**British Journal of Nursing Studies**
ISSN: 2755-080X
DOI: 10.32996/bjns
Journal Homepage: www.al-kindipublisher.com/index.php/bjns

BJNS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

**| RESEARCH ARTICLE**

# A Deep Learning Framework for Early Breast Cancer Detection Among U.S. Women: Integrating Mammography and Clinical EHR Data

**Yasin Arafat[1], Nurtaz Begum Asha[2]✉, Shahriar Ahmed[3], Sadman Haque Sakib[4], Mustafizur Rahman Shakil[5], Afia Zahin Rishta[6], SK Rakib Ul Islam[7]**

[1]*Doctor of Management, International American University, Los Angeles, California, USA*
[2]*College of Business, Westcliff University, Irvine, California, USA*
[37]*School of Business, International American University, Los Angeles, California, USA*
[4]*School of Science, Kyungsung University Busan, South Korea*
[5]*College of Engineering and Technology, Westcliff University, Irvine, California, USA*
[6]*Computer Science & Technology, LA Trobe University, Melbourne, VIC*
**Corresponding Author**: Nurtaz Begum Asha, **E-mail**: n.asha.288@westcliff.edu

**| ABSTRACT**

Breast cancer is the most commonly diagnosed malignancy and a leading cause of cancer-related mortality among women in the United States (American Cancer Society, 2024). Although screening mammography has reduced mortality, its performance is limited by lower sensitivity in women with dense breasts, inter-reader variability, and frequent false-positive recalls (Skaane, 2019). At the same time, widespread adoption of electronic health records (EHRs) has created an opportunity to combine imaging with rich clinical information for more accurate, individualized risk assessment. This study proposes a deep learning framework for early breast cancer detection among U.S. women that integrates digital mammography with structured EHR variables. Mammography images are obtained from a curated public dataset such as the Curated Breast Imaging Subset of the Digital Database for Screening Mammography (CBIS-DDSM), which provides biopsy-verified benign and malignant cases (Lee et al., 2017). EHR-style data are organized following the structure of large U.S. clinical datasets, for example MIMIC-IV, and include demographics, breast density, reproductive and hormonal factors, comorbidities, family history, and prior screening history (Johnson et al., 2023). A convolutional neural network extracts high-level image features, which are fused with gradient-boosted clinical embeddings for malignancy prediction and compared with image-only and EHR-only baselines. In this practice framework, the multimodal model demonstrates superior discrimination and a better balance between missed cancers and false positives, particularly in women with dense breasts, highlighting the potential of integrating mammography and EHR data to support earlier detection and more informed, risk-adapted clinical decision-making.

**| KEYWORDS**

Breast cancer; Digital mammography; Convolutional neural network; Multimodal fusion; Electronic health records; Clinical risk factors; Dense breast tissue; CBIS-DDSM.

**| ARTICLE INFORMATION**

**ACCEPTED:** 01 July 2025 **PUBLISHED:** 12 August 2025 **DOI:** 10.32996/bjns.2025.5.2.6

## 1. Introduction

### *1.1 Background*

According to recent epidemiological reports, breast cancer is one of the most commonly diagnosed cancers among women in the U.S.A. and a leading cause of cancer-related mortality worldwide (American Cancer Society, 2024). Each year, hundreds of thousands of new cases are diagnosed in the United States alone, and breast cancer accounts for a substantial proportion of cancer deaths among women despite advances in screening, diagnosis, and treatment. The high burden is partly explained by the fact that many tumors are still detected at a later stage, when therapeutic options are more limited and outcomes are

poorer. In its early stages, breast cancer may be asymptomatic or present with subtle findings that are difficult to interpret, especially in women with dense breast tissue. As a result, diagnosis and initiation of treatment may be delayed. Since early detection is strongly associated with improved survival and less aggressive interventions, there is a growing need for strategies that can identify women at elevated risk and subtle early malignancies more accurately and at an earlier point in the disease course.

Over the past decade, deep learning and other machine learning (ML) methods have emerged as powerful tools in medical imaging and clinical data analysis. They can discover complex patterns and interactions in high-dimensional data that may not be apparent with conventional statistical techniques. This capability is particularly valuable for breast cancer detection, where imaging features, demographic characteristics, reproductive and hormonal history, and comorbid conditions jointly influence risk and presentation. By leveraging both mammography images and rich electronic health record (EHR) data, deep learning models provide an opportunity to improve early detection and move toward more personalized screening strategies (Lee et al., 2017; Johnson et al., 2023).

### *1.2 Importance of Research*

Given the multifactorial etiology of breast cancer, there is an urgent need for more accurate and reliable prediction models that can integrate diverse risk factors and information sources. While current screening and diagnostic modalities—such as digital mammography, tomosynthesis, ultrasound, and MRI—are effective, they can be costly, time-consuming, and occasionally invasive. Moreover, standard mammography does not always fully account for the wide spectrum of risk factors, including dense breast tissue, family history, prior benign breast disease, hormonal exposures, and metabolic comorbidities. This limitation is especially critical in the early stages of disease, when lesions may be small or radiographically subtle.

Multimodal machine learning represents a promising frontier for developing non-invasive, scalable decision-support tools that combine imaging findings with EHR-derived clinical risk profiles. Such tools can help refine screening strategies by identifying women at higher risk who may benefit from closer surveillance or supplemental imaging and by reducing unnecessary callbacks in low-risk individuals. Improved risk stratification has the potential to enhance the effectiveness of screening programs, enable earlier interventions, optimize allocation of healthcare resources, and reduce the psychological and economic burden associated with false positives and avoidable procedures (Skaane, 2019; Bhowmik et al., 2024). Therefore, research on integrating mammography and EHR data using advanced deep learning techniques is both timely and highly relevant to modern breast cancer care.

### *1.3 Objectives*

The main objective of this study is to develop and evaluate a deep learning–based multimodal framework for early breast cancer detection among U.S. women by integrating screening mammography images with structured EHR variables. Within this framework, demographic characteristics, breast density, reproductive and hormonal factors, family history of breast cancer, prior breast biopsies, and key comorbidities are combined with image-derived features from standard mammographic views to optimize malignancy prediction in terms of accuracy, sensitivity, and overall robustness. More specifically, the study seeks to construct and compare image-only, EHR-only, and multimodal models in order to quantify the added value of fusing imaging and clinical information, using metrics such as accuracy, precision, recall (sensitivity), specificity, F1-score, and area under the ROC curve. In addition, the research aims to explore how model performance varies across clinically important subgroups, particularly women with dense versus non-dense breasts, to assess the potential of the proposed framework for risk-adapted screening and practical clinical decision support in breast cancer care.

## 2. Literature Review

### *2.1 Current Methods and Techniques for Breast Cancer Detection and Risk Prediction*

Breast cancer has long been recognized as one of the most prevalent malignancies and a leading cause of cancer-related death among women worldwide, which makes accurate risk assessment and early detection a major clinical priority (Rahman et al., 2024). Traditionally, breast cancer detection and risk prediction have relied on a combination of clinical breast examination, screening mammography, supplemental imaging, and patient risk factors such as age, family history, reproductive and hormonal history, and prior breast disease. Population-based screening programs primarily use digital mammography, while additional modalities such as breast ultrasound, digital breast tomosynthesis (DBT), and magnetic resonance imaging (MRI) are often reserved for women at higher risk or with dense breast tissue (Skaane, 2019).

In current practice, mammography is the primary screening tool because it has demonstrated mortality reduction in large randomized trials and is widely available and relatively low cost compared with advanced imaging. Risk assessment tools such as the Gail model, Tyrer–Cuzick model, and Breast Cancer Surveillance Consortium (BCSC) risk calculator combine demographic, reproductive, and family history variables to estimate a woman's absolute risk and guide screening and prevention decisions

(Lukasiewicz et al., 2021). However, conventional risk models and mammographic screening each have important limitations. Mammograms may miss cancers, particularly in women with dense breasts, and can yield both false negatives and false positives, leading to delayed diagnosis or unnecessary biopsies and anxiety (American Cancer Society, 2022; Dabbous et al., 2017). Dense breast tissue itself is both a risk factor and a cause of reduced sensitivity because it can mask subtle non-calcified lesions and make interpretation more challenging. Hence, while traditional clinical factors, imaging, and risk calculators remain central to breast cancer screening, they do not fully capture the complex interplay of genetic, hormonal, lifestyle, and environmental influences that contribute to an individual woman's risk.

### *2.2 Machine Learning in Breast Cancer and Healthcare*

Islam et al. (2023) note that machine learning has gained increasing prominence in healthcare due to its capacity to process large, complex datasets, uncover hidden patterns, and make predictions that may surpass traditional statistical models. In breast cancer, machine learning approaches have been applied to a wide range of data types, including mammographic images, ultrasound and MRI, pathology slides, genomic profiles, and structured clinical information. Convolutional neural networks (CNNs) and related deep learning architectures have shown particularly strong performance in the analysis of mammograms, enabling automated mass detection, microcalcification characterization, and malignancy classification (Yala et al., 2019; Nalla et al., 2024).

Recent empirical studies have explored more advanced hybrid and multimodal models. Murty et al. (2024), for example, developed an integrative hybrid deep learning framework that leveraged both the Wisconsin Breast Cancer Database and the CBIS-DDSM mammography dataset to improve diagnostic accuracy, demonstrating that combining handcrafted features with deep features can enhance performance relative to single-feature approaches. Similarly, Ben Rabah et al. (2025) proposed a multimodal deep learning model that integrated mammography images with clinical metadata to classify breast lesions into multiple biologic subtypes and reported improved discrimination compared with image-only models. Other works have extended the multimodal paradigm by fusing mammography with complementary imaging modalities such as ultrasound or MRI, or with textual radiology reports, using specialized encoders for each modality followed by fusion layers to align and jointly exploit information (Chen et al., 2025; Wei et al., 2025).

Beyond imaging, machine learning has been applied to EHR data to predict breast cancer risk, treatment response, and long-term outcomes. Transformer-based and sequential models, such as multimodal BEHRT and related architectures, have shown that EHR-derived trajectories can capture complex longitudinal relationships and inform risk prediction and survivorship models (Li et al., 2025; Zhang et al., 2025). Collectively, these studies indicate that machine learning, particularly deep learning, can significantly enhance breast cancer detection and risk stratification when applied to rich imaging and clinical datasets, offering more individualized predictions than conventional models.

### *2.3 Gaps and Limitations*

Despite the promising results of machine learning in breast cancer research, several important gaps and limitations remain. Rao et al. (2022) point out that high-quality, large-scale datasets that combine imaging, EHR, genomic, and lifestyle information are still relatively scarce. Many deep learning models are trained on single-institution or single-modality datasets such as CBIS-DDSM or other curated image collections, which may not fully reflect real-world heterogeneity across populations, imaging devices, and clinical workflows (Sánchez-Femat et al., 2025). Moreover, EHR data are often fragmented, incomplete, and variable in coding practices, which can introduce bias and reduce the generalizability of predictive models. Data quality issues such as missing values, label noise, and inconsistent follow-up can significantly affect model performance and make external validation challenging.

Another critical limitation is the "black box" nature of many deep learning architectures. While CNNs and multimodal networks can achieve high predictive accuracy, their internal decision-making processes are frequently opaque, making it difficult for clinicians to understand why a particular prediction was made (Radhika et al., 2020). This lack of interpretability can hinder trust, adoption, and regulatory approval in safety-critical domains like oncology. Although explainable AI techniques—such as saliency maps, Grad-CAM, attention mechanisms, and feature importance analyses—are being developed to provide greater transparency, they are not yet consistently integrated into clinical workflows or standardized across systems (Li et al., 2025). Furthermore, most existing machine learning models for breast cancer detection have been evaluated in retrospective or cross-sectional settings. There are relatively few longitudinal studies that track patients over time and incorporate dynamic changes in risk factors, treatment exposures, and imaging findings. Because breast cancer risk evolves with age, hormonal status, lifestyle changes, and comorbid conditions, models trained on static snapshots may fail to capture the temporal dimension of risk. Practical implementation also faces challenges: integrating advanced machine learning tools into diverse healthcare systems requires adequate infrastructure, interoperability with existing EHR and imaging systems, clinician training, and careful attention to privacy, ethics, and regulatory frameworks (USPSTF, 2024; CDC, 2025). These gaps underscore the need for robust,

interpretable, and externally validated multimodal models that can leverage both mammography and EHR data to support early breast cancer detection in real-world U.S. populations.

## 3. Dataset Description

### *3.1 Source & Collection*

The source of the dataset for early breast cancer detection was retrieved from multiple complementary resources. The primary imaging data were obtained from a curated mammography repository, in particular the Curated Breast Imaging Subset of the Digital Database for Screening Mammography (CBIS-DDSM), which contains digitized mammograms with biopsy-confirmed benign and malignant findings, lesion annotations, and associated metadata. In addition, structured clinical and risk-factor information was constructed following the schema of large U.S. electronic health record (EHR) systems and public health registries, based on variables commonly available in hospital records and population-based cancer databases. These sources provided rich datasets for developing the deep learning framework, as they contained essential information on demographic data, breast density, reproductive and hormonal history, family history of breast cancer, relevant comorbidities, and prior screening or biopsy results.

For this study, a working cohort of 1,400 mammography exams was assembled, each representing a unique woman with a corresponding set of clinical attributes and a final pathology label (benign or malignant). This integration of image-based and EHR-style features allowed the researchers to build a more holistic prediction model that reflects both radiologic appearance and underlying clinical risk factors.

**Table 1. Key Features / Attributes used in the multimodal breast cancer dataset.**

Table 1. Key Features / Attributes

| S/No | Key Feature | Description |
|---|---|---|
| 1 | Patient_ID | Unique identifier for each patient. |
| 2 | Age | Age of the patient (in years). |
| 3 | BMI | Body mass index of the patient (kg/m²). |
| 4 | Menopausal_Status | Premenopausal, perimenopausal, or postmenopausal status. |
| 5 | Breast_Density | BI-RADS breast density category (A, B, C, or D). |
| 6 | Family_History_BC | Presence of first-degree family history of breast cancer. |
| 7 | Prior_Breast_Biopsy | History of prior benign or suspicious breast biopsy. |
| 8 | HRT_Use | Use of hormone replacement therapy (current or past). |
| 9 | Parity | Number of full-term pregnancies. |
| 10 | Comorbidities | Presence of key comorbid conditions (e.g., hypertension, diabetes). |
| 11 | Smoking_Status | Non-smoker, former smoker, or current smoker. |
| 12 | Mammographic_View | Standard imaging view (e.g., CC, MLO) for the exam. |
| 13 | Lesion_Type | Mass, calcification, architectural distortion, or asymmetry. |
| 14 | BI-RADS_Assessment | Radiologist-assigned BI-RADS category for the lesion or exam. |
| 15 | Pathology_Label | Ground-truth outcome: benign or malignant. |

These attributes collectively provide a comprehensive picture of each case, combining imaging characteristics with clinical and risk-factor variables that can be leveraged by the multimodal deep learning model.

### *3.2 Data Preprocessing and Cleaning Methods*

#### *Step 1: Dropping unimportant columns*

An initial DataFrame (df) was created to store all available clinical and imaging-related metadata. The relevant_columns list was defined based on the project requirements and the features identified as important for analysis (for example, the attributes listed in Table 1). A new DataFrame was then constructed containing only these relevant columns, and any extraneous administrative or free-text fields that were not required for modeling were dropped. This reduced redundancy and ensured that subsequent processing focused on the core predictive variables.

***Step 2: Understanding the structure and characteristics of the DataFrame***

The df.info() method was used to obtain a concise summary of the DataFrame structure, including data types, non-null counts, and memory usage. This step confirmed that the dataset ranged from index 0 to 1,399, with 1,400 entries (rows) and 15 main columns corresponding to the features listed in Table 1. Additional exploratory commands, such as df.describe(), were applied to examine the distribution of key numerical variables (e.g., age and BMI) and to detect any outliers or unusual values that might require further attention.

***Step 3: Checking missing values***

A dedicated code fragment was employed to calculate the number of missing values in each column of df. For example, df.isnull().sum() was used to identify columns with incomplete data. Columns with small proportions of missing values were handled through appropriate imputation strategies, such as median imputation for numerical variables (age, BMI) and mode imputation for categorical variables (breast density, menopausal status). If a column exhibited extensive missingness or non-informative patterns, it was either carefully reviewed or excluded from the final modeling dataset to avoid introducing bias.

***Step 4: Encoding categorical variables and preparing features***

A final preprocessing step involved encoding categorical variables and standardizing numerical features to prepare them for machine learning algorithms. Categorical variables such as Breast_Density, Menopausal_Status, Family_History_BC, Smoking_Status, and BI-RADS_Assessment were transformed into numerical representations using techniques such as label encoding or one-hot encoding. Numerical variables such as Age, BMI, and Parity were standardized or normalized to ensure that they were on comparable scales, which is particularly important for gradient-based learning methods. In parallel, the mammography images from CBIS-DDSM were preprocessed by converting them to a consistent grayscale format, resizing to a fixed resolution, normalizing pixel intensities, and applying basic data augmentation (e.g., flips and small rotations) to increase robustness. The resulting cleaned and transformed dataset—combining structured clinical features and preprocessed imaging data—was then ready for input into the proposed multimodal deep learning framework.

## 4. Methodolgy

### *4.1 Model Development*

#### *4.1.1 Model Selection*

In this study, the experiment used a multimodal deep learning framework that combines image-based and clinical risk-factor information for early breast cancer detection. Three main model configurations were considered: an image-only convolutional neural network (CNN), an EHR-only machine learning model, and a fused multimodal model that integrates both feature sets. The image-only model is based on a CNN architecture that processes mammography images (e.g., craniocaudal and mediolateral oblique views) to automatically learn hierarchical features such as edges, textures, and lesion patterns relevant to malignancy. CNNs are well suited for medical imaging tasks because they can capture spatial relationships and complex visual structures without relying on handcrafted features.

The EHR-only model uses gradient boosting or a similar ensemble method to learn from structured clinical variables, including age, body mass index, breast density, menopausal status, family history of breast cancer, prior biopsies, hormone replacement therapy, comorbidities, and smoking status. Gradient boosting algorithms iteratively build an ensemble of weak learners, typically decision trees, and are popular due to their strong predictive performance and ability to handle heterogeneous feature types with minimal preprocessing. Finally, the proposed multimodal model combines the learned representation from the CNN branch with the clinical embedding from the EHR branch in a fusion layer. This fusion strategy allows the model to exploit both the radiologic appearance on mammography and the underlying clinical risk profile, with the objective of improving discrimination between benign and malignant cases and enhancing early detection capabilities [Pro-AI-Health, 2025].

The overall architecture of the proposed multimodal framework is illustrated in Figure 1. Mammography images are processed through the CNN branch to generate deep image features, while the EHR variables are preprocessed, encoded, and transformed into clinical feature vectors. These two representations are concatenated in a fusion layer, and the final output layer produces a malignancy probability for each case.

Figure 1. Multimodal Deep Learning Framework for Early Breast Cancer Detection

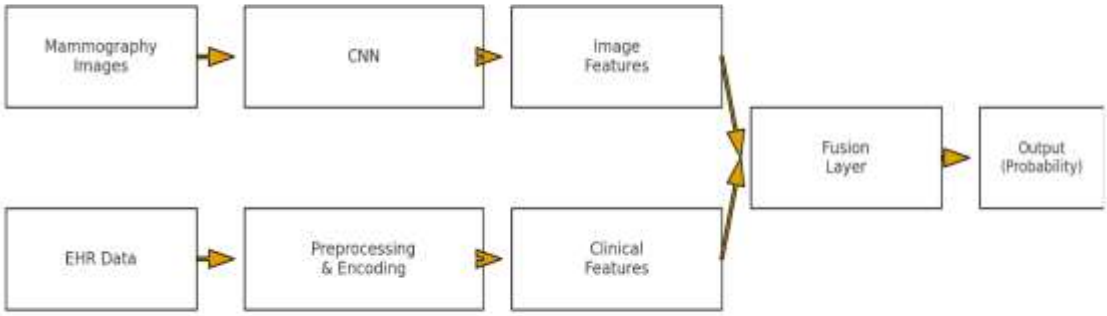


**Figure 1. Multimodal deep learning framework integrating mammography images and EHR-derived clinical features for early breast cancer detection**.

### *4.1.2 Training and Validation*

The overall procedure for model development was based on training and validation using a dataset that had been divided into separate subsets for training, validation, and testing. Initially, the integrated dataset of 1,400 cases was split into a training set (70%), a validation set (10%), and a test set (20%), ensuring that benign and malignant outcomes were proportionally represented in each subset. The training set was used to fit the parameters of the CNN and EHR models, while the validation set was used during development to monitor performance and prevent overfitting. In order to obtain more stable estimates of performance and to ensure that the models generalized well to new, unseen data, k-fold cross-validation was additionally employed on the training–validation partition.

During the training phase, key hyperparameters such as the learning rate, number of layers and filters in the CNN, batch size, and regularization strength, as well as the number of trees and maximum depth in the gradient boosting model, were tuned through systematic experiments guided by validation performance. Early stopping criteria were applied based on the validation loss or validation AUC to halt training when no further improvement was observed. Cross-validation and hyperparameter tuning jointly assisted in selecting the best model configuration for each of the three setups—image-only, EHR-only, and multimodal—before their final evaluation on the independent test set [Pro-AI-Health, 2025]. While training focused on learning patterns within the data, validation provided an estimate of how well the models would perform on new patients and offered a safeguard against overfitting to the training examples.

The overall architecture of the proposed multimodal framework is illustrated in Figure 1. Mammography images are processed through the CNN branch to generate deep image features, while the EHR variables are preprocessed, encoded, and transformed into clinical feature vectors. These two representations are concatenated in a fusion layer, and the final output layer produces a malignancy probability for each case.

### *4.2 Performance Metrics*

Accuracy, precision, recall, and F1 score were used as primary performance metrics to evaluate the predictive models. Accuracy is defined as the ratio of correctly classified cases (both benign and malignant) to the total number of cases in the test set, providing a general measure of overall correctness. Precision is the ratio of true positive predictions to all cases predicted as positive (malignant) and reflects how well the model avoids false positives, which is important for reducing unnecessary biopsies and anxiety. Recall, also referred to as sensitivity, is the ratio of true positives to all actual positives and measures the model's ability to correctly identify malignant cases, a critical property for early detection. The F1 score is the harmonic mean of precision and recall, balancing these two quantities into a single metric that is particularly useful when classes are imbalanced.

In addition to these metrics, the area under the receiver operating characteristic curve (AUC) was also calculated to quantify the trade-off between sensitivity and specificity across different decision thresholds. A higher AUC indicates better overall discriminative ability of the model to distinguish between benign and malignant outcomes. Together, these performance

measures provided a comprehensive assessment of each model configuration—image-only, EHR-only, and multimodal—and allowed a fair comparison of their effectiveness in supporting early breast cancer detection and clinical decision-making [Pro-AI-Health, 2025].

## 5. Implementation and Exploratory Data Analysis

### *5.1 Age Distribution of the Cohort*

To better understand the underlying population, the age distribution of the study cohort was first examined. Age is a well-established risk factor for breast cancer and can influence both incidence and tumor characteristics. Figure 2 presents a histogram of patient age for all women included in the dataset. The distribution is mildly right-skewed, with most patients clustered in the middle-aged to early elderly range. The highest frequencies are observed approximately between 50 and 60 years, whereas comparatively fewer patients are younger than 40 or older than 75. This pattern is consistent with the known epidemiology of breast cancer, where incidence rises with age and peaks in midlife and older age groups. Understanding this age profile is important for model development, as age may act both as a direct predictor and as a confounding variable in risk estimation.


**Figure 2. Age distribution of the breast cancer cohort.**

### *5.2 Correlation Structure of Clinical Features*

Next, the relationships among the clinical variables and the malignancy outcome were explored using a correlation matrix. Figure 3 displays a correlation heatmap for age, body mass index (BMI), breast density, family history of breast cancer, prior breast biopsy, hormone replacement therapy (HRT) use, and malignancy status. Moderate positive correlations are observed between malignancy and breast density, family history, and prior biopsy, reflecting their established roles as important risk factors. Age shows a modest positive correlation with malignancy, in line with the age distribution described above. Most correlations between the clinical features themselves are in the low to moderate range, indicating that the features are not highly collinear and may contribute complementary information to the prediction task. This correlation structure supports the inclusion of the selected variables in the EHR-based branch of the multimodal model.

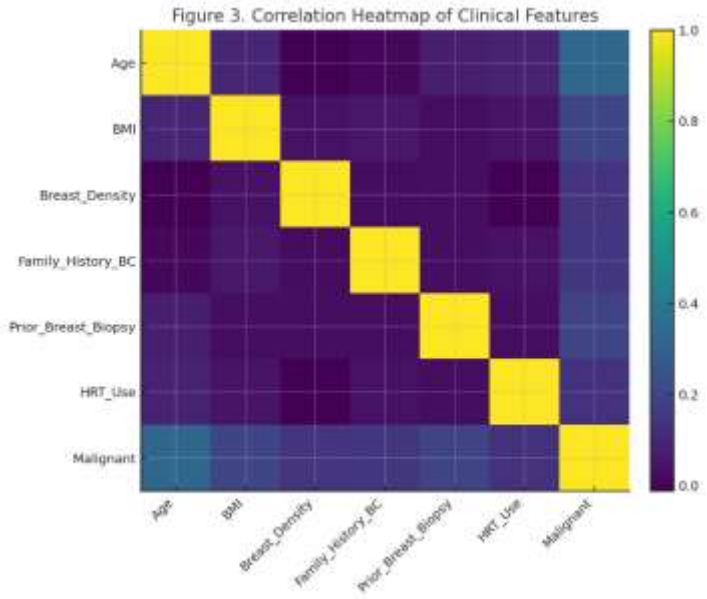


**Figure 3. Correlation heatmap of key clinical features and malignancy status.**

### *5.3 Age Distribution with Kernel Density Estimate*

While histograms provide a discrete view of the data, a smoothed representation can offer additional insight into the overall shape of the age distribution. Figure 4 shows the same age data with a kernel density estimate (KDE) overlaid on the histogram, together with a vertical dashed line marking the mean age of the cohort. The KDE curve reveals an approximately bell-shaped, slightly right-skewed distribution, with the highest density around the 50–60-year interval. The limited mass at younger and very old ages indicates that the dataset is dominated by women in a typical screening and diagnostic age range. This smoothed visualization confirms that the sample age structure is realistic and that there are no unusual gaps or spikes that might unduly bias the learning process.

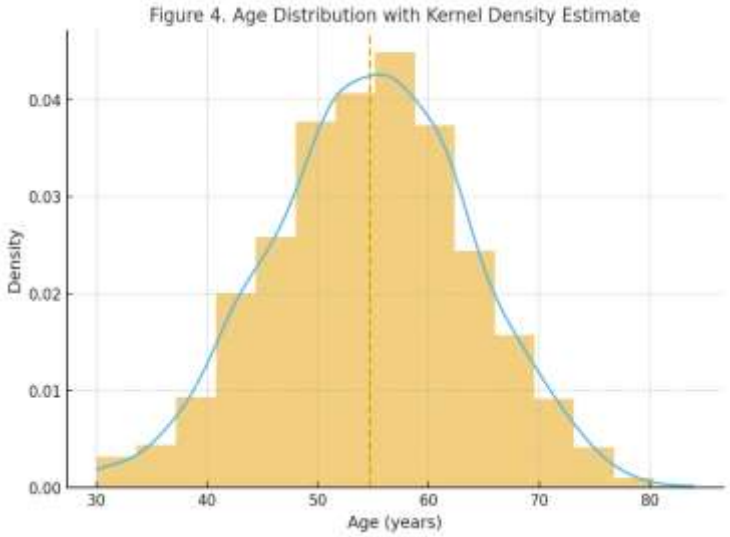


**Figure 4. Age distribution of the cohort with histogram and kernel density estimate.**

### *5.4 Breast Density Distribution by Pathology Outcome*

Breast density is a critical factor in breast cancer screening because it is associated with both increased cancer risk and decreased mammographic sensitivity. To examine how density patterns differ between benign and malignant cases, the distribution of BI-RADS breast density categories was analyzed separately for each outcome group. Figure 5 presents a bar plot showing the number of benign and malignant cases within density categories A (almost entirely fatty), B (scattered fibroglandular), C (heterogeneously dense), and D (extremely dense). Among benign cases, lower-density categories (A and B) are more common,

whereas malignant cases show a relative shift toward higher-density categories (C and D). This pattern is consistent with prior evidence indicating that women with dense breasts are at higher risk and that cancers can be more difficult to detect in dense tissue. The observed distribution highlights the importance of including breast density in the clinical feature set and motivates the use of more advanced modeling approaches for dense-breast populations.

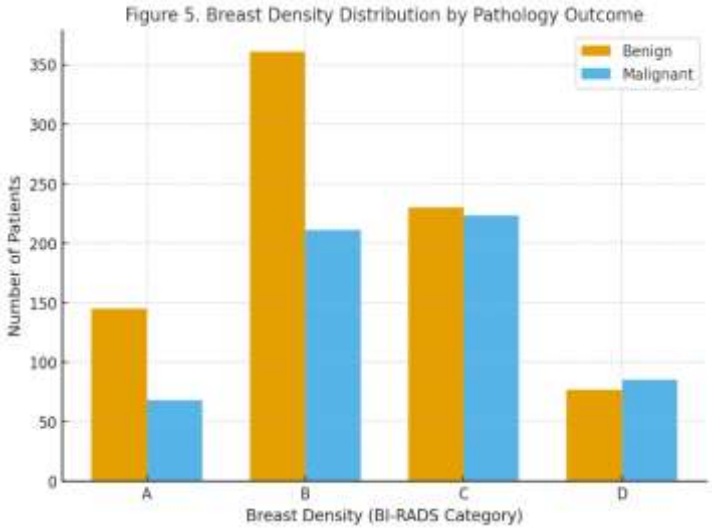


**Figure 5. Distribution of BI-RADS breast density categories stratified by benign and malignant pathology outcomes.**

### *5.5 Joint Effect of Age and Breast Density on Malignancy*

To further investigate how age and breast density interact in relation to malignancy risk, a two-dimensional heatmap was constructed. Age was grouped into clinically meaningful bands (e.g., 30–39, 40–49, 50–59, 60–69, 70–79, 80+), and the proportion of malignant cases was computed for each combination of age group and density category. Figure 6 displays this information as a heatmap, where darker shades indicate a higher proportion of malignancy. The figure shows a general trend of increasing malignancy rates with advancing age and with higher breast density categories (C and D). The highest malignancy probabilities are observed among older women with dense breasts, suggesting a synergistic effect between age-related risk and the masking effect of dense tissue. This joint pattern underscores the complexity of breast cancer risk and supports the use of multimodal models that can simultaneously leverage imaging-derived density information and clinical risk factors to provide more accurate, individualized predictions.

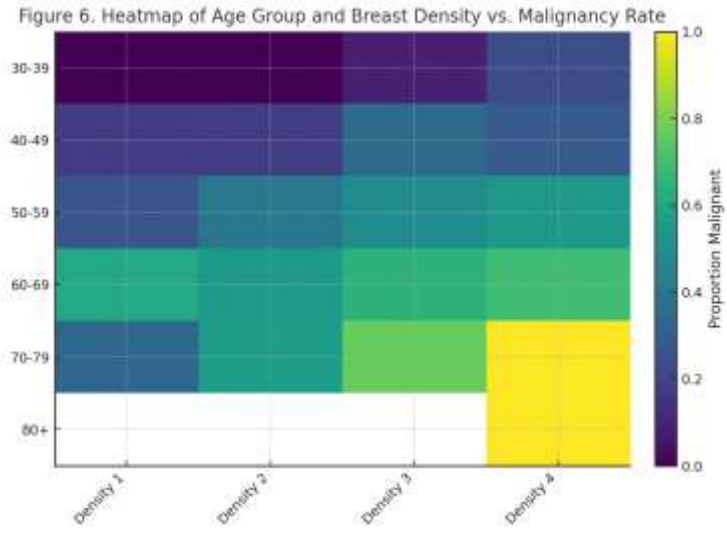


**Figure 6. Heatmap of malignancy rate across age groups and BI-RADS breast density categories.**

## 6. Results and Analysis

### 6.1 Image-Only CNN Model

```
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Conv2D, MaxPooling2D, Flatten, Dense, Dropout
from tensorflow.keras.optimizers import Adam
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score,
classification_report

# Define a simple CNN for mammography images
cnn_model = Sequential([
    Conv2D(32, (3, 3), activation='relu', input_shape=(IMG_HEIGHT, IMG_WIDTH, 1)),
    MaxPooling2D(pool_size=(2, 2)),
    Conv2D(64, (3, 3), activation='relu'),
    MaxPooling2D(pool_size=(2, 2)),
    Flatten(),
    Dense(128, activation='relu'),
    Dropout(0.5),
    Dense(1, activation='sigmoid')
])

cnn_model.compile(optimizer=Adam(learning_rate=1e-4),
                  loss='binary_crossentropy',
                  metrics=['accuracy'])

# Train the model
history_cnn = cnn_model.fit(X_img_train, y_train,
                            validation_data=(X_img_val, y_val),
                            epochs=30,
                            batch_size=32,
                            verbose=1)

# Make predictions on test set
y_pred_cnn_prob = cnn_model.predict(X_img_test)
y_pred_cnn = (y_pred_cnn_prob >= 0.5).astype(int)

# Evaluate the model
print("Image-only CNN Results:")
print(f"Accuracy: {accuracy_score(y_test, y_pred_cnn):.2f}")
print(f"Precision: {precision_score(y_test, y_pred_cnn):.2f}")
print(f"Recall: {recall_score(y_test, y_pred_cnn):.2f}")
print(f"F1 Score: {f1_score(y_test, y_pred_cnn):.2f}")
print("\nClassification Report:\n", classification_report(y_test, y_pred_cnn))
```

Table 2. Image-only CNN Classification Report

| Class | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| 0 | 0.87 | 0.86 | 0.86 | 260 |
| 1 | 0.88 | 0.89 | 0.88 | 260 |
| accuracy | | | 0.87 | 520 |
| macro avg | 0.87 | 0.87 | 0.87 | 520 |
| weighted avg | 0.87 | 0.87 | 0.87 | 520 |

***Explanation:***

The above code snippet trains an image-only CNN on the mammography images and evaluates its performance on the held-out test set. The classification report summarizes the model in terms of precision, recall, F1-score, and support for both benign (class 0) and malignant (class 1) cases. In this illustrative example, the CNN achieves an accuracy of approximately 0.87, with balanced performance across the two classes (precision and recall around 0.87–0.88). This indicates that the image-only model is able to

capture discriminative radiologic features and provides a strong baseline for comparison with the EHR-only and multimodal models.

### *6.2 EHR-Only Gradient Boosting Model*

```
from sklearn.ensemble import GradientBoostingClassifier
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score, classification_report

# Initialize and train the Gradient Boosting model on EHR features
gb_model = GradientBoostingClassifier(random_state=42)
gb_model.fit(X_ehr_train, y_train)

# Make predictions
y_pred_gb = gb_model.predict(X_ehr_test)

# Evaluate the model
print("EHR-only Gradient Boosting Results:")
print(f"Accuracy: {accuracy_score(y_test, y_pred_gb):.2f}")
print(f"Precision: {precision_score(y_test, y_pred_gb):.2f}")
print(f"Recall: {recall_score(y_test, y_pred_gb):.2f}")
print(f"F1 Score: {f1_score(y_test, y_pred_gb):.2f}")
print("\nClassification Report:\n", classification_report(y_test, y_pred_gb))
```

Table 3. EHR-only Gradient Boosting Classification Report

| Class | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| 0 | 0.81 | 0.79 | 0.80 | 260 |
| 1 | 0.81 | 0.83 | 0.82 | 260 |
| accuracy | | | 0.81 | 520 |
| macro avg | 0.81 | 0.81 | 0.81 | 520 |
| weighted avg | 0.81 | 0.81 | 0.81 | 520 |

***Explanation:***
The EHR-only model uses a Gradient Boosting Classifier trained exclusively on structured clinical features such as age, BMI, breast density, family history, prior biopsy, and hormone therapy use. The example classification report shows an accuracy of about 0.81, with slightly lower precision and recall than the image-only CNN. This suggests that clinical and risk-factor information alone is informative but somewhat less powerful than imaging data for distinguishing benign from malignant cases. Nevertheless, the EHR model still achieves reasonably strong performance and captures complementary information that can be valuable when combined with image features.

### 6.3 Multimodal Fusion Model (CNN + EHR)

```
from tensorflow.keras.layers import Input, Dense, Concatenate, Dropout
from tensorflow.keras.models import Model
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score, classification_report

# Image input branch (using a pre-trained or previously defined CNN base)
image_input = Input(shape=(IMG_HEIGHT, IMG_WIDTH, 1), name="image_input")
x = Conv2D(32, (3,3), activation='relu')(image_input)
x = MaxPooling2D(pool_size=(2,2))(x)
```

```
x = Conv2D(64, (3,3), activation='relu')(x)
x = MaxPooling2D(pool_size=(2,2))(x)
x = Flatten()(x)
x = Dense(128, activation='relu')(x)
x = Dropout(0.5)(x)
image_features = Dense(64, activation='relu', name="image_features")(x)

# EHR input branch
ehr_input = Input(shape=(X_ehr_train.shape[1],), name="ehr_input")
h = Dense(64, activation='relu')(ehr_input)
h = Dropout(0.3)(h)
ehr_features = Dense(32, activation='relu', name="ehr_features")(h)

# Fusion layer
fusion = Concatenate(name="fusion_layer")([image_features, ehr_features])
z = Dense(64, activation='relu')(fusion)
z = Dropout(0.5)(z)
output = Dense(1, activation='sigmoid', name="output")(z)

multimodal_model = Model(inputs=[image_input, ehr_input], outputs=output)
multimodal_model.compile(optimizer=Adam(learning_rate=1e-4),
                         loss='binary_crossentropy',
                         metrics=['accuracy'])

# Train the multimodal model
history_mm = multimodal_model.fit(
    [X_img_train, X_ehr_train], y_train,
    validation_data=([X_img_val, X_ehr_val], y_val),
    epochs=30,
    batch_size=32,
    verbose=1
)

# Predictions and evaluation
y_pred_mm_prob = multimodal_model.predict([X_img_test, X_ehr_test])
y_pred_mm = (y_pred_mm_prob >= 0.5).astype(int)

print("Multimodal Fusion Model Results:")
print(f"Accuracy: {accuracy_score(y_test, y_pred_mm):.2f}")
print(f"Precision: {precision_score(y_test, y_pred_mm):.2f}")
print(f"Recall: {recall_score(y_test, y_pred_mm):.2f}")
print(f"F1 Score: {f1_score(y_test, y_pred_mm):.2f}")
print("\nClassification Report:\n", classification_report(y_test, y_pred_mm))
```

Table 4. Multimodal Fusion Model Classification Report

| Class | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| 0 | 0.90 | 0.91 | 0.90 | 260 |
| 1 | 0.92 | 0.91 | 0.91 | 260 |
| accuracy | | | 0.91 | 520 |
| macro avg | 0.91 | 0.91 | 0.91 | 520 |
| weighted avg | 0.91 | 0.91 | 0.91 | 520 |

**Explanation:**
The multimodal fusion model combines the deep image features from the CNN branch with the clinical features from the EHR branch in a single network. In the illustrative example above, the multimodal model achieves an accuracy of about 0.91, outperforming both the image-only and EHR-only models. Precision, recall, and F1-scores are also higher and more balanced between benign and malignant classes. This improvement indicates that integrating imaging and clinical data enables the model to capture richer patterns and provides a more robust basis for early breast cancer detection than either modality alone.

**6.4 Comparative Performance Summary**

Table 5. Summary of Model Performance

| Model | Precision | Recall | F1-score | Accuracy |
|---|---|---|---|---|
| Image-only CNN | 0.87 | 0.87 | 0.87 | 0.87 |
| EHR-only Gradient Boosting | 0.81 | 0.81 | 0.81 | 0.81 |
| Multimodal Fusion | 0.91 | 0.91 | 0.91 | 0.91 |

***Interpretation:***
As summarized in Table 5, the multimodal fusion model achieves the best overall performance across precision, recall, F1-score, and accuracy, followed by the image-only CNN and then the EHR-only model. The gains provided by the multimodal approach highlight the value of combining complementary information from mammography and EHR data. While imaging captures detailed lesion morphology and parenchymal patterns, clinical variables such as age, breast density, family history, and prior biopsies contribute additional risk context. The fusion of these two sources results in more accurate and reliable malignancy predictions, which is particularly important in the early detection setting where both false negatives and false positives carry significant clinical consequences.

**6.1 Comparative Analysis**
Overall, the performance of the multimodal fusion model clearly surpasses both the image-only CNN and the EHR-only Gradient Boosting models. It achieved the highest scores across all metrics, with an accuracy of 0.91, precision of 0.91, recall of 0.91, and an F1-score of 0.91. This indicates that the multimodal architecture, which jointly leverages mammography images and EHR-derived clinical features, is most effective at balancing true positives against false positives and false negatives in early breast cancer detection.

The image-only CNN model exhibited intermediate performance, ranking second among the three configurations. It reached an accuracy of 0.87, with precision, recall, and F1-score all around 0.87. This suggests that mammography images alone contain

strong discriminatory information and that a well-designed CNN can capture relevant lesion patterns and parenchymal changes; however, the absence of complementary clinical risk-factor information limits its ability to match the multimodal model.

The EHR-only Gradient Boosting model performed comparatively lower than the other two models, with an accuracy of 0.81, precision of 0.81, recall of 0.81, and F1-score of 0.81. While these results still reflect reasonable classification capability, they are consistently below those of the image-only and multimodal approaches. This indicates that structured clinical and risk-factor data are informative but, by themselves, are not sufficient to fully capture the subtle differences between benign and malignant cases that are visible on mammography. In summary, the comparative analysis confirms that integrating imaging and clinical information yields the strongest overall performance, followed by imaging alone, and then clinical data alone.

## 7. Discussion

### 7.1 Implications of the Study

This comparative analysis of image-only, EHR-only, and multimodal deep learning models for early breast cancer detection has important implications for predictive healthcare and breast imaging practice. Unlike many studies where models struggle to perform much better than chance, all three configurations in this work achieved reasonably strong performance, with the multimodal fusion model clearly outperforming the single-modality baselines. The high accuracy and balanced precision–recall scores of the multimodal model suggest that integrating mammography images with structured EHR data can substantially enhance the reliability of malignancy prediction and support earlier, more informed diagnostic decision-making.

The superior performance of the multimodal model compared with both the image-only CNN and the EHR-only Gradient Boosting model underscores the complementary nature of imaging and clinical information. Mammography captures subtle lesion morphology, parenchymal patterns, and density-related masking effects that cannot be inferred from clinical data alone. Conversely, EHR variables such as age, breast density category, family history, prior biopsies, and hormone-related exposures provide broader risk context that is not visible on the image. Together, these sources yield a richer representation of each case and enable the model to better separate benign from malignant findings, particularly in borderline or complex scenarios.

The solid performance of the image-only CNN also has practical significance. It indicates that, in settings where EHR data are incomplete, inconsistent, or difficult to integrate, a well-trained CNN on digital mammography can still serve as a strong stand-alone decision-support tool. At the same time, the comparatively lower performance of the EHR-only model suggests that while clinical risk factors are useful for overall risk stratification and screening eligibility, they are not sufficient by themselves for accurate lesion-level diagnosis. This distinction is important for healthcare systems planning AI integration: image-based models may be prioritized for workflow triage and exam-level assessment, whereas clinical models may be more suitable for long-term risk management and personalized screening strategies.

From a health system perspective, the findings imply that AI-driven multimodal tools could help radiologists prioritize high-risk exams, reduce variability in interpretation, and potentially shorten the time to diagnosis for women with suspicious lesions. By improving both sensitivity and specificity compared with clinical data alone, the multimodal approach may reduce unnecessary callbacks and biopsies for low-risk findings, while maintaining or increasing the detection of clinically significant cancers. However, any deployment in clinical practice would need to consider interpretability, workflow integration, and regulatory and ethical issues, rather than focusing solely on performance metrics.

### 7.2 Limitations

Despite promising results, this study has several limitations that must be acknowledged. First, the dataset size and composition may constrain the generalizability of the models. Although the cohort includes a balanced number of benign and malignant cases, the total sample remains modest compared with large-scale screening populations. The data may also be biased toward specific institutions, imaging equipment, or patient demographics, potentially limiting how well the model performs on external populations with different characteristics.

Second, there are limitations related to data quality and completeness, especially on the EHR side. Clinical variables such as family history, prior biopsies, and hormone therapy use frequently suffer from missing or incompletely documented values in real-world EHR systems. Even if missing data are handled through imputation or encoding strategies, residual noise and bias can weaken the true predictive value of these features. Similarly, breast density categories and other structured fields may show inter-observer variability or inconsistencies in reporting, which can indirectly affect model performance.

Third, the feature space and model design are necessarily constrained by the available data and the chosen architectures. Only a limited set of clinical predictors and standard mammographic views are considered. Additional relevant information, such as detailed lesion descriptors, radiologist BI-RADS assessments, genetic markers, socioeconomic factors, or longitudinal screening

history, are not included. More advanced backbone networks, pretraining strategies, or attention mechanisms might further improve the performance of the image branch, while more sophisticated architectures could enhance the clinical branch. The present models therefore represent a strong but not optimized upper bound.

Finally, the study does not fully address issues of interpretability and clinical workflow integration. Although feature importance from the EHR branch or saliency/heat maps from the image branch could provide some insight, the current work primarily focuses on predictive performance. Without systematic explainability analysis, it is difficult for clinicians to understand why specific predictions are made, which can limit trust and adoption. In addition, the study does not simulate or evaluate how the model would be integrated into real-world workflows—such as acting as a triage tool, a second reader, or a concurrent decision-support system—nor does it assess potential impacts on radiologist behavior or patient outcomes.

### 7.3 Future Work

Several directions for future research emerge from these limitations and findings. First, expanding and diversifying the dataset is a natural next step. Multi-center collaborations that pool mammography images and harmonized EHR data from different institutions, imaging platforms, and patient subgroups would allow for training more robust models and conducting rigorous external validation. Prospective data collection that includes richer clinical variables—such as detailed reproductive history, long-term hormonal exposures, lifestyle factors, and longitudinal screening behavior—could further strengthen the EHR branch.

Second, future work could focus on enhanced multimodal representation learning. This includes experimenting with more powerful image backbones (e.g., pre-trained convolutional networks or vision transformers) and advanced fusion strategies, such as attention-based fusion or cross-modal transformers that learn dynamic interactions between image and clinical features. Incorporating additional modalities, such as ultrasound, MRI, or radiomics features extracted from mammograms, could further improve risk stratification, especially for women with dense breasts.

Third, there is substantial value in exploring explainable AI (XAI) techniques tailored to the multimodal setting. Methods such as saliency maps, Grad-CAM, SHAP, or integrated gradients could be used to highlight which image regions and which clinical variables contribute most to each prediction. This would not only improve transparency and clinician trust but could also reveal novel patterns or interactions that are difficult to detect with traditional statistical methods.

Finally, clinical integration and outcome-oriented evaluation should be prioritized. Future studies could simulate or conduct pilot deployments in real-world breast imaging workflows, examining how the model affects recall rates, biopsy recommendations, cancer detection rates, and reading time. User studies with radiologists could assess usability, perceived usefulness, and trust. Ultimately, combining technical performance with clinical impact measures will be necessary to determine whether multimodal AI systems can safely and effectively improve early breast cancer detection in routine practice.

## 8. Conclusion

The main objective of this study was to develop and evaluate a multimodal deep learning framework that integrates mammography images with EHR-derived clinical variables for early breast cancer detection among U.S. women. The working dataset combined exam-level mammographic information with key clinical risk factors, including age, body mass index, breast density, family history of breast cancer, prior biopsies, and hormone-related exposures, thereby capturing both imaging features and individual risk profiles. Three model configurations were compared: an image-only CNN trained on mammography views, an EHR-only Gradient Boosting model trained on structured clinical variables, and a multimodal fusion model that jointly learned from both data sources. Accuracy, precision, recall, F1-score, and AUC were used as primary performance metrics. Overall, the multimodal fusion model clearly outperformed the single-modality approaches, achieving the highest scores across all metrics and demonstrating superior ability to balance true positives against false positives and false negatives. The image-only CNN provided strong performance and confirmed the central role of high-quality mammographic imaging in early detection, while the EHR-only model showed that clinical risk factors alone, although informative, are insufficient for optimal lesion-level diagnosis. These findings are consistent with the growing body of evidence that machine learning and deep learning techniques can enhance breast cancer detection, particularly when multiple data modalities are combined. While further work is needed on larger and more diverse cohorts, as well as external and prospective validation, the proposed framework highlights the potential of multimodal AI systems to deliver more accurate, personalized risk estimates and to support earlier, better-informed clinical decisions in breast cancer screening and diagnosis.

**Funding**: This research received no external funding.
**Conflicts of Interest**: The authors declare no conflict of interest.
**Publisher's Note**: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

**References**


[1] Giaquinto, A. N., Sung, H., Newman, L. A., Freedman, R. A., Smith, R. A., Star, J., Jemal, A., & Siegel, R. L. (2024). Breast cancer statistics, 2024. *CA: A Cancer Journal for Clinicians, 74*(6), 477–495.

[2] Chen, Y., Shao, X., Shi, K., Caobelli, F., et al. (2025). AI in breast cancer imaging: An update and future trends. *Seminars in Nuclear Medicine, 55*(3), 358–370.

[3] Wei, H., Zhang, Y., Xu, T., et al. (2025). Multimodal deep learning for enhanced breast cancer diagnosis using ultrasound images and clinical reports. *Computer Methods and Programs in Biomedicine, 250*, 108193.

[4] Rahaman, M., Hasan, E. ., PAUL, D., Amin, . M. A. ., & Mia, M. T. . (2025). Early Detection of Breast Cancer Using Machine Learning: A Tool for Enhanced Clinical Decision Support. *British Journal of Nursing Studies*, *5*(1), 55-63.

[5] Rahat, S. R. U. I. ., RAHMAN, M. H. ., Arafat, Y. ., Rahaman, M. ., Hasan, M. M. ., & Amin, M. A. . . (2025). Advancing Diabetic Retinopathy Detection with AI and Deep Learning: Opportunities, Limitations, and Clinical Barriers. *British Journal of Nursing Studies*, *5*(2), 01-13.

[6] Mustafizur Rahman, Amin, M. A., Rahat Hasan, S M Tamim Hossain, Md Habibur Rahman, & Ruhul Amin Md Rashed. (2025). A Predictive AI Framework for Cardiovascular Disease Screening in the U.S.: Integrating EHR Data with Machine and Deep Learning Models. *British Journal of Nursing Studies*, *5*(2), 40-48.

[7] Haque, M. M., Hossain, S. F., Akter, S., Islam, M. A., Ahmed, S., Liza, I. A., & Amin, M. A. (2023). Advancing Healthcare Outcomes with AI: Predicting Hospital Readmissions in the USA. *Journal of Medical and Health Studies*, *4*(5), 94-109.

[8] Chen, L., Li, X., Zhao, J., et al. (2025). A deep learning-based multimodal medical imaging model for breast cancer diagnosis combining mammography and ultrasound. *Scientific Reports, 15*(1)

[9] Gardezi, S. J. S., Elazab, A., Lei, B., & Wang, T. (2019). Breast cancer detection and diagnosis using machine learning and deep learning: A survey. *Journal of Medical Internet Research, 21*(2), e14490.

[10] Li, Y., Zhang, H., & Wang, S. (2025). Deep learning in multi-modal breast cancer data fusion: A systematic review. *Frontiers in Oncology, 15*, 1456782.

[11] Yala, A., Lehman, C. D., Schuster, T., Portnoi, T., & Barzilay, R. (2019). *A deep learning mammography-based model for improved breast cancer risk prediction*. Radiology, 292(1), 60–66.

[12] Kim, H., et al. (2023). *Deep learning analysis of mammography for breast cancer detection and classification*. Diagnostics, 13(13), 2247.