasks.

Table 1 presents a comparative overview of different MLP-Mixer model configurations, highlighting their key architectural parameters. The models vary in patch resolution (P × P), hidden size (C), number of layers, and the MLP dimensions used for channel and token mixing. The sequence length (S), which depends on the patch resolution, affects how spatial features are processed, while the total number of parameters (in millions) indicates the computational complexity of each model.

TABLE 1. Summary of MLP mixer architectures: B/16, B/32, L/16, and L/32.

Model	B/16	B/32	L/16	L/32
Patch Resolution (P×P)	16×16	32×32	16×16	32×32
Hidden Size (C)	768	768	1024	1024
Number of Layers	12	12	24	24
MLP Dimension (Channel Mixing, $\boldsymbol{D_C}$)	3072	3072	4096	4096
Sequence Length (S)	196	49	196	49
MLP Dimension (Token Mixing, $\boldsymbol{D_S}$)	384	384	512	512
Parameters (Millions)	59	60	207	206

The MLP-Mixer architecture processes images by dividing them into patches, with patch resolution (P × P) defining the number and size of these patches. Models with smaller patch resolutions (16 × 16), such as B/16 and L/16, produce a higher sequence length (S = 196), enabling finer-grained feature extraction at the cost of increased computation, whereas larger patch resolutions (32 × 32), as in B/32 and L/32, reduce sequence length (S = 49) and computational demands but may lose spatial granularity. The B/16 and B/32 models feature a hidden size of 768 with 12 layers, while the L/16 and L/32 models employ a larger hidden size of 1024 with 24 layers, allowing for more complex feature representation and deeper hierarchical learning. The MLP dimensions (D_C and D_S) govern two essential operations: channel-mixing MLP (D_C), which enhances feature interactions across channels, and token-mixing MLP (D_S), which enables spatial feature integration across tokens. The L-series models have larger MLP dimensions (4096 for channel mixing and 512 for token mixing) compared to the B-series models (3072 and 384, respectively), further contributing to their increased capacity and number of parameters. Larger models, such as L/16 (207M parameters), offer higher

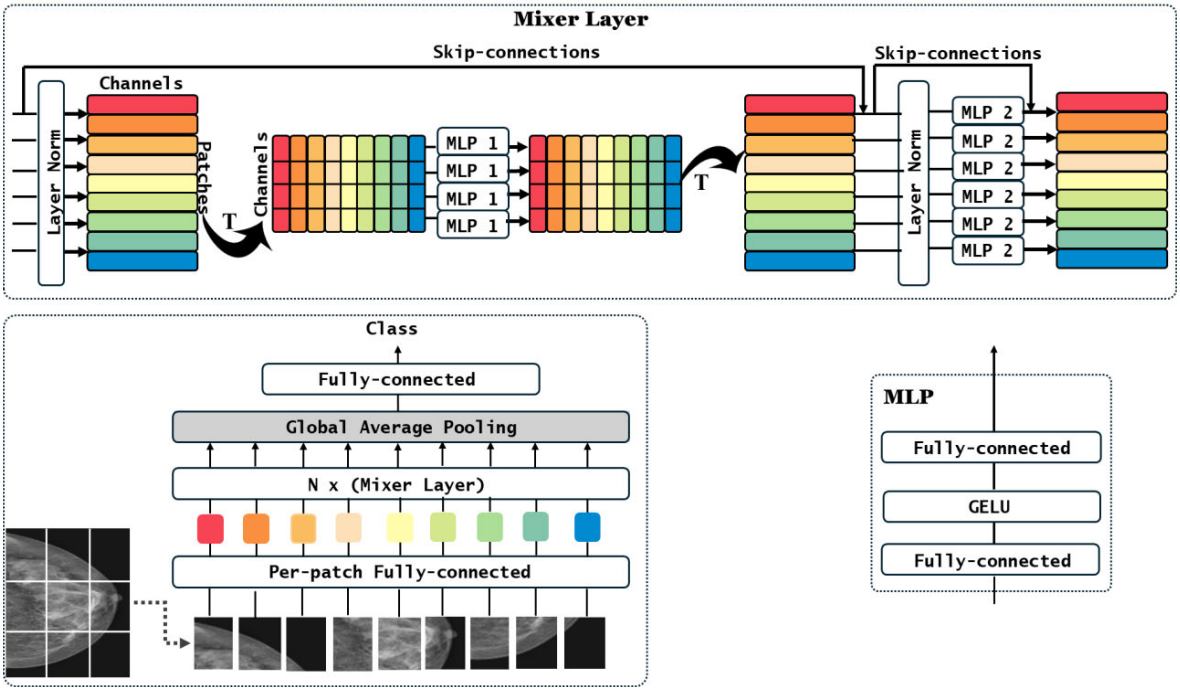

FIGURE 2. MLP-mixer-based framework for breast cancer detection using mammography [39].

learning capacity but require greater computational resources than smaller configurations like B/16 (59M parameters).

Understanding these adjustments is particularly important in breast cancer detection using mammography, where diagnostic accuracy and computational efficiency play key roles. A key objective of this study is to evaluate these compromises in the context of breast cancer detection using mammography and develop a practical guide for selecting the most effective MLP-Mixer model based on diagnostic performance and computational feasibility.

Figure 2 illustrates the MLP-Mixer-based framework for breast cancer diagnosis using mammograms, highlighting its key components. The architecture includes patch embedding, which transforms mammographic images into tokenised representations, followed by mixer layers that facilitate both spatial (token) and feature (channel) interactions. The final classification pipeline is augmented for extracting and analysing critical patterns in mammographic images to enhance diagnostic accuracy.

C. OVERVIEW OF CNN-BASED PRETRAINED MODELS FOR MAMMOGRAPHIC ANALYSIS

Deep learning models have played a transformative role in breast cancer detection using mammography, with CNN-based pre-trained models being the most widely used due to their proven feature extraction capabilities [37], [38]. These architectures employ convolutional layers, pooling operations, and residual connections to learn hierarchical representations, making them highly effective for image classification and segmentation in medical imaging. By capturing local and broader structural patterns, CNNs enable robust tumor and anomaly detection in mammographic images.

TABLE 2. An overview of commonly used CNN-based pre-trained models.

Model	Input Resolution	Number of Layers	Filter Size	Number of Parameters (Millions)
Xception	299×299	36	3×3	22.9
VGG16	224×224	16	3×3	138
VGG19	224×224	19	3×3	144
ResNet50V2	224×224	50	7×7 (initial), 3×3 (residual)	25.6
ResNet52V2	224×224	152	7×7 (initial), 3×3 (residual)	60.4
MobileNet	224×224	Variable	3×3 (depthwise separable)	4.2
MobileNetV2	224×224	Variable	3×3 (depthwise separable)	3.4
DenseNet121	224×224	121	3×3	8.0
DenseNet169	224×224	169	3×3	14.3
DenseNet201	224×224	201	3×3	20.0

Table 2 provides an overview of commonly used CNN-based pre-trained models, detailing their input resolution, depth (number of layers), filter sizes, pooling strategies, and parameter count. The table includes lightweight architectures such as MobileNetV2 (3.4M parameters), deeper networks like ResNet152V2 (60.4M parameters), and high-capacity models like VGG19 (144M parameters).

These models are introduced to serve as a baseline for comparison against MLP-Mixer-based architectures, allowing for a comprehensive evaluation of their effectiveness in mammographic image classification. By analysing their performance across key metrics, this study aims to determine whether MLP-Mixer models provide a viable alternative to traditional CNN-based approaches in breast cancer detection.

A key difference between CNN-based and MLP-Mixer-based models lies in how they process spatial dependencies, which is particularly important in medical imaging applications such as breast cancer detection using mammography. CNNs rely on convolutional kernels, which slide over an image to capture local patterns such as edges, textures, and fine details. These kernels extract hierarchical representations, where shallow layers detect low-level features (e.g., contours and gradients), while deeper layers combine these into complex structures (e.g., tissue patterns and tumours). Generally, convolutional layers inherently focus on local receptive fields, but capturing long-range dependencies and global spatial relationships requires deeper architectures. This is why architectures such as ResNet152V2 (152 layers) and VGG19 (19 layers) are significantly deeper, to enable them to combine gradually local features into a more holistic understanding of the image.

While CNNs excel in localised pattern recognition, their reliance on convolutional layers with fixed kernel sizes makes it challenging to model long-range dependencies efficiently without significantly increasing depth. This can be a limitation in mammography, where both fine-grained details such as microcalcifications and tumour margins, as well as broader structural patterns such as breast tissue asymmetry and lesion localisation, are equally important.

MLP-Mixer models consider spatial feature extraction in a different way by separating an image into fixed-size patches and processing them using fully connected (MLP) layers. Instead of relying on local kernels, they use token-mixing MLP layers, which allow direct global interactions between different image regions, even in early layers. This means that spatial dependencies across distant regions of the mammogram can be learned more efficiently without requiring deep hierarchies like CNNs. For example, in MLP-Mixer models, an image is first divided into patches of size 16×16 or 32×32 pixels, forming a sequence of tokenised representations. Each token (patch) is then processed using channel-mixing MLP layers, which extract feature relationships across different image regions, and token-mixing MLP layers, which allow the model to exchange information across the entire image at every stage. This structure inherently supports long-range dependencies without needing depth-intensive architectures, making it possibly useful for capturing localised lesions and broader mammographic patterns.

Given the widespread use of CNN-based pretrained models for mammographic analysis, this study investigates whether MLP-Mixer models, with their distinct architectural design, can offer comparable or improved performance. We focus on evaluating their accuracy, computational efficiency, and feature extraction capabilities across multiple benchmark mammography datasets. This sets the foundation for a more in-depth comparative analysis in the following sections.

III. EXPERIMENTAL STUDY

A. DATASETS

This section details the validation of four pre-trained MLP-Mixer architectures, B/16, B/32, L/16, and L/32, in the context of breast cancer diagnosis utilising mammographic imaging. The evaluation is performed using three prominent publicly accessible datasets: the MIAS (Mammographic Image Analysis Society database), the CBIS-DDSM (Curated Breast Imaging Subset of the Digital Database for Screening Mammography), and the INbreast dataset. The datasets include a diverse range of mammographic images, enabling a thorough assessment of model performance across various imaging conditions and diagnostic challenges.

1) CURATED BREAST IMAGING SUBSET OF DDSM (CBIS-DDSM)

The Curated Breast Imaging Subset of DDSM (CBIS-DDSM) is an optimised and standardised version of the Digital Database for Screening Mammography (DDSM), aimed at improving the quality and accessibility of mammographic data for CAD research [42], [43]. Unlike its original DDSM, CBIS-DDSM offers full-field digital mammograms in DICOM format, providing high-resolution and decompressed images that are consistent with current computer vision datasets. Additionally, the dataset features a selection of cases meticulously reviewed by expert mammographers, enhanced mass segmentations, and structured annotations. These attributes show that CBIS-DDSM is an invaluable

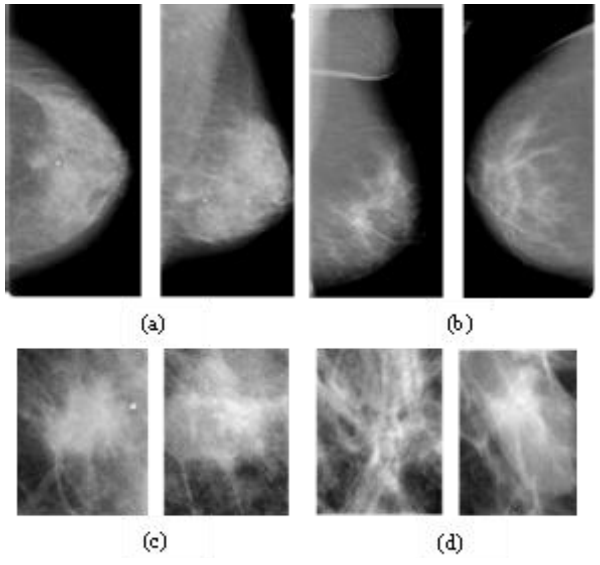

FIGURE 3. Sample images from the CBIS-DDSM dataset: Full and cropped mammograms for malignant (a, c) and benign masses (b, d).

resource for the development and benchmarking of deep learning algorithms in breast cancer detection.

The CBIS-DDSM dataset consists of 1,644 cases, systematically classified into four main categories: Benign Calcification, Benign Mass, Malignant Calcification, and Malignant Mass. Within this collection, there are 753 instances of calcifications and 891 instances of masses. Each case comes with comprehensive annotations detailing the lesion type and corresponding pathology, to enable in-depth analysis and study in the field of mammography. This study is focused on mass lesions, utilising a training set comprising 355 benign and 336 malignant cases, along with a testing set containing 117 benign and 83 malignant cases. Cases with calcifications have been excluded from this phase of investigation to enable a more concentrated analysis of mass-based breast cancer classification.

The CBIS-DDSM dataset is essential to our research owing to its high-quality imaging and the diverse presentation of cases, complemented by expert-reviewed segmentations critical for the development of robust computer-aided detection (CAD) algorithms. Figure 3 shows representative examples from the CBIS-DDSM mammogram dataset, highlighting the variety of breast tissue patterns, lesion classifications, and imaging modalities encompassed within the dataset.

2) INBREAST DATABASE

The INbreast dataset is an integral resource in the domain of breast cancer research and CAD systems, comprising high-resolution full-field digital mammography (FFDM) images [44]. The INbreast dataset was collected from the Centro Hospitalar de S. João in Porto, Portugal, it includes 410 mammographic images derived from 115 distinct cases that represent a comprehensive array of both pathological and normal findings. A key aspect of the INbreast dataset is its meticulous annotation of lesions, which encompasses various types including masses, calcifications, asymmetries, and architectural distortions. Each abnormality is accurately delineated by radiologists to ensure the dataset offers high-quality ground truth annotations. These annotations are available in XML format, streamlining their integration into machine learning workflows for model training and validation. This organised annotation format not only enhances the dataset's usability for algorithm development but is particularly beneficial for the detection and classification of subtle mammographic features essential for early breast cancer diagnosis. The combination of high-resolution imaging, detailed lesion annotations, and the representation of diverse pathological entities solidifies INbreast's importance as an invaluable asset for developing CAD systems. Figure 3 illustrates representative examples from the INbreast dataset.

3) MAMMOGRAPHIC IMAGE ANALYSIS SOCIETY (MIAS) DATABASE

The Mammographic Image Analysis Society (MIAS) database is a cornerstone resource in breast cancer research,

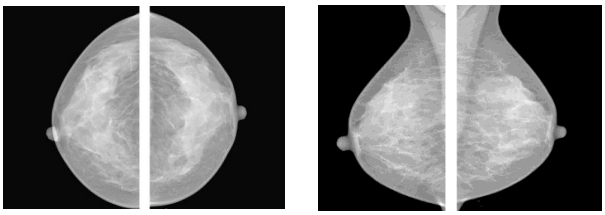

FIGURE 4. Typical images from the INbreast dataset: (a) Craniocaudal (CC) and (b) Mediolateral Oblique (MLO) Views of both breasts.

well-recognised for its essential role in benchmarking CAD systems. MIAS was developed collaboratively by a consortium of UK research groups. It offers a thoroughly structured and standardised dataset that allows comprehensive evaluation of image-based diagnostic algorithms [45], [46]. The dataset comprises 322 mammographic images sourced from 161 cases, each containing comprehensive annotations on lesion type and anatomical location. Specifically, the collection includes 63 benign, 52 malignant, and 207 normal images, all acquired in mediolateral oblique (MLO) views from both the left and right breast. Initially, the images were digitised using a high-precision scanning microdensitometer, achieving a spatial resolution of $50\mu\text{m} \times 50\mu\text{m}$, with each pixel represented in 8-bit depth. For improved accessibility and computational efficiency, the images were subsequently downscaled to a pixel resolution of $200\mu\text{m}$, consistent with methodologies outlined in prior studies. The MIAS provides high-resolution images at 1024×1024 pixels in PGM format, which serves as a critical tool for algorithm development, performance validation, and comparative studies. Figure 5 depicts a selection of images from the MIAS dataset, illustrating its application in research.

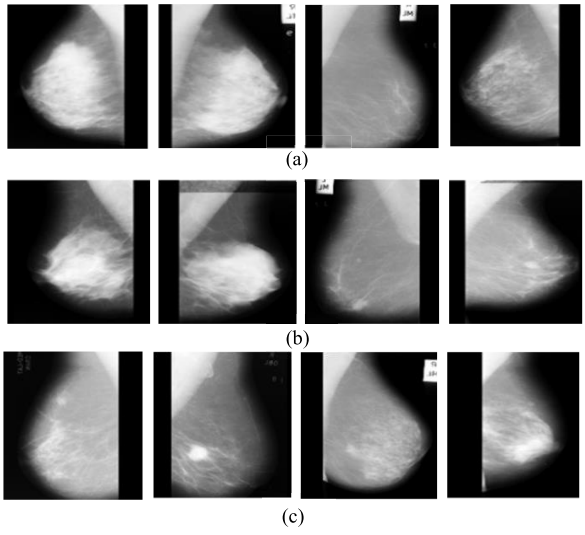

FIGURE 5. Typical samples from the MIAS mammogram dataset: Illustrating (a) Normal, (b) Benign, and (c) Malignant cases.

B. DATA PRE-PROCESSING

To ensure uniform input dimensions for compatibility with CNNs and MLP-mixer architectures, all mammogram images were resized while preserving their aspect ratio. This approach preserved the original proportions, necessary for avoiding distortion and maintaining critical diagnostic details. Bilinear interpolation was utilised to facilitate smooth pixel transitions, ensuring high-quality feature representation. Furthermore, padding and cropping techniques were employed to normalise the images to the required resolution, employing zero padding to safeguard the integrity of the core mammographic content. These preprocessing steps are important in enhancing the model's ability to extract relevant features effectively, leading to consistent training and evaluation outcomes.

To enhance the generalisation capabilities of the model, we implemented a range of data augmentation techniques specifically for the training dataset. These transformations are designed to simulate real-world variations commonly encountered in mammographic imaging, thus improving the model's robustness to positional differences and intensity variations. The augmentation strategies employed include rotations of $-10°$, $0°$, and $10°$ to account for minor misalignments, as well as translations by -11, 0, and $+11$ pixels to reflect potential patient movement during imaging. Scaling transformations at factors of $0.9\times$, $1.0\times$, and $1.1\times$ were applied to approximate variations in imaging distances. To further refine the model's adaptability, we incorporated horizontal flipping to accommodate laterality differences. Adjustments to contrast and brightness, specifically with parameters alpha $= 1.1$ and beta $= 10$, facilitated the learning of illumination-invariant features [34]. These augmentation techniques significantly increased the diversity of the dataset, enabling MLP-Mixer and CNN models to analyse a wider array of patterns. This is particularly critical in medical imaging scenarios where labelled data is often limited. Importantly, the testing dataset was kept unchanged to guarantee its reliability as a benchmark for evaluating the model. By implementing these preprocessing strategies, this study aims to develop a robust and adaptable deep-learning framework to enhance breast cancer detection systems.

C. MODEL TRAINING AND EVALUATION SETUP

In this study, we build on the preprocessing and data augmentation strategies previously outlined to establish a comprehensive training and evaluation framework for breast cancer detection using mammographic images. Given the essential variability in mammographic imaging, the training of deep learning models requires a well-structured approach that optimises predictive accuracy while ensuring computational efficiency.

Four pre-trained MLP-Mixer architectures, specifically B/16, B/32, L/16, and L/32, were fine-tuned for the task of binary breast cancer classification, utilising a patch-based approach to effectively capture spatial and structural features in mammography images. These models were accessed through the PyTorch Image Models (timm) library, which is a robust open-source resource that provides an extensive array of state-of-the-art pre-trained deep learning models optimised for various computer vision applications [47], [48]. The timm library offers highly optimised implementations and a range of pre-trained weights while incorporating advanced training utilities. This makes it an important tool for transfer learning and model benchmarking, particularly within the deep learning research landscape.

To ensure input consistency across the datasets, a detailed pre-processing was applied to all images, which involved resizing them to a uniform resolution of 224×224 pixels to ensure compatibility with the input structure of pre-trained models. Selective data augmentation techniques were systematically applied to the training dataset to reinforce the model's ability to generalise and enhance its robustness, as outlined in the data pre-processing section above. These techniques included rotations, translations, scaling, and contrast adjustments, each contributing to the introduction of variability in the training data. The dataset was randomly partitioned into three subsets, where 64% was allocated for training, 16% for validation, and 20% reserved for testing. Augmentations were confined exclusively to the training set to guarantee an unbiased assessment of the model's performance.

The model was trained for 50 epochs with early stopping based on validation loss to ensure convergence without overfitting. The Adam optimiser, which is known for its efficient convergence properties, was used alongside a specified learning rate of 0.00001 to ensure stable fine-tuning of pre-trained models on relatively small medical datasets. A cross-entropy loss function was implemented to measure the difference between the predicted and actual outcomes. A batch size of 16 was selected to provide an optimal balance between computational efficiency and the stability of convergence during training. Throughout this process, a comprehensive set of performance metrics was diligently monitored, including accuracy, specificity, precision, recall (sensitivity), F1-score, area under the ROC curve (AUC), and distance from the ideal position (DIP). This rigorous tracking enabled a thorough evaluation of the model's predictive capabilities. Additionally, to support reproducibility, we trained each model 10 times independently. We carefully combined the results to provide a strong performance assessment. This detailed process strengthened the credibility of our findings and gave us more insights into how reliable and effective the model is under different training conditions.

D. PERFORMANCE EVALUATION METRICS

To evaluate thoroughly the performance of the MLP-Mixer models in breast cancer detection, we employed a comprehensive set of evaluation metrics, ensuring a balanced assessment of predictive accuracy, reliability, and clinical applicability. These metrics include accuracy, specificity, precision, recall (sensitivity), F1-score, area under the ROC curve (AUC), and distance from the ideal position (DIP). Each of these measures provides a unique perspective on

model performance, capturing different aspects of predictive capability. The foundation of these metrics is based on true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), which serve as fundamental components for evaluating classification models [34]:

- True Positives (TP): Correctly classified positive cases (e.g., correctly identifying malignant tumours).
- False Negatives (FN): Cases where the model incorrectly classifies positive instances as negative (e.g., missing a malignant tumour).
- True Negatives (TN): Correctly classified negative cases (e.g., correctly identifying benign cases).
- False Positives (FP): Cases where the model incorrectly classifies negative instances as positive (e.g., falsely predicting cancer in a benign case).

Utilising these definitions, the following key metrics were computed to assess the effectiveness of the trained models.

1) Accuracy: A fundamental measure of overall classification performance, representing the proportion of correctly classified instances across both classes. It is defined as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

2) Specificity: Also known as the true negative rate, specificity quantifies the model's ability to correctly identify negative cases, minimising false positives. High specificity is very necessary in medical diagnostics to reduce unnecessary interventions for patients without disease. Specificity can be calculated as:

$$Specificity = \frac{TN}{TN + FP} \tag{11}$$

3) Precision: Precision measures how many of the predicted positive cases are positive, helping to evaluate the reliability of positive classifications. Precision can be computed as follows:

$$Precision = \frac{TP}{TP + FP} \tag{12}$$

4) Recall (Sensitivity): Sensitivity assesses how effectively the model captures all actual positive cases, ensuring that malignant cases are not overlooked:

$$Recall = \frac{TP}{TP + FN} \tag{13}$$

5) F1-Score: The F1-score provides a balanced metric by incorporating both precision and recall, ensuring that neither metric is favoured disproportionately:

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{14}$$

6) Area Under the ROC Curve (AUC-ROC): AUC-ROC evaluates the model's ability to distinguish between positive and negative cases. The ROC curve plots the true positive rate (recall) against the false positive rate, and the AUC quantifies the area under this curve.

7) Distance from the Ideal Position (DIP): Unlike traditional classification metrics, DIP provides a holistic performance measure by considering multiple independent metrics. It is calculated using the Euclidean distance from the ideal performance value (1 for each metric), normalised by the number of metrics, and transformed into a score ranging from 0 to 1:

$$DIP = 1 - \frac{\sqrt{\sum_{i=1}^{N} (1 - m_i)^2}}{\sqrt{N}} \tag{15}$$

Here m_i is the value of the *ith* metric, and N is the number of metrics. A higher DIP value indicates better performance, with 1 being the best possible score. DIP has been proven to be superior to F1-score in high-performance settings, making it particularly useful for fine-grained model evaluation [49].

E. EXPERIMENTAL RESULTS AND ANALYSIS

To systematically assess the efficacy of MLP-Mixer pre-trained models in breast cancer diagnosis through mammography, we carried out a series of experiments utilising the three established benchmark datasets described above: CBIS-DDSM, INbreast, and MIAS. Each dataset presents a unique set of imaging features, enabling a comprehensive evaluation of the model's performance across varied mammographic data characteristics.

Table 3 provides a detailed performance assessment of MLP-Mixer pre-trained models across three datasets. This evaluation considers validation accuracy, test accuracy, and testing time, offering valuable insights into the efficiency of different model configurations in breast cancer classification tasks.

TABLE 3. Performance comparison of MLP-Mixer models across CBIS-DDSM, INbreast, and MIAS datasets.

Dataset	Model	Validation Accuracy (%) (Mean ± Std)	Test Accuracy (%) (Mean ± Std)	Testing Time (Mean ± Std)
CBIS-DDSM	B/16	99.86 ± 0.10	99.73 ± 0.15	0.85 ± 0.03s
	L/16	99.90 ± 0.07	99.81 ± 0.12	0.98 ± 0.04s
	B/32	99.57 ± 0.15	99.54 ± 0.18	0.78 ± 0.03s
	L/32	99.43 ± 0.20	99.43 ± 0.22	0.92 ± 0.05s
INbreast	B/16	99.72 ± 0.05	99.71 ± 0.03	0.82 ± 0.02s
	L/16	99.62 ± 0.13	99.62 ± 0.11	0.94 ± 0.04s
	B/32	87.65 ± 0.80	87.35 ± 1.25	0.76 ± 0.02s
	L/32	88.97 ± 6.83	88.75 ± 6.76	0.89 ± 0.03s
MIAS	B/16	99.02 ± 0.08	99.09 ± 0.10	0.65 ± 0.02s
	L/16	99.21 ± 0.05	99.22 ± 0.07	0.74 ± 0.03s
	B/32	98.89 ± 0.12	98.90 ± 0.15	0.58 ± 0.01s
	L/32	98.82 ± 0.16	98.82 ± 0.18	0.68 ± 0.02s

The results highlight how patch size, model complexity, and dataset characteristics notably affect classification performance, highlighting the balance between diagnostic accuracy and computational efficiency. Across all tested datasets, models utilising smaller patch sizes (B/16 and L/16) consistently demonstrated enhanced accuracy compared to

their larger patch variants (B/32 and L/32). This trend is particularly clear in the CBIS-DDSM dataset, where the L/16 model achieved a validation accuracy of 99.90% and a test accuracy of 99.81%, outperforming all other configurations. The B/16 model closely followed with a validation accuracy of 99.86% and a test accuracy of 99.73%, supporting the opinion that reduced patch sizes facilitate superior feature extraction and generalisation capabilities. In contrast, the B/32 and L/32 models showed lower test accuracies of 99.54% and 99.43%, respectively, indicating that larger patch sizes may reduce the model's capability to capture fine-grained mammographic structures. Moreover, testing times for CBIS-DDSM varied from 0.78 seconds per image for B/32 to 0.98 seconds for L/16, reflecting the increased computational demands required by deeper and larger models, which produce only marginal performance improvements.

A similar trend was observed in the INbreast dataset, where B/16 and L/16 models achieved the highest test accuracies of 99.71% and 99.62%, respectively. B/32 and L/32 struggled with test accuracies dropping to 87.35% and 88.75%, respectively. The substantial decline in performance suggests that larger patch models may fail to capture critical mammographic patterns essential for accurate classification. The high variability in L/32 performance ($\pm$ 6.76%) further indicates model instability, possibly due to increased sensitivity to dataset variations or inefficient feature representations. This finding highlights the challenge of using large patch sizes in high-resolution mammography, where complex details such as microcalcifications and subtle architectural distortions are important for distinguishing malignant and benign cases. Testing times for INbreast remained within a similar range, with B/16 being the most efficient (0.82s per image), while L/16 required 0.94s.

The MIAS dataset results align with the stated trends, where L/16 and B/16 outperformed the larger patch models. The L/16 model recorded 99.21% validation accuracy and 99.22% test accuracy, slightly outperforming B/16, which achieved 99.02% and 99.09% accuracy, respectively. In contrast, B/32 and L/32 displayed slightly lower accuracy (98.90% and 98.82%), supporting the fact that larger patches may lead to feature loss and reduced classification effectiveness. Testing times were slightly lower for MIAS compared to CBIS-DDSM and INbreast, ranging from 0.58s (B/32) to 0.74s (L/16), likely due to the smaller dataset size and lower complexity of mammographic images in MIAS.

These results clearly compromise model complexity, patch size, and classification performance. The L/16 model consistently outperformed others across all datasets, demonstrating the benefits of a deeper model utilising smaller patch sizes. This indicates that fine-grained feature extraction is crucial in breast cancer detection, where texture and structural patterns are critical for accurate diagnosis. However, the increased computational demands of L/16 (0.98s per image in CBIS-DDSM and 0.94s in INbreast) suggest that B/16 may be a strong alternative, offering comparable performance

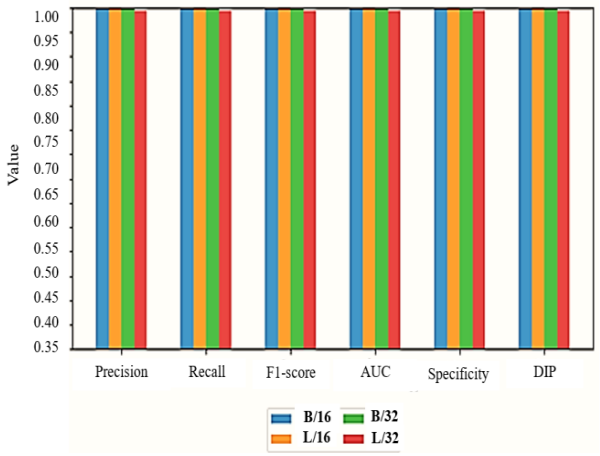

FIGURE 6. Comparison of key performance metrics for MLP-Mixers: Evaluating precision, Recall, F1-Score, AUC, specificity, and DIP across different MLP-Mixer pretrained models using the CBIS-DDSM Dataset.

with reduced inference time. The B/32 and L/32 models were computationally efficient but showed poor classification accuracy, especially on the INbreast dataset, which highlights the limitations of using larger patch sizes in mammographic imaging. The findings highlight the importance of selecting the appropriate deep-learning architectures for mammographic analysis. While deep and complex models provide superior accuracy, their computational cost must be considered, particularly for real-time applications and resource-constrained settings. The superior performance of B/16 and L/16 highlights the importance of preserving local structural details in mammograms, which are often lost when using larger patch-based models. Additionally, the reduced accuracy of B/32 and L/32 in INbreast indicates that certain datasets require enhanced spatial representations to maintain classification performance.

To perform a detailed evaluation of the MLP-Mixer models for breast cancer diagnosis, we implemented a comprehensive validation process employing additional performance metrics such as precision, recall, F1-score, AUC, specificity, and the DIP. These metrics help assess how well each model detects malignant cases, reduces false positives, and maintains consistent performance across different datasets. Given the critical nature of early breast cancer detection, this extended validation ensures that the selected models not only achieve high accuracy but also demonstrate reliability, and generalisability.

Figures 6, 7, and 8 provide a comparative analysis of the four MLP-Mixer architectures (B/16, L/16, B/32, and L/32) using the six additional validation metrics described earlier, including Precision, Recall, F1-score, AUC, Specificity, and DIP. These evaluations were conducted across the three mammographic datasets utilised in our study. As shown in Figure 6, the four MLP-Mixer architectures demonstrate consistently high values (generally $\geq$ 0.985), reflecting the efficacy of MLP-Mixer architectures in medical image classification. Among them, L/16 achieves the highest performance

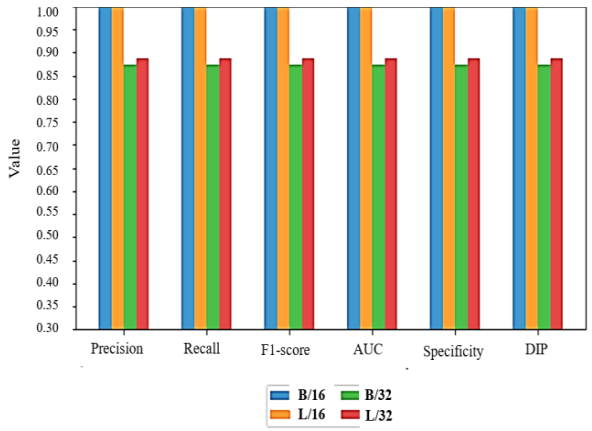

FIGURE 7. Comparison of key performance metrics for MLP-Mixers: Evaluating precision, Recall, F1-Score, AUC, Specificity, and DIP across different MLP-mixer pretrained models using the INbreast dataset.

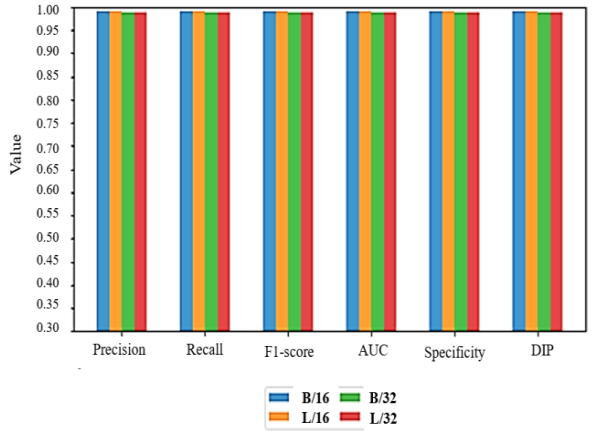

FIGURE 8. Comparison of key performance metrics for MLP-Mixers: Evaluating precision, Recall, F1-Score, AUC, Specificity, and DIP across different MLP-mixer pretrained models using the MIAS dataset.

across all metrics, achieving a Precision of 0.998, a Recall of 0.997, an F1-score of 0.997, an AUC of 0.998, a Specificity of 0.998, and a DIP of 0.992. These results indicate exceptional classification stability and strong sensitivity for detecting malignancies. Following closely is B/16, which displays a Recall of 0.994 and a DIP of 0.990. While slightly behind L/16 in precision and F1-score, B/16 offers a practical efficiency-performance trade-off. Its testing time of 0.85 seconds is approximately 13% lower than L/16's 0.98 seconds, making it a competitive and computationally efficient alternative. In contrast, B/32 and L/32 demonstrate reduced performance, particularly in recall (0.985) and DIP (0.984). These results suggest that larger patch sizes may overlook critical localised features, such as microcalcifications or irregular lesion boundaries, resulting in a higher rate of false negatives. Although their AUC values remain high ($\geq$0.991), the decrease in recall and DIP underscores the trade-off between computational simplicity and spatial sensitivity.

The results presented in Figure 7, which focuses on the INbreast dataset, illustrate that the classification task becomes more difficult due to the limited sample size and higher image resolution of the dataset. Consequently, there is a more significant variance in model performance. The L/16 model outperforms the others, achieving a Recall of 0.985, an F1-score of 0.985, and a DIP of 0.980, which highlights its robustness even under more complex and variable imaging conditions. The B/16 model remains competitive but scores slightly lower across all metrics. In contrast, the performance of the B/32 and L/32 models declines sharply, with the Recall for L/32 dropping to 0.930 and the DIP to 0.915. This 6.5-point decrease in DIP from L/16 indicates considerable instability and poor generalisation, likely due to the reduced spatial granularity resulting from larger patches. These models seem less capable of capturing the significant patterns essential for accurate malignancy detection in complex cases.

As depicted in Figure 8, the performance disparities among models utilising the MIAS dataset are comparatively minimal. All architectures demonstrate commendable performance, with Precision, Recall, and F1-scores all equal to or exceeding 0.988. Notably, the L/16 model maintains a slight advantage with a DIP of 0.991, closely followed by the B/16 model. Conversely, while the B/32 and L/32 models continue to produce satisfactory results, they demonstrate a subtle yet consistent decline in both recall and DIP. This observation proves the conclusion that larger patches, even within simpler datasets, correspond to a slight reduction in performance.

The performance trends observed across all three datasets highlight the advantages of MLP-Mixer models with smaller patch sizes (16×16). The L/16 model consistently achieves the highest performance across all metrics, demonstrating its ability to capture both local and global mammographic features effectively. The B/16 model follows closely, highlighting the magnitude of fine-grained feature extraction. In contrast, the larger patch-size models (B/32 and L/32) present reduced performance, particularly in recall and F1-score, indicating challenges in detecting sensitive malignancy patterns. The high AUC values across all models suggest that MLP-Mixers retain strong classification capabilities, but the DIP metric reveals that smaller patch models offer greater stability and robustness.

Table 4 summarises the performance of MLP-Mixer models across three mammographic datasets, highlighting the best models, key trends, and the impact of patch size on classification accuracy. Smaller patch sizes (16×16) consistently outperform larger ones, enhancing feature extraction and diagnostic precision. The findings highlight the impact of model architecture and patch resolution on the efficacy of mammographic analysis. By employing smaller patch sizes, MLP-Mixer models are capable of extracting more complex and detailed feature representations from mammographic data. This capability not only amplifies their diagnostic accuracy but also enables a deeper understanding of the visual patterns associated with breast tissue abnormalities.

The results suggest that a focused optimisation of patch resolution, combined with the innovative MLP-Mixer architecture, can lead to significant enhancements in breast cancer

TABLE 4. Performance summary of MLP-Mixer models across mammographic datasets.

Dataset	Best Performing Model	Key Observations	Performance Insights
CBIS-DDSM	L/16	Consistently high precision, recall, F1-score, and specificity.	Smaller patch sizes (16×16) allow for better feature extraction, enhancing diagnostic performance.
INbreast	L/16, B/16	Significant performance gap between 16×16 and 32×32 models, with larger patch sizes showing reduced recall and F1-score.	Finer patch resolutions are crucial for handling high-resolution mammograms and capturing subtle malignancies effectively.
MIAS	L/16, B/16	Models with smaller patch sizes outperform larger ones in all metrics, particularly recall and F1-score.	Large patch-size models struggle with subtle feature extraction, impacting malignancy detection accuracy.
Overall Performance Trend	L/16 consistently leads, followed by B/16	AUC values remain high across all models, indicating strong classification ability. DIP scores confirm that smaller patch-size models are more stable and reliable.	MLP-Mixer models benefit from fine-grained feature extraction, and smaller patches significantly enhance classification performance in mammography.

detection rates. This advancement has profound implications for the future of AI-driven mammography analysis, as it provides critical insights that could support the development of more advanced diagnostic tools and offer important directions for future developments in AI-driven mammography methodologies.

Interpretation of Observed Improvements and Trade-Offs

The observed improvements in L/16 stem from its ability to capture both global context and local detail through finer patch representation and early spatial mixing. This is particularly beneficial in mammography, where fine-grained structural cues are vital. However, this comes at the cost of a longer inference time of 0.98 seconds per image. The B/16 model provides a practical compromise with slightly lower performance but faster inference, taking just 0.85 seconds per image. This makes it ideal for resource-limited environments or high-throughput screening settings. On the other hand, the 32×32 patch models, while potentially quicker to train, sacrifice important spatial resolution. This loss results in poorer recall and reduced classification stability, which can be particularly concerning in clinical scenarios where it is crucial to minimise missed malignancies.

Implications for Clinical Applications and the Broader Field

The findings indicate that token-based MLP-Mixer models, particularly those utilising 16×16 patches, can be highly effective components in breast cancer diagnosis. Their ability to generalise across datasets of varying sizes and complexities suggests that these models could be applied in both high-resource diagnostic settings and low-resource environments, such as mobile units or telehealth platforms. Furthermore, the strong performance in metrics such as AUC, F1-score, and DIP indicates that MLP-Mixer architectures may serve as a lighter-weight alternative to convolutional and attention-based models. This could potentially reduce model complexity while maintaining accuracy.

Limitations and Future Research Directions

While our results are promising, it is important to acknowledge certain limitations. The higher inference cost of L/16 might present deployment challenges in large-scale systems without GPU acceleration. Furthermore, our models were evaluated on public datasets; external validation with varied real-world data is essential to confirm their generalisability. Future research should investigate adaptive patch sizing or multi-scale token mixing, as the optimal patch size may vary across imaging modalities and resolutions. Additionally, this study did not delve into explainability or uncertainty estimation, which are crucial for clinical trust and integration. We aim to explore a hybrid MLP architecture, develop interpretability mechanisms (e.g., attention heatmaps), and extend our evaluation to other modalities such as ultrasound and CT in future work.

F. EVALUATION OF CNN-BASED PRETRAINED MODELS FOR BREAST CANCER DETECTION

This section presents a series of experiments conducted using CNN-based pretrained models on the same preprocessed mammographic datasets used for MLP-Mixer models. The objective is to establish a direct performance comparison between traditional convolutional architectures and the MLP-Mixer framework. Various CNN architectures, including lightweight models (e.g., MobileNetV2), deep networks (e.g., ResNet152V2), and high-capacity architectures (e.g., VGG19), were fine-tuned on the mammographic data to assess their classification accuracy, computational efficiency, and overall diagnostic potential. The experimental results provide insights into the strengths and limitations of CNN-based feature extraction in breast cancer detection, forming a foundation for evaluating the advantages of MLP-Mixer models in this domain.

Table 5 presents the classification performance of various deep learning models across the CBIS-DDSM, INbreast, and MIAS datasets, revealing substantial differences in generalisation and robustness. Models trained on CBIS-DDSM achieve consistently high accuracy, with DenseNet201 (99.83% $\pm$ 0.15), MobileNet (99.80% $\pm$ 0.17), and

TABLE 5. Classification performance of CNN-based pretrained models on CBIS-DDSM, INbreast and MIAS datasets.

Model	CBIS-DDSM	INbreast	MIAS
Xception	99.77 ± 0.22	99.25 ± 0.05	98.96 ± 0.37
VGG16	54.59 ± 9.24	67.06 ± 4.18	98.72 ± 0.55
VGG19	58.31 ± 7.47	67.06 ± 8.27	92.84 ± 7.42
ResNet50V2	99.78 ± 0.13	91.19 ± 1.18	99.02 ± 0.25
ResNet52V2	99.79 ± 0.11	98.67 ± 0.35	98.8 ± 0.12
MobileNet	99.80 ± 0.17	97.36 ± 0.60	98.78 ± 0.15
MobileNetV2	99.20 ± 0.63	98.59 ± 0.45	98.70 ± 0.17
DenseNet121	99.71 ± 0.23	98.90 ± 0.39	99.01 ± 0.14
DenseNet169	99.57 ± 0.05	94.52 ± 0.42	98.9 ± 0.29
DenseNet201	99.83 ± 0.15	84.80 ± 0.45	98.8 ± 0.28

ResNet52V2 (99.79% ± 0.11) outperforming others. These architectures leverage deep feature extraction, residual connections, and efficient parameterisation, contributing to superior discrimination of mammographic abnormalities. In contrast, VGG16 (54.59% ± 9.24) and VGG19 (58.31% ± 7.47) exhibit significantly lower accuracy, underscoring the limitations of networks lacking feature reuse mechanisms.

The INbreast dataset presents a more challenging evaluation scenario due to its high-resolution images and limited sample size, increasing the risk of overfitting. While Xception (99.25% ± 0.05) and DenseNet121 (98.90% ± 0.39) maintain strong performance, deeper architectures such as DenseNet201 (84.80% ± 0.45) and ResNet50V2 (91.19% ± 1.18) exhibit a marked decline. This degradation suggests that complex architectures with extensive parameterisation may struggle to generalise when dataset size is constrained. The performance of VGG16 (67.06% ± 4.18) and VGG19 (67.06% ± 8.27) remains notably weak, further reinforcing their limited suitability for mammographic analysis in high-resolution settings.

On the MIAS dataset, a smaller collection of lower-resolution mammograms, models achieve uniformly high accuracy. ResNet50V2 (99.02% ± 0.25), DenseNet121 (99.01% ± 0.14), and Xception (98.96% ± 0.37) rank among the top performers, with VGG16 (98.72% ± 0.55) performing unexpectedly well compared to its results on INbreast. This suggests that smaller datasets with lower complexity may mitigate the architectural deficiencies of early convolutional models. MobileNetV2 (98.70% ± 0.17) and DenseNet169 (98.90% ± 0.29) maintain competitive performance, while the higher variance in VGG19 (92.84% ± 7.42) indicates instability during training.

Collectively, the results highlight that lightweight architectures such as MobileNetV2 and Xception consistently perform well across datasets, demonstrating adaptability and efficiency. In contrast, deep residual networks (ResNet52V2, DenseNet201) excel on large datasets but show vulnerability to overfitting on smaller ones. The VGG family remains largely unsuitable for modern mammography tasks, lacking the architectural innovations necessary for robust feature extraction. These findings highlight the importance of selecting models based on dataset-specific characteristics to optimise deep learning applications in breast cancer detection. Furthermore, Table 6 presents model testing times per image across CBIS-DDSM, INbreast, and MIAS, revealing distinct computational demands. MobileNet (0.85s–0.65s) and MobileNetV2 (0.94s–0.72s) demonstrate the highest efficiency, leveraging depthwise separable convolutions to minimise processing costs. Xception (1.1s–0.85s) remains competitive due to its factorised convolutional structure.

TABLE 6. Testing time (seconds) for CNN-based models on CBIS-DDSM, INbreast, and MIAS datasets based on model complexity.

Model	CBIS-DDSM Testing Time (s)	INbreast Testing Time (s)	MIAS Testing Time (s)
MobileNet	0.85	0.82	0.65
MobileNetV2	0.94	0.9	0.72
Xception	1.1	1.07	0.85
VGG16	1.27	1.23	0.98
VGG19	1.36	1.31	1.04
ResNet50V2	1.44	1.39	1.1
ResNet52V2	1.49	1.43	1.14
DenseNet121	1.53	1.48	1.17
DenseNet169	1.7	1.64	1.3
DenseNet201	1.95	1.89	1.49

In contrast, VGG16 (1.27s–0.98s) and VGG19 (1.36s–1.04s) exhibit increased latency, reflecting the inefficiencies of early CNN architectures. ResNet50V2 (1.44s–1.10s) and ResNet52V2 (1.49s–1.14s) show moderate computational demands, benefiting from residual connections that enhance feature propagation. The DenseNet family, particularly DenseNet201 (1.95s–1.49s), acquires the highest latency due to extensive feature reuse, which, while improving accuracy, increases inference time.

Inference times remain consistent across datasets, with MIAS exhibiting slightly lower values, likely due to reduced image resolution and dataset complexity. MobileNet and MobileNetV2 offer the most promising balance between accuracy and efficiency, while DenseNet models, despite their superior classification performance, impose computational constraints that may limit real-time clinical deployment.

G. COMPARATIVE ANALYSIS OF MLP-MIXER AND CNN-BASED PRETRAINED MODELS

The comparative analysis of MLP-Mixer and CNN-based pretrained models across CBIS-DDSM, INbreast, and MIAS underscores key differences in classification performance and computational efficiency. On CBIS-DDSM, MLP-Mixer variants B/16 (99.73% ± 0.15) and L/16 (99.81% ± 0.12) achieved accuracy comparable to DenseNet201 (99.83% ± 0.15), while maintaining significantly lower testing times (0.85s and 0.98s vs. 1.95s). Similarly, in INbreast, B/16 (99.71% ± 0.03) and L/16 (99.62% ± 0.11) surpassed CNN counterparts, including Xception (99.25% ± 0.05) and DenseNet121 (98.90% ± 0.39), demonstrating superior generalisation despite dataset limitations. B/32

(87.35% ± 1.25) and L/32 (88.75% ± 6.76) demonstrated substantial accuracy degradation, suggesting that larger patch sizes compromise fine-grained spatial learning. On MIAS, where dataset complexity is lower, MLP-Mixer models performed on same level with ResNet50V2 (99.02% ± 0.25) and DenseNet121 (99.01% ± 0.14), with B/16 (99.09% ± 0.10) and L/16 (99.22% ± 0.07) matching or exceeding CNN accuracy. Importantly, inference times for MLP-Mixer models remained consistently lower across all datasets, particularly in MIAS (B/32: 0.58s vs. ResNet50V2: 1.10s), highlighting their computational efficiency.

These findings highlight MLP-Mixer's capacity to deliver CNN-level accuracy with reduced inference time (30 – 50% faster), particularly on large-scale datasets such as CBIS-DDSM, where hierarchical convolutions are less critical. The token-mixing design allows for early global context aggregation without the computational cost of deep convolutional hierarchies. On the INbreast, MLP-Mixer models showed greater generalisation and stability, outperforming complex CNNs such as DenseNet201, which suffered from performance drops due to data size constraints. The results suggest that MLP-Mixer models, especially those utilising smaller patches (B/16, L/16), provide an effective alternative to deep CNNs in mammogram analysis, offering a promising balance between accuracy and speed, filling a significant gap in current literature where most models favour depth over efficiency. CNNs, particularly DenseNet and ResNet, remain viable for applications requiring deeper feature hierarchies, but their increased inference times pose limitations for real-time clinical workflows. The efficiency of MLP-Mixer models, particularly in high-resolution mammography, supports their potential for scalable deployment in resource-constrained medical imaging applications, where rapid and precise diagnosis is essential.

H. PERFORMANCE EVALUATION OF MLP-MIXER B/16 AND L/16 MODELS AGAINST STATE-OF-THE-ART METHODS IN MAMMOGRAPHY

Key Architectural Differentiators and Advantages of MLP-Mixer Models:

Having demonstrated the superior performance of the MLP-Mixer B/16 and L/16 models across multiple mammographic datasets, we now examine the architectural foundations that drive their effectiveness. The MLP-Mixer introduces a novel paradigm in medical image analysis by entirely replacing traditional convolutional and attention-based mechanisms with a pure multilayer perceptron. This design presents a series of structural and clinical advantages that are particularly relevant for the classification of mammographic images.

1. **Fundamentally Different Design**: Beyond Convolutions and Attention: Unlike CNNs and Vision Transformers (VTs), MLP-Mixer models eliminate the utilisation of convolutions and attention mechanisms. Instead, they employ a methodology of token-mixing and channel-mixing MLP layers. This allows for direct spatial and cross-channel interactions across the entire input, thereby simplifying architectural depth and reducing parameter complexity while maintaining the model's capacity to capture and integrate both local and global features. In the context of mammography, where abnormalities may occupy substantial spatial areas or manifest as subtle distributed patterns, this direct and comprehensive processing approach is particularly effective.

2. **Early Integration of Global Context**: CNNs typically develop a global spatial understanding through progressively deeper layers, which may result in increased latency and heightened risks of overfitting when applied to smaller datasets. Conversely, MLP-Mixers facilitate global information flow from the initial token-mixing layer, which ensures that long-range spatial relationships are captured early and efficiently. This capability significantly enhances the detection of dispersed features, such as microcalcifications, architectural distortions, or subtle masses, which are challenging to model with localised convolution filters.

3. **Robust Generalisation Across Datasets**: The MLP-Mixer models show strong and consistent performance across mammographic datasets that vary in size, resolution, and acquisition characteristics. Our evaluation revealed that both the B/16 and L/16 models maintained consistently high accuracy and sensitivity across the MIAS, CBIS-DDSM, and INbreast datasets, without the need for dataset-specific architectural adjustments. In contrast, deeper CNNs like DenseNet201 tend to overfit on smaller datasets such as INbreast, which limits their generalisability. The MLP-Mixer's ability to generalise without modifications demonstrates its architectural robustness and suitability for various real-world applications.

4. **Diagnostic Metrics Beyond Accuracy**: While classification accuracy is often used as a standard metric in deep learning studies, it does not fully reflect the clinical diagnostic performance of models. To provide a more comprehensive evaluation, we assess our models using additional clinically relevant metrics, including the Area Under the Curve (AUC) and Distance from the Ideal Position (DIP). The DIP measures how much a model's performance deviates from that of an ideal classifier across several independent metrics, specifically sensitivity, specificity, and accuracy. This approach offers a more complete perspective on diagnostic reliability. The MLP-Mixer models consistently demonstrate low DIP values, signifying a strong alignment with optimal performance and effective balances between true positive and true negative rates. Furthermore, their high sensitivity, which is at least 99.2%, highlights their effectiveness in reliably detecting malignant cases, an essential factor in breast cancer screening aimed at minimising missed diagnoses.

5. **Reduced Complexity and Deployment Efficiency**: When compared to CNNs, which often comprise millions of parameters and intricate connectivity (e.g., ResNet152V2 with over 60 million parameters or DenseNet201 with densely connected skip connections), MLP-Mixer models provide a lightweight and efficient alternative. They are simpler to implement and fine-tune, require fewer

hyperparameters, and are well-suited for optimisation in low-power or real-time clinical environments. Their reduced architectural burden and rapid inference times render them particularly promising for scalable deployment in resource-constrained settings, such as mobile diagnostics or rural screening units.

Comparison of Results

To further evaluate the effectiveness of MLP mixers-based pre-trained models, we compare their performance with recently published results [34], [50], [51], [52], [53], [54], [55], [56], [57], [58], [59], [60], utilising the same datasets used in this study. Table 7 presents the datasets in the first column, followed by reference numbers in the second. The subsequent columns show accuracy, sensitivity, and AUC, which provide a structured evaluation of classification performance across methods. This comparison allows for a comprehensive assessment of MLP-Mixer models (B/16 and L/16) against various CNN-based architectures and hybrid machine-learning techniques used for breast cancer detection with mammography.

In [50], a method employs a 2D-Fourier Bessel decomposition technique (2D-FBDM) to extract texture features from mammograms, followed by linear regression to classify benign and malignant masses. In [51], a method employs Deep Multiple Instance Learning (MIL) to integrate bidimensional empirical mode decomposition (BEMD) to improve ROI-based analysis, demonstrating robust performance on INbreast. Moreover, Wavelet-transform-based models, utilising 2D-DWT and statistical feature selection, and classification using a back-propagation neural network (BPNN) to differentiate between normal, benign, and malignant breast tissues [52]. In [53], transfer learning approaches where features from mammogram images are extracted using pre-trained CNN architectures such as VGG16, ResNet50, and InceptionV3. These features are then fine-tuned and used in combination with classifiers such as SoftMax and SVM to enhance the diagnostic performance across various evaluation metrics, particularly on the MIAS dataset.

In [54], a method is presented that integrates multi-feature fusion for breast mass classification. This approach extracts complementary features including SIFT, GIST, HOG, LBP, ResNet, DenseNet, and VGG. It also involves mining cross-modal pathological semantics and applying dynamic weight computation for mid-level fusion. Finally, it employs ensemble learning with voting strategies for the final classification. In [55], a technique for classifying breast tumours involves two automated methods. The first method utilises region-growing segmentation, with thresholds set by a trained artificial neural network (ANN). The second method employs cellular neural network (CNN) segmentation, where parameters are optimised using a genetic algorithm (GA). This is followed by feature extraction and classification using ANN and other classifiers. A method integrates pre-trained CNN models, such as EfficientNet, with ensemble learning that employed majority and soft voting strategies to classify mammogram images [56]. Furthermore, a three-stage transfer learning process utilising EfficientNet for breast cancer diagnosis in two-view mammography is described in [57]. The model is trained sequentially on natural images, mammogram patches, and complete mammogram views. This approach successfully achieves high accuracy by using complementary information from both views. Reference [58] introduces a method that utilises a modified U-Net model for

TABLE 7. Comparative performance of MLP-mixer models and state-of-the-art methods across three benchmark mammography datasets (MIAS, CBIS-DDSM, INbreast) using accuracy, sensitivity, and AUC metrics.

		Validation Metrics		
Dataset	**Method/Ref**	**Accuracy (%)**	**Sensitivity (%)**	**AUC**
MIAS	2D-FBDM [50]	96.2	96.02	0.96
	MIL+BEMD [51]	98.04	98.12	0.9817
	2D-DWT + BPNN [52]	94.2	100	0.95
	VGG16 Transfer [53]	98.96	97.83	0.995
	MoEffNet [34]	99.4	99.2	0.992
	MLP-mixers			
	(B/16)	**99.09**	**99.07**	**99.08**
	(L/16)	**99.22**	**99.2**	**99.3**
CBIS-DDSM	2D-FBDM [50]	99.06	98.48	0.99
	MIL+BEMD [51]	98.62	98.60	0.9818
	DL Fusion [54]	90.91	82.96	0.983
	ANN + CNN [55]	96.47	96.87	-
	EfficientNet Ensemble [56]	96.05	-	-
	EfficientNet Transfer + 2-View [57]	92.98	85.13	93.44
	Segmentation + Pretrained CNN [58]	98.87	98.98	09888
	MoEffNet [34]	99.6	99.5	0.995
	MLP-mixers			
	(B/16)	**99.73**	**99.71**	**99.71**
	(L/16)	**99.81**	**99.78**	**99.81**
INbreast	MIL+BEMD [51]	98.26	97.60	0.9823
	Discriminative CNN Transfer Learning [59]	99.8	-	-
	CNN Transfer Learning [60]	95.5	-	0.97
	MoEffNet [34]	99.8	99.8	0.997
	MLP-mixers			
	(B/16)	**99.71**	**99.72**	**99.74**
	(L/16)	**99.62**	**99.62**	**99.65**

the segmentation of mammogram images, followed by classification using pre-trained CNN models such as InceptionV3, DenseNet121, ResNet50, VGG16, and MobileNetV2. The technique employs transfer learning and data augmentation to improve performance.

A study in [59] describes a technique for training a deep-learning model to diagnose breast cancer. This involves using discriminative fine-tuning, which assigns different learning rates to each layer of the deep convolutional neural network (CNN), and mixed-precision training to lower computational requirements. The approach also incorporates data augmentation to improve the model's performance on a small dataset, resulting in rapid convergence and high accuracy. A method that employs deep Convolutional Neural Networks (CNNs) to classify mammogram images is described in [60]. The approach employed transfer learning and fine-tuning strategies using pre-trained architectures, including VGG16, ResNet50, and Inception v3. Finally, MoEffNet [34], a hybrid architecture integrating EfficientNet with a Mixture of Experts (MoE) framework, dynamically allocates features to specialised subnetworks.

Table 7 provides a detailed comparison of the proposed MLP-Mixer models (B/16 and L/16) against a wide range of state-of-the-art methods [34], [50], [51], [52], [53], [54], [55], [56] described above. The listed references can be summarised into a variety of model categories:

- Traditional methods (e.g., 2D-FBDM [50], 2D-DWT + BPNN [52]),
- CNN-based transfer learning approaches (e.g., VGG16 [53], ResNet [60]),
- Hybrid and ensemble models (e.g., MoEffNet [34], EfficientNet Ensembles [56]),
- Segmentation and fusion techniques (e.g., DL-Fusion [54], U-Net + CNN [58]),
- Advanced deep learning pipelines (e.g., MIL+BEMD [51], Discriminative Transfer Learning [59]).

Across the MIAS, CBIS-DDSM, and INbreast datasets, the results consistently demonstrate that the MLP-Mixer models (B/16 and L/16) either outperform or match the best results achieved by the other listed studies. For instance, on the MIAS dataset, the MLP-Mixer models achieved an accuracy of 99.09% to 99.22%, which is closely aligned with the highest-performing model in the literature, which reported an accuracy of 99.4%. Additionally, they demonstrated remarkable sensitivity (99.07% to 99.2%) and AUC values (99.08% to 99.3%). In the CBIS-DDSM dataset, MLP-Mixers surpassed most CNN-based approaches with accuracy figures ranging from 99.73% to 99.81%, and sensitivity between 99.71% to 99.78%. These results exceed those of previous methods and are in line with MoEffNet, which achieved an accuracy of 99.6%. The AUC values, ranging from 99.71% to 99.81%, further confirm their robust performance in malignancy detection. For the INbreast dataset, MLP-Mixers achieved accuracy rates ranging from 99.71% to 99.62%. These results are comparable to the highest-performing model reported at 99.8% and exceed previous CNN-based models. Their sensitivity ranged from 99.72% to 99.62%, with AUC values between 99.74% and 99.65%, reinforcing their diagnostic capabilities. These results show that MLP-Mixer models generalise effectively across various datasets, achieving classification precision comparable to leading CNNs while remaining computationally efficient. This makes them particularly suitable for breast cancer detection using mammograms.

IV. CONCLUSION

This study systematically evaluated the effectiveness of pre-trained MLP-Mixer models for breast cancer diagnosis using mammographic images. We assessed the performance of four MLP-Mixer variants (B/16, L/16, B/32, L/32) across three benchmark datasets, CBIS-DDSM, INbreast, and MIAS, and compared them with state-of-the-art CNN-based models. The findings indicate that MLP-Mixer pre-trained models using smaller patch sizes (B/16 and L/16) consistently achieve classification accuracy, sensitivity, and AUC comparable to or exceeding those of CNN-based approaches, while also maintaining lower computational costs and inference time. Unlike CNNs, which approximate global context through deep hierarchies, the MLP-Mixers model long-range dependencies directly through token mixing, allowing for early and efficient global spatial feature integration.

The consistent performance of B/16 and L/16 across datasets of varying resolution and complexity confirms their strong generalisation capabilities without requiring dataset-specific adjustments. In contrast, larger patch models (B/32 and L/32) underperformed, particularly on high-resolution datasets, highlighting the importance of fine-grained spatial representation in mammographic analysis and the need to balance computational efficiency with diagnostic precision.

The comparative evaluation confirms that MLP-Mixer models offer a simpler yet more effective alternative to traditional CNNs and hybrid ensemble methods. Their reduced architectural complexity, faster inference, and single-stream design eliminate the need for deep or multi-branch networks while preserving diagnostic reliability. This work establishes token-mixing MLP architectures as a promising new direction in AI-assisted breast cancer diagnosis, with strong potential for integration into clinical decision support systems and CAD workflows. Future research should explore the integration of MLP-Mixers with multimodal imaging modalities, hybrid feature extractors, and explainability frameworks to further enhance clinical interpretability, diagnostic accuracy, and deployment readiness.

Data Access Statement

In this study, we use three publicly available datasets: MIAS (Mammographic Image Analysis Society database) (https://www.repository.cam.ac.uk/items/b6a97f0c-3b9b-40ad-8f18-3d121eef1459), CBIS-DDSM (Curated Breast Imaging Subset of the Digital Database for Screening Mammography) (https://www.cancerimagingarchive.net/collection/cbis-ddsm/), and INbreast (http://medicalresearch.inescporto.pt/breastresearch/index.php/Get_INbreast_Database).

REFERENCES

[1] *Breast Cancer: World Health Organisation (WHO)*. Accessed: Dec. 24, 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/breast-cancer

[2] *Cancer: World Health Organisation (WHO)*. Accessed: Dec. 24, 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cancer

[3] N. Cabioglu, E. Yavuz, and A. Aydiner, *Breast Cancer: A Guide To Clinical Practice*. Cham, Switzerland: Springer, Jul. 2019, pp. 99–122, doi: 10.1007/978-3-030-16792-9_3.

[4] M. Akram, M. Iqbal, M. Daniyal, and A. U. Khan, "Awareness and current knowledge of breast cancer," *Biol. Res.*, vol. 50, no. 1, pp. 1–23, Oct. 2017, doi: 10.1186/s40659-017-0140-9.

[5] L. F. Ellison and N. Saint-Jacques, "Five-year cancer survival by stage at diagnosis in Canada. Health reports," *Health Rep.*, vol. 34, no. 1, pp. 3–15, Jan. 2023, doi: 10.25318/82-003-x202300100001-eng.

[6] A. N. Giaquinto, H. Sung, K. D. Miller, J. L. Kramer, L. A. Newman, A. Minihan, A. Jemal, and R. L. Siegel, "Breast cancer statistics," *Cancer J. Clinicians*, vol. 72, no. 6, pp. 524–541, Oct. 2022, doi: 10.3322/caac.21754.

[7] D. J. van der Meer, I. Kramer, M. C. van Maaren, P. J. van Diest, S. C. Linn, J. H. Maduro, L. J. A. Strobbe, S. Siesling, M. K. Schmidt, and A. C. Voogd, "Comprehensive trends in incidence, treatment, survival and mortality of first primary invasive breast cancer stratified by age, stage and receptor subtype in The Netherlands between 1989 and 2017," *Int. J. Cancer*, vol. 148, no. 9, pp. 2289–2303, May 2021, doi: 10.1002/ijc.33417.

[8] M. Tariq, S. Iqbal, H. Ayesha, I. Abbas, K. T. Ahmad, and M. F. K. Niazi, "Medical image based breast cancer diagnosis: State of the art and future directions," *Expert Syst. Appl.*, vol. 167, Apr. 2021, Art. no. 114095, doi: 10.1016/j.eswa.2020.114095.

[9] C. K. Kuhl, S. Schrading, C. C. Leutner, N. Morakkabati-Spitz, E. Wardelmann, R. Fimmers, W. Kuhn, and H. H. Schild, "Mammography, breast ultrasound, and magnetic resonance imaging for surveillance of women at high familial risk for breast cancer," *J. Clin. Oncol.*, vol. 23, no. 33, pp. 8469–8476, Nov. 2005, doi: 10.1200/jco.2004.00.4960.

[10] C. H. Lee, D. D. Dershaw, D. Kopans, P. Evans, B. Monsees, D. Monticciolo, R. J. Brenner, L. Bassett, W. Berg, S. Feig, E. Hendrick, E. Mendelson, C. D'Orsi, E. Sickles, and L. W. Burhenne, "Breast cancer screening with imaging: Recommendations from the society of breast imaging and the ACR on the use of mammography, breast MRI, breast ultrasound, and other technologies for the detection of clinically occult breast cancer," *J. Amer. College Radiol.*, vol. 7, no. 1, pp. 18–27, Jan. 2010, doi: 10.1016/j.jacr.2009.09.022.

[11] A. Trentham-Dietz, C. H. Chapman, J. Jayasekera, K. P. Lowry, B. Heckman-Stoddard, J. M. Hampton, J. Caswell-Jin, J. Y. Lu, R. E. Gangnon, L. Sun, and H. Huang, "Breast cancer screening with mammography: An updated decision analysis for the U.S. preventive services task force," Agency Healthcare Research Quality (AHRQ), Rockville, MD, USA, Tech. Rep. 23-05303-EF-2, Apr. 2024.

[12] A. Shahi, "Risk factors of breast cancer among patients from a lower- and middle-income country: A case-control study from Nepal," *J. Clin. Oncol.*, vol. 42, no. 16, Jun. 2024, Art. no. e12502, doi: 10.1200/jco.2024.42.16_suppl.e12502.

[13] S. Y. Song, B. Park, S. Hong, M. J. Kim, E. H. Lee, and J. K. Jun, "Comparison of digital and screen-film mammography for breast-cancer screening: A systematic review and meta-analysis," *J. Breast Cancer*, vol. 22, no. 2, p. 311, Jun. 2019, doi: 10.4048/jbc.2019.22.e24.

[14] K. Freeman, J. Geppert, C. Stinton, D. Todkill, S. Johnson, A. Clarke, and S. Taylor-Phillips, "Use of artificial intelligence for image analysis in breast cancer screening programmes: Systematic review of test accuracy," *BMJ*, vol. 374, no. 1872, pp. 1–15, Sep. 2021.

[15] 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*, vol. 14, no. 21, p. 5334, Oct. 2022, doi: 10.3390/cancers14215334.

[16] A. Gangwal and R. K. Gautam, "Artificial intelligence-driven decisions in breast cancer diagnosis," *Drug Therapy Develop. Triple Negative Breast Cancer*, vol. 2023, pp. 131–151, Jun. 2023, doi: 10.1002/9783527841165.ch8.

[17] J. Tang, R. M. Rangayyan, J. Xu, I. El Naqa, and Y. Yang, "Computer-aided detection and diagnosis of breast cancer with mammography: Recent advances," *IEEE Trans. Inf. Technol. Biomed.*, vol. 13, no. 2, pp. 236–251, Mar. 2009, doi: 10.1109/TITB.2008.2009441.

[18] K. Ganesan, U. R. Acharya, C. K. Chua, L. C. Min, K. T. Abraham, and K. H. Ng, "Computer-aided breast cancer detection using mammograms: A review," *IEEE Rev. Biomed. Eng.*, vol. 6, pp. 77–98, 2012, doi: 10.1109/RBME.2012.2232289.

[19] D. A. Zebari, D. A. Ibrahim, D. Q. Zeebaree, H. Haron, M. S. Salih, R. Damaševič ius, and M. A. Mohammed, "Systematic review of computing approaches for breast cancer detection based computer aided diagnosis using mammogram images," *Appl. Artif. Intell.*, vol. 35, no. 15, pp. 203–2157, Dec. 2021, doi: 10.1080/08839514.2021.2001177.

[20] M. N. Yeasmin, M. Al Amin, T. J. JotiJ, Z. Aung, and M. A. Azim, "Advances of AI in image-based computer-aided diagnosis: A review," *Array*, vol. 6, pp. —, Jul. 2024, Art. no. 100357, doi: 10.1016/j.array.2024.100357.

[21] N. I. Yassin, S. Omran, E. M. El Houby, and H. Allam, "Machine learning techniques for breast cancer computer aided diagnosis using different image modalities: A systematic review," *Comput. Methods Programs Biomed.*, vol. 156, pp. 25–45, Mar. 2018, doi: 10.1016/j.cmpb.2017.12.012.

[22] H. O. A. Ahmed and A. K. Nandi, "Colour clustering and deep transfer learning techniques for breast cancer detection using mammography images," in *The Latest Developments and Challenges in Biomedical Engineering*, vol. 746. Springer, pp. 105–119, doi: 10.1007/978-3-031-38430-1_9.

[23] R. M. Rangayyan, T. M. Nguyen, F. J. Ayres, and A. K. Nandi, "Effect of pixel resolution on texture features of breast masses in mammograms," *J. Digit. Imag.*, vol. 23, no. 5, pp. 547–553, Sep. 2009.

[24] A. Rojas-Domínguez and A. K. Nandi, "Development of tolerant features for characterization of masses in mammograms," *Comput. Biol. Med.*, vol. 39, no. 8, pp. 678–688, Aug. 2009.

[25] A. Rojas Domínguez and A. K. Nandi, "Toward breast cancer diagnosis based on automated segmentation of masses in mammograms," *Pattern Recognit.*, vol. 42, no. 6, pp. 1138–1148, Jun. 2009.

[26] A. Rojas Domínguez and A. K. Nandi, "Detection of masses in mammograms via statistically based enhancement, multilevel-thresholding segmentation, and region selection," *Computerized Med. Imag. Graph.*, vol. 32, no. 4, pp. 304–315, Jun. 2008.

[27] T. Mu, "Analysis of breast tumors in mammograms using the pairwise Rayleigh quotient classifier," *J. Electron. Imag.*, vol. 16, no. 4, Oct. 2007, Art. no. 043004, doi: 10.1117/1.2803834.

[28] A. R. Domínguez and A. K. Nandi, "Improved dynamic-programming-based algorithms for segmentation of masses in mammograms," *Med. Phys.*, vol. 34, no. 11, pp. 4256–4269, Nov. 2007.

[29] R. J. Nandi, A. K. Nandi, R. M. Rangayyan, and D. Scutt, "Classification of breast masses in mammograms using genetic programming and feature selection," *Med. Biol. Eng. Comput.*, vol. 44, no. 8, pp. 683–694, Aug. 2006.

[30] T. Mu, A. K. Nandi, and R. M. Rangayyan, "Classification of breast masses via nonlinear transformation of features based on a kernel matrix," *Med. Biol. Eng. Comput.*, vol. 45, no. 8, pp. 769–780, Jul. 2007.

[31] T. Mu, A. K. Nandi, and R. M. Rangayyan, "Classification of breast masses using selected shape, edge-sharpness, and texture features with linear and kernel-based classifiers," *J. Digit. Imag.*, vol. 21, pp. 69–153, Feb. 2008, doi: 10.1007/s10278-007-9102-z. [Online]. Available: https://doi.org/10.1007/s10278-007-9102-z

[32] D. Abdelhafiz, C. Yang, R. Ammar, and S. Nabavi, "Deep convolutional neural networks for mammography: Advances, challenges and applications," *BMC Bioinf.*, vol. 20, pp. 1–20, Jun. 2019, doi: 10.1186/s12859-019-2823-4.

[33] L. Abdelrahman, M. Al Ghamdi, F. Collado-Mesa, and M. Abdel-Mottaleb, "Convolutional neural networks for breast cancer detection in mammography: A survey," *Comput. Biol. Med.*, vol. 131, Apr. 2021, Art. no. 104248, doi: 10.1016/j.compbiomed.2021.104248.

[34] H. O. Ahmed and A. K. Nandi, "High performance breast cancer diagnosis from mammograms using mixture of experts with EfficientNet features (MoEffNet)," *IEEE Access*, vol. 12, pp. 133703–133725, 2024, doi: 10.1109/ACCESS.2024.3461360.

[35] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vision*, Jun. 2017, pp. 618–626.

[36] S. Lundberg and S. Lee, "A unified approach to interpreting model predictions," in *Proc. NeurIPS*, Jan. 2017, pp. 1–10.

[37] S. M. Ganie and P. K. Pramanik, "Comparative analysis of transfer learning models for breast cancer detection," in *Proc. AIC*, Jul. 2024, pp. 830–836, doi: 10.1109/AIC61668.2024.10731032.
[38] M. B. Ammar, F. L. Ayechi, R. Ksantini, and H. Mahjoubi, "Pre-trained deep convolutional neural network architectures for breast cancer diagnosis in mammography: Current State-Of-The-Art," in *Proc. INISTA*, Sep. 2023, pp. 1–7, doi: 10.1109/INISTA59065.2023.10310612.
[39] I. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, A. Steiner, D. Keysers, J. Uszkoreit, and M. Lucic, "Mlp-mixer: An all-mlp architecture for vision," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2021, pp. 72–24261.
[40] H. Zhang, Z. Dong, B. Li, and S. He, "Multi-scale MLP-mixer for image classification," *Knowledge-Based Syst.*, vol. 258, Dec. 2022, Art. no. 109792, doi: 10.1016/j.knosys.2022.109792.
[41] H. Liu, Z. Dai, D. So, and Q. V. Le, "Pay attention to mlps," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2021, pp. 1–15.
[42] R. S. Lee, F. Gimenez, A. Hoogi, K. K. Miyake, M. Gorovoy, and D. L. Rubin, "A curated mammography data set for use in computer-aided detection and diagnosis research," *Scientific Data*, vol. 4, no. 1, pp. 1–9, Dec. 2017.
[43] M. D. Heath, K. W. Bowyer, D. B. Kopans, P. Kegelmeyer, R. H. Moore, K.-H. Chang, and S. Munishkumaran, "Current status of the digital database for screening mammography," in *Digital Mammography*. Dordrecht, The Netherlands: Springer, Jan. 1998, pp. 457–460.
[44] I. C. Moreira, I. Amaral, I. Domingues, A. Cardoso, M. J. Cardoso, and J. S. Cardoso, "Inbreast: Toward a full-field digital mammographic database," *Academic Radiol.*, vol. 19, no. 2, pp. 236–248, 2012.
[45] J. Suckling, "The mammographic images analysis society digital mammogram database," in *Proc. Exerpta Medica. Int. Congr. Ser.*, vol. 1069, 1994, pp. 375–378.
[46] J. Suckling, J. Parker, D. Dance, S. Astley, I. Hutt, C. Boggis, I. Ricketts, E. Stamatakis, N. Cerneaz, S. Kok, P. Taylor, D. Betal, and J. Savage, "Mammographic image analysis society (MIAS) database v1.21," Dept. Psychiatry, Univ. Cambridge Repository, Cambridge, U.K., Tech. Rep., 2015. [Online]. Available: https://doi.org/10.17863/CAM.105113
[47] R. Wightman. (2019). *Pytorch Image Models*. [Online]. Available: https://github.com/rwightman/
[48] R. Wightman. *Timm: PyTorch Image Models*. Accessed: Jan. 6, 2025. [Online]. Available: https://pypi.org/project/timm/
[49] A. K. Nandi, "From multiple independent metrics to single performance measure based on objective function," *IEEE Access*, vol. 11, pp. 3899–3913, 2023.
[50] P. K. Chaudhary and R. B. Pachori, "Differentiation of benign and malignant masses in mammogram using 2D-Fourier-Bessel intrinsic band functions and improved feature space," *IEEE Trans. Artif. Intell.*, vol. 5, no. 11, pp. 5442–5451, May 2024.
[51] A. Elmoufidi, "Deep multiple instance learning for automatic breast cancer assessment using digital mammography," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–13, 2022.
[52] S. Beura, B. Majhi, and R. Dash, "Mammogram classification using two dimensional discrete wavelet transform and gray-level co-occurrence matrix for detection of breast cancer," *Neurocomputing*, vol. 154, pp. 1–14, Apr. 2015.
[53] A. Saber, M. Sakr, O. M. Abo-Seida, A. Keshk, and H. Chen, "A novel deep-learning model for automatic detection and classification of breast cancer using the transfer-learning technique," *IEEE Access*, vol. 9, pp. 71194–71209, 2021.
[54] H. Zhang, R. Wu, T. Yuan, Z. Jiang, S. Huang, J. Wu, J. Hua, Z. Niu, and D. Ji, "DE-Ada: A novel model for breast mass classification using cross-modal pathological semantic mining and organic integration of multi-feature fusions," *Inf. Sci.*, vol. 539, pp. 461–486, Oct. 2020.
[55] R. Rouhi, M. Jafari, S. Kasaei, and P. Keshavarzian, "Benign and malignant breast tumors classification based on region growing and CNN segmentation," *Expert Syst. Appl.*, vol. 42, no. 3, pp. 990–1002, Feb. 2015.
[56] F. Azour and A. Boukerche, "An efficient transfer and ensemble learning based computer aided breast abnormality diagnosis system," *IEEE Access*, vol. 11, pp. 21199–21209, 2023.
[57] D. G. P. Petrini, C. Shimizu, R. A. Roela, G. V. Valente, M. A. A. K. Folgueira, and H. Y. Kim, "Breast cancer diagnosis in two-view mammography using end-to-end trained EfficientNet-based convolutional network," *IEEE Access*, vol. 10, pp. 77723–77731, 2022.
[58] W. M. Salama and M. H. Aly, "Deep learning in mammography images segmentation and classification: Automated CNN approach," *Alexandria Eng. J.*, vol. 60, no. 5, pp. 4701–4709, Oct. 2021.
[59] A. P. Adedigba, S. A. Adeshina, and A. M. Aibinu, "Performance evaluation of deep learning models on mammogram classification using small dataset," *Bioengineering*, vol. 9, no. 4, p. 161, Apr. 2022.
[60] H. Chougrad, H. Zouaki, and O. Alheyane, "Deep convolutional neural networks for breast cancer screening," *Comput. Methods Programs Biomed.*, vol. 157, pp. 19–30, Apr. 2018.

HOSAMELDIN O. A. AHMED received the Ph.D. degree in electronic and computer engineering from the University of Brunel London. He is currently a Distinguished Researcher and the author specializing in signal and image processing, machine learning, and condition monitoring. He is also focusing on advancing breast cancer diagnosis by applying deep transfer learning and image processing techniques. His recent work involves developing innovative methods for detecting breast cancer using mammography images, demonstrating his commitment to improving healthcare outcomes through technology. In addition to his healthcare-focused research, he has collaborated extensively with Prof. Asoke Nandi, contributing to numerous publications on various aspects of machine condition monitoring. He co-authored the book *Condition Monitoring with Vibration Signals: Compressive Sampling and Learning Algorithms for Rotating Machines* (IEEE—John Wiley & Sons, 2020) which explores advanced methods for monitoring rotating machinery health using compressive sampling and machine learning. Furthermore, he has published widely on topics, such as bearing fault diagnosis, intelligent fault diagnosis frameworks for modular multilevel converters, and internet addiction disorder detection using machine learning. His interdisciplinary research has also contributed to cultural heritage preservation through digital image inpainting and 3-D visual interaction techniques. The H-index of his publications is 13 and the i10-index is 19 (Google Scholar).

ASOKE K. NANDI (Life Fellow, IEEE) received the Ph.D. degree in physics from the Trinity College, University of Cambridge.

He held academic positions in several universities, including Oxford, Imperial College London, Strathclyde, and Liverpool and Finland Distinguished Professorship. In 2013, he moved to Brunel University London. In 1983, he co-discovered the three fundamental particles known as W^+, W^-, and Z^0, providing the evidence for the unification of the electromagnetic and weak forces, for which the Nobel Prize for Physics, in 1984, was awarded to two of his team leaders for their decisive contributions. He made fundamental theoretical and algorithmic contributions to many aspects of signal processing and machine learning. He has much expertise in "big data." He has authored over 650 technical publications, including 320 journal articles and six books, titled *Image Segmentation: Principles, Techniques, and Applications* (Wiley, 2022), *Condition Monitoring with Vibration Signals: Compressive Sampling and Learning Algorithms for Rotating Machines* (Wiley, 2020), *Automatic Modulation Classification: Principles, Algorithms and Applications* (Wiley, 2015), *Integrative Cluster Analysis in Bioinformatics* (Wiley, 2015), *Blind Estimation Using Higher-Order Statistics* (Springer, 1999), and *Automatic Modulation Recognition of Communications Signals* (Springer, 1996). The H-index of his publications is 91 (Google Scholar) and ERDOS number is 2. His current research interests include signal processing and machine learning, with applications to machine health monitoring, functional magnetic resonance data, gene expression data, communications, and biomedical data.

Prof. Nandi is a fellow of the Royal Academy of Engineering and six other institutions. In 2023, he was honoured by the Academia Europaea and the Academia Scientiarum et Artium Europaea. He has received many awards, including the IEEE Heinrich Hertz Award, in 2012, the Glory of Bengal Award for his outstanding achievements in scientific research, in 2010, the Water Arbitration Prize of the Institution of Mechanical Engineers, in 1999, and the Mountbatten Premium of the Institution of Electrical Engineers, in 1998. He is an IEEE Distinguished Lecturer (EMBS), from 2018 to 2019.