RESEARCH & INNOVATION DIVISION
IJIST

AI-Powered Detection: Implementing Deep Learning for Breast Cancer Prediction

Anmol Tanveer[1], Saima Munawar[1*], Nasir Naveed[1]
[1]Faculty of Computer Science and Information Technology, Virtual University of Pakistan, Lahore
*Correspondence: Dr. Saima Munawar, saima.munawar@vu.edu.pk

Citation| Tanveer. A, Munawar. S, Naveed. N, "AI-Powered Detection: Implementing Deep Learning for Breast Cancer Prediction", IJIST, Vol. 6 Issue. 3 pp 1488-1504, Sep 2024
Received| Aug 30, 2024 **Revised**| Sep 20, 2024 **Accepted**| Sep 21, 2024 **Published**| Sep 22, 2024.

Breast cancer remains a critical global health issue, affecting millions of women worldwide. According to the World Health Organization (WHO), there were 2.3 million new cases and 685,000 deaths from breast cancer in 2020 alone. This makes breast cancer the most prevalent cancer globally, with 7.8 million cases diagnosed over the past five years. As the prevalence of breast cancer continues to rise, the need for accurate and efficient diagnostic tools becomes increasingly urgent. Artificial Intelligence (AI) has shown considerable promise in enhancing breast cancer detection and diagnosis. Over the past two decades, AI tools have increasingly aided physicians in interpreting mammograms, offering the potential for automated, precise, and early cancer detection. However, significant challenges remain, particularly concerning data imbalance in datasets—where cancerous images are often underrepresented—and the issue of low pixel resolution, which can obscure crucial details in medical images. This work utilizes a subset of the data called Mini-DDSM, a lightweight version of the Digital Database for Screening Mammography. To address these challenges, our research employed the Neighborhood Cleaning Rule (NCR) algorithm from the imbalance library, designed to mitigate data imbalance by refining the dataset through the selective removal of noisy and borderline examples. This method enhances the quality of training data, enabling AI models to learn more effectively. We developed a deep learning model that incorporates a transfer learning layer (DenseNet121), dense layers, a global pooling layer, and a dropout layer to optimize performance. This model demonstrated promising results, effectively addressing the challenges of data imbalance and low image resolution. Our approach underscores the potential of AI to significantly improve breast cancer detection and diagnosis, ultimately leading to better patient outcomes. Continued research and refinement of AI techniques will be crucial in overcoming remaining challenges and fully realizing the potential of these technologies in healthcare.

Keywords: Breast Cancer, Deep Learning, Artificial Intelligence, DenseNet121, Mammography Images

Introduction:

Breast Cancer (BC) remains a significant global health challenge, particularly affecting women. Each year, an alarming number of invasive breast cancer cases are diagnosed, with the mortality rate steadily increasing, making it the second most life-threatening illness among women. By 2040, the number of new breast cancer cases is predicted to exceed 3 million annually, with the death toll potentially reaching 1 million due to factors such as population growth and aging. Research indicates that benign tumors, such as adenomas, fibroids, and lipomas, are non-cancerous, while malignant tumors, including those associated with breast, lung, and colorectal cancers, are cancerous. Treatment strategies vary depending on the tumor type. Breast cancer, a malignant tumor, typically forms in or near breast tissue, particularly within the milk ducts and glands. It often begins as a lump or calcium deposit resulting from abnormal cell growth. Although most breast lumps are benign, some can be malignant or premalignant, potentially leading to cancer if not properly diagnosed and treated [1] [2]. A recent study conducted at the Sindh Institute of Urology and Transplantation in Karachi, covering data from March 2017 to December 2021, revealed that out of 690 patients, 99% were female, with a mean age of 49.3 years. The most common stage at presentation was stage II (48.6%), and grade II invasive ductal carcinoma was the predominant histopathological finding (57.2%). Traditional diagnostic methods struggle with issues such as limited resolution and variability in interpretation, impacting the accuracy and timeliness of diagnoses. AI offers a promising solution by enhancing the diagnostic process through deep learning models and automated image analysis. By integrating AI into medical imaging, there is potential for more precise and early detection of breast cancer, ultimately leading to better-targeted treatments and improved patient outcomes.

Breast cancer is characterized by uncontrolled cell division in breast tissue, leading to tumor formation. Common symptoms include discomfort, changes in skin color, mass formation, and alterations in breast shape and size. To diagnose breast cancer, medical practitioners commonly use mammography, which involves brief-intensity X-rays to scan the breasts for abnormalities. Additionally, Magnetic Resonance Imaging (MRI) and ultrasound scans are employed as supplementary imaging techniques [3]. The techniques have, however, their setbacks; apart from the variations that may arise in the interpretation of a case, human error may result in misdiagnosis. In addition, analysis of medical images by humans may further result in incorrect diagnosis in 10% to 30% of cases.

The current rapid development of AI, especially by deep learning techniques, promises huge enhancements in the imaging of breast cancer. During the past decade, deep learning applications have shown tremendous performance regarding complex medical image analysis, diagnostic performances, and workflow simplification. These models span a wide range of imaging modalities and are being extended to probe into risk assessment, prognosis, and therapeutic response monitoring. Yet, despite such promise, several challenges with rigorous validation and model interpretability still need to be addressed. Nonetheless, integrating AI into medical imaging holds great potential in the hopes of upgrading the capability of detection, diagnostic errors, and overall outcomes in patients with breast cancer. The increasing prevalence of breast cancer has drawn significant attention from both medical and technological experts. While doctors focus on finding effective treatments, AI engineers are developing innovative early intervention techniques aimed at improving survival rates among women.

AI engineers now play a critical role in advancing the prediction and detection of breast cancer. Through deep learning and advanced feature extraction techniques, coupled with vast amounts of data, neural models have been developed to predict and detect breast cancer with exceptional precision and accuracy. Artificial intelligence has rapidly evolved, with researchers making significant contributions to societal welfare. Research shows that increased

public awareness, early detection, and effective community treatment can significantly prolong the survival of over 50% of breast cancer patients worldwide. Strategies for early detection include Breast Self-Examination (BSE), physical examination by healthcare professionals, and mammography [4].

In our study, we implemented a deep learning algorithm using Neural Networks to build classification models aimed at identifying cancerous tissues in the breast. This method enhances early detection capabilities. By employing a transfer learning technique for feature extraction, we were able to effectively detect and predict breast cancer. Transfer learning improves learning in new tasks by transferring knowledge from a related task. Transfer learning technique in machine learning where a model developed for a particular task is reused as the starting point for a model on a second task. For example, a pre-trained image recognition model can be adapted to analyze medical images. Machine learning algorithms focus on facilitating transfer learning, with current research covering inductive and reinforcement learning, negative transfer, and task mapping issues. Open problems remain in this area [5]. While feature extraction (FE) plays an important part in image retrieval, image processing, data mining, and computer vision. In medical imaging, this involves extracting meaningful patterns from images, like the shape or texture of tissues. It involves extracting relevant information from raw data, revealing unique features like contrast, homogeneity, entropy, mean, and energy. Despite challenges, FE techniques are versatile and can be applied to various applications [6]. Deep learning methods were used to classify breast tissues into categories such as normal, benign, and malignant. A dataset containing mammograms of breast cancer-related images was used to determine whether patients had breast cancer. Deep learning networks, composed of multiple processing layers, create various levels of abstraction to represent the data, achieving high accuracy and other statistical metrics. Keras, an advanced deep learning library, was utilized for prediction.

One approach involved building a neural network for breast cancer detection using a list of cellular properties from a breast mass biopsy. Breast cancer is the second most common cancer and the fifth leading cause of death among women, following lung cancer [7] [8]. Our method encompasses the development and validation of a sophisticated deep learning model capable of accurately detecting and predicting breast cancer from medical imaging data. The process involved exploring various deep learning algorithms, image preprocessing techniques, and feature extraction methods to enhance model performance. Comprehensive testing on diverse datasets was conducted to ensure the model's generalizability and reliability. Artificial intelligence holds the potential to revolutionize breast cancer diagnosis and prognosis, contributing to early detection and improved patient outcomes in medical science.

Research Questions:

RQ1: How can a deep learning framework be designed to effectively detect breast cancer using low-resolution mammogram images?

RQ1.1: How can data imbalance in breast cancer datasets (with fewer cancerous images) be addressed using machine learning techniques like the Neighbourhood Cleaning Rule (NCR)?

RQ1.2: What is the impact of JPEG compression and reduced pixel resolution on the performance of deep learning models in mammogram analysis?

Research Objectives:

The primary objective of this research is:

- To investigate the effectiveness of using low-resolution mammogram images for breast cancer detection and optimize model performance using deep learning approaches.
- To address data imbalance by applying the NCR algorithm to improve the quality of the training dataset.

- To assess the impact of using the Mini-DDSM dataset and evaluate the model's performance with JPEG-compressed mammogram images.
- To propose an optimized deep learning model that incorporates transfer learning with DenseNet121 for enhanced detection accuracy.

Novelty Statement:

This research presents a novel approach to breast cancer detection by leveraging low-resolution, JPEG-compressed mammogram images from the Mini-DDSM dataset. By utilizing the Neighbourhood Cleaning Rule (NCR) algorithm to address data imbalance and incorporating transfer learning with DenseNet121, the proposed deep learning model demonstrates the capability to effectively detect cancerous patterns despite challenges like low pixel resolution and reduced dataset size. This framework is validated using the Mini-DDSM dataset and provides significant potential for improving breast cancer diagnosis, especially in resource-constrained environments where high-resolution imaging and large dataset storage are limited.

Literature Review

The literature review emphasizes the critical importance of early detection of breast cancer due to its high prevalence and mortality rate. Given that one in eight women may develop breast cancer, this health issue demands serious attention. The literature analyzes the potential of deep learning for diagnosing breast cancer, discussing the limitations of screening methods, performance measures, and datasets, and providing new insights for future research [9] [10]. Machine learning and deep learning algorithms have been developed to distinguish between benign and malignant tumors, addressing the significant impact of breast cancer on women's mortality. Studies have shown that Support Vector Machines (SVM) and Random Forest classifiers offer the best prediction performance, while Convolutional Neural Networks (CNN) and Artificial Neural Networks (ANN) have achieved remarkable accuracy. Activation functions like ReLU and sigmoid have been employed to predict outcomes in terms of probabilities [10]. Traditional machine learning methods, such as Support Vector Machines (SVM), rely on manually defined features, while deep learning models like Convolutional Neural Networks (CNNs) automatically learn hierarchical features from data, offering greater flexibility and accuracy for complex tasks like image classification. Traditional machine learning automates the process of analytical model building and solving tasks, while deep learning, based on artificial neural networks, outperforms shallow models and traditional data analysis approaches. Understanding these fundamentals helps to develop a deep understanding of intelligent systems and their applications [11].

Deep learning, a powerful tool in artificial intelligence, uses feature learning to map input features to output. This process occurs in multiple connected layers, each containing multiple neurons, each a mathematical processing unit designed to learn the relationship between input features and output. Deep learning algorithms are complex mathematical structures with multiple processing layers that separate data features into abstraction layers. In supervised learning, a Deep Neural Network (DNN) sequentially passes input feature data from neurons in one layer to neurons in the next layer during repeated cycles, known as epochs. Each neuron accepts weighted inputs from multiple other neurons, summating them and passing them to an internal activation function. If the activation threshold is exceeded, the neuron generates an output combined with a weight value before passing it to multiple neurons in the next layer. The model's knowledge is captured in its weight values, which are analogs to traditional statistical models. DenseNet121, a deep learning model, enhances information flow across layers through dense connections, improving performance in medical image analysis. Compared to traditional CNNs, DenseNet reduces the risk of vanishing gradients and requires fewer parameters, making it ideal for medical applications like breast cancer detection [17].

In studies using the Diagnostic Breast Cancer Wisconsin dataset, researchers have applied machine learning (ML) methods such as decision trees, SVM, naive Bayes, and K-nearest neighbors to predict breast cancer, with SVM emerging as the most accurate classifier [12]. Various cell analysis frameworks reportedly allow for the integration of cell-level data to examine multiscale diseased images [13]. Deep learning algorithms have proven effective in handling the complexities of automatic BC diagnosis. While numerous review studies have addressed BC classification, few have provided clear guidance for future researchers. Most review studies on BC have focused primarily on general artificial neural networks (ANNs) or conventional ML techniques [14].

Although numerous methods such as Logistic Regression, Support Vector Machine (SVM) and K Nearest Neighbour (KNN), Multi-Layer perceptron classifier, Artificial Neural Network (ANN)) etc. have been proposed by researchers, further improvements are still needed [10]. Researchers continue to develop deep learning methods for early breast cancer detection through imaging, though reviews of new architectures and modalities remain scarce. The advantages and disadvantages of existing deep learning models have been examined, alongside discussions of various imaging techniques, measurements, results, challenges, and future research directions [15]. A particular study aims to improve early breast cancer detection and diagnosis by utilizing a proposed deep learning model (CNN) and five pre-trained models. The research classifies BC into eight categories based on MRI images. The models were trained and evaluated using a dataset collected from Kaggle, which was enhanced through Generative Adversarial Networks (GAN) techniques. With evaluations based on F1-score, recall, precision, and accuracy, the BCCNN achieved the highest accuracy, outperforming other models. Notably, dataset boosting and magnification significantly improved the proposed model's performance, especially with 400X magnification images [16].

To forecast the clinical prognosis of breast cancer, researchers have applied deep learning, feature selection, and extraction methods. Their research suggests that using these techniques can enhance the precision of breast cancer outcome prediction [17]. One paper investigates eight classification models, including single and ensemble classifiers, using a dataset refined by five feature selection methods. SVM, MLP, and stack classifiers achieved high accuracy, with SVM outperforming others. The main contribution of this study is the ranking of SVM as the best classifier, with a comparative analysis categorizing classifiers into performance tiers. The enhanced dataset and methodologies significantly improve accuracy and computational efficiency, highlighting SVM's superiority even over the ensemble stack classifier [18].

A study focusing on BC detection using a computer-aided diagnosis (CAD) system based on CNNs leverages advancements in deep learning to enhance cancer cell identification, achieving a high accuracy rate in classifying benign and malignant cells [19]. Another study discusses the application of transfer learning techniques to train CNNs for automated diagnosis using ultrasound images, demonstrating that models like MobileNet and DenseNet121 show promise in detecting breast cancer [20]. The study aims to develop a robust breast cancer classification model using meta-learning and multiple CNNs on the Breast Ultrasound Images (BUSI) dataset. The proposed model employs meta-learning, transfer learning, and data augmentation to optimize learning, enhance feature extraction, and increase dataset diversity. By combining CNN outputs through meta-ensemble learning, the model achieves high accuracy [21]. Deep learning techniques for breast cancer diagnosis using various medical image modalities, including X-ray (Mammography), ultrasound, and histopathology images, have been investigated. Different strategies, including VGGNet19, ResNet50, DenseNet121, and EfficientNet v2 for image classification, and UNet, ResUNet++, and DeepLab v3 for image segmentation, have been evaluated. Results indicate that ResNet50 performs best in image classification, while UNet excels in X-ray and

ultrasound image segmentation. The proposed strategies significantly enhance accuracy, with improvements of 33.3% in X-ray segmentation, 29.9% in ultrasound segmentation, and 22.8% in histopathology image classification [22].

Researcher Yuan suggested a CNN-based model for identifying various cancers using information retrieved from many levels to analyze the spatial distribution of lymphocytes in Whole Slide Images (WSIs) [23]. Machine learning automates analytical model building and solving tasks, while deep learning, based on artificial neural networks, outperforms shallow models and traditional data analysis approaches. These concepts provide a broader understanding of the systematic underpinning of intelligent systems, addressing challenges in electronic markets and networked business, human-machine interaction, and artificial intelligence servitization.icientNet v2, UNet, ResUNet++, and DeepLab v3, along with loss functions like binary cross-entropy, dice loss, and Tversky loss, as well as data augmentation techniques. Results reveal that ResNet50 excels in image classification, while UNet performs optimally in image segmentation for X-ray and ultrasound images [19].

One study employed a web crawler to download age variables and thumbnail images from the Digital Database for Screening Mammography to build an AI-based model for age estimation from mammography images. Using a Random Forest regressor, the model estimates age automatically with an average error value of 8 years. The method introduces the free-access Mini-DDSM dataset, which has been validated using logistic and linear regression models on another independent dataset [24]. Another study uses mammography images for breast cancer identification, applying deep learning techniques like CNN, VGG19, Inception-Net, and ResNet50 for image classification [25]. Additionally, a study conducted for breast cancer detection using Computer-Aided Diagnosis (CAD) utilized technologies like CNNs, Generative Adversarial Networks (GANs), and reinforcement learning, resulting in an accuracy rate of 99.58% [26]. Finally, a study on deep learning-based breast cancer diagnosis trained various CNN models, including MobileNet and DenseNet121, through transfer learning techniques. It found that DenseNet121 was the most effective in detecting breast cancer [27].

The literature underscores the complexity of applying deep learning techniques to breast cancer detection and emphasizes the need for continuous exploration to improve diagnostic accuracy and patient outcomes.

Material and Methods:

Data Collection:

A dataset is organized, transformed, and modeled during this process by a data analyst or data scientist. The dataset from Kaggle, Complete MINI-DDSM, last updated on 2021-03-23, has been used in this study as shown in Figure 1 and Figure 2. Figure 1 age distribution in the dataset While figure 2 explains the density (amount of Fibroglandular tissue) distribution in Complete MINI-DDSM using Bi-Rads scoring [28]. Its format is in jpeg for usage at smaller levels thus pixel resolution is low. This is the lightweight version of the Digital Database for Screening Mammography (DDSM) which originated in the USA and the year of building is 1999. With 26, 20 cases and 10,480 photos, DDSM is the largest public database for screening mammography as shown in Table 1. It has MLO and CC images of various lesions, from benign to malignant [29]. CBIS-DDSM, an updated and condensed version of the DDSM dataset for the evaluation of CAD procedures in mammography, is also available to use [9]. It is one of the most famous and profoundly used datasets to date. This dataset has 3 types of images for classification (Normal, Benign, and malignant).

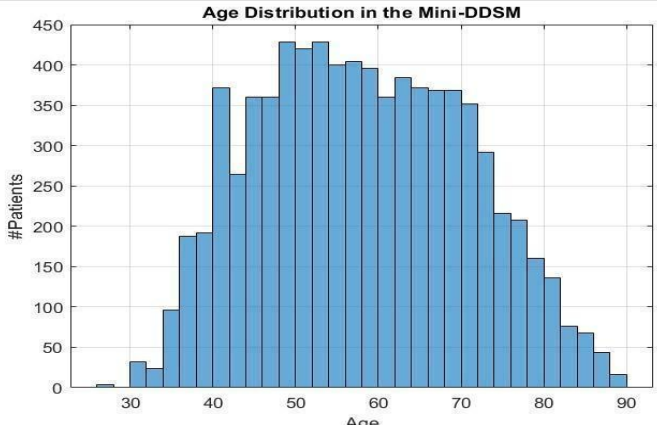

Figure 1. Age Distribution

Using feature extraction, we developed a model for detecting and predicting tumors with a focus on accurately distinguishing between benign and malignant types. In our study, we segmented the tumor samples into five categories: benign, malignant, and atypical. To enhance the model's accuracy in distinguishing between these different tumor types, we extracted key features from the cell nucleus, including area, perimeter, eccentricity, compactness, and circularity. These features were crucial in improving the model's performance, allowing for more precise tumor classification.

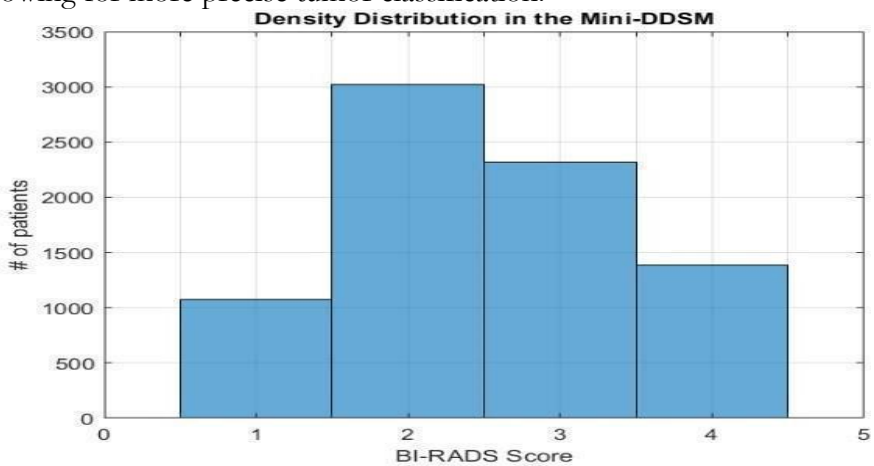

Figure 2. Density Distribution

Dataset Characteristics: The dataset used in this study, consisting of breast mammography images, was sourced from Kaggle. The images were saved in JPEG format, which is suitable for storage and distribution but results in lower pixel resolution due to compression. This aspect of the dataset was taken into account throughout the research and modeling process.

Training and Testing Dataset:

Table 1 provides the metadata for the MINI-DDSM dataset, which serves as the foundation for our analysis. Various statistical metrics—including mean, median, mode, standard deviation, range, skewness, and kurtosis—were computed for this dataset, as detailed in Table 2. During feature extraction, features that did not show significant changes in their average values when comparing benign to malignant lesions were excluded from further analysis. Among the features examined, area, perimeter, and circularity demonstrated significantly different average values between benign and malignant cases. Consequently, these three features were prioritized for further research, as they play a crucial role in distinguishing benign tumors, tissues, or cells from malignant ones.

Table 1. Metadata for the mini-ddsm dataset

Metadata	Value

Metadata	Value
Image Type	JPEG
Size	4 GB
Mode of Image Acquisition	Screen Film
Number of Images	10,758
Resolution	8 bits
Originated From	DDSM

Table 2. Statistics for training and testing data

Statistic	Training Data	Testing Data
Images (%)	79.9962818368	20.0037181632
Mean	0.15	0.15
Median	0.04	0.04
Mode	0.00	0.00
Variance	0.05	0.05
Standard Deviation	0.21	0.21

Data Analysis:

Statistical Analysis of Training and Testing Data:

The close alignment of mean and median values suggests a normalized pixel distribution, while the mode value of 0.00 indicates the presence of many low pixel values, which is typical in image data. This is particularly expected in mammograms, where normalization often results in lower pixel values. The low variance and standard deviation observed in both training and testing data further indicate a relatively uniform distribution, with pixel values clustered around the mean.

- Mean: Both datasets have a mean of 0.15, indicating that the average value of the data points is consistent across training and testing sets. This consistency suggests that the model will likely perform similarly on both datasets.
- Median: The median is 0.04 for both datasets, which represents the middle value in the data distribution. This value is less sensitive to outliers than the mean and confirms that the central tendency of the data is stable across both datasets.
- Mode: The mode is 0.00 for both datasets, showing the most frequently occurring value. A common mode of 0.00 suggests that this value appears often, which may indicate a prevalence in the dataset and could influence model behavior.
- Variance: With a variance of 0.05 for both datasets, the data points are closely distributed around the mean. Low variance indicates that the data is relatively stable and uniform, which helps in achieving a stable model training process.
- Standard Deviation: The standard deviation is 0.21 for both datasets, reflecting the average distance of data points from the mean. This moderate spread suggests that while there is some variability in the data, it is consistent between the training and testing datasets, supporting reliable model evaluation.

These metrics help in understanding the data's distribution and stability, which is crucial for assessing the effectiveness of data preprocessing and the model's ability to generalize to new data.

Preprocessing techniques:

Data Preprocessing:

Before feature extraction from mammograms, data analysis and preprocessing were performed to prepare the raw data for machine learning or deep learning algorithms.

Preprocessing involves various transformations aimed at converting the input data into meaningful floating-point tensors suitable for feature extraction. During this stage, data normalization was applied to the batch. The data was then split using the Scikit-learn (sklearn) library. Additionally, a technique known as "Neighborhood Cleaning" was employed, which is a data preprocessing method used to enhance dataset quality by removing noisy or irrelevant samples. This technique analyzes the relationships between samples to identify and remove outliers or mislabeled instances, thereby improving model training by reducing irrelevant noise. Furthermore, normalization, specifically Min-Max scaling, was utilized to bring data values into a common range, minimizing the influence of features with larger magnitudes. Code snippets for data preprocessing are provided in the implementation section of the appendix.

1) Mathematical Formulation: The Neighborhood Cleaning Rule (NCR) algorithm improves data quality in datasets with samples represented by x_i for analysis. Calculate the distance, identify neighbours within a defined threshold distance, compute a neighborhood metric based on its neighbours, and finally, if the computed neighbourhood metric $M(x_i)$ meets a predefined criterion, remove xi and its neighbourhood from the dataset.

$$d(x_i, x_j) = some - distance - metric(x_i, x_j)\ N(x_i) = x_j \mid d(x_i, x_j) threshold$$

$$M(x_i) = some - function\ (N(x_i))$$

Feature Extraction:

In Figure 3, the model system is shown through the class structure view to understand the steps taken. Mammogram images possess a complex structure that presents significant challenges for radiologists in feature extraction and accurate disease classification. To address these challenges, we employed the DenseNet121 model through transfer learning to extract features from our preprocessed dataset. Transfer learning leverages a previously trained model as the foundation for a new task, optimizing performance by applying knowledge from one domain to another. This approach not only enhances model efficiency but also enables effective training with a limited amount of data, leading to time savings and successful outcomes [10] [11]. DenseNet121, a popular pre-trained deep learning model, serves as the core architecture for our feature extraction. Compared to other pre-trained models like VGG or ResNet, DenseNet121 offers several advantages for image classification tasks, making it particularly suitable for our study.

1)DenseNet121: DenseNet121 is a Deep Convolutional Neural Network (CNN) architecture that leverages dense connections to enhance information flow across layers [15]. DenseNet121 offers superior parameter efficiency, feature reuse, gradient flow, and accuracy compared to other models. Its design enables the efficient learning of relevant features in medical images, which is particularly advantageous for breast cancer detection and prediction projects.

The DenseNet architecture utilizes L layers to address the vanishing gradient problem, reduce the number of parameters, and promote feature reuse. In DenseNet, L layers refer to multiple densely connected layers where each layer receives inputs from all preceding layers, enhancing feature reuse, reducing parameters, and mitigating the vanishing gradient problem. Dense connections within the network minimize the parameter count and ensure effective feature reuse while mitigating the risk of vanishing gradients. The network is structured into dense blocks, where the feature map dimensions remain constant within each block, but the number of filters varies [14].

Proposed model:

Training and Testing System:

In this stage, feature extraction is combined with the CNN model, leveraging DenseNet121 as the pre-trained base model. The architecture incorporates a feature extraction model, dense layers, a global pooling layer, and a dropout layer, as illustrated in Figure 4. To minimize model overfitting, we applied the train-validate-test procedure, commonly known as

early stopping. Overfitting is a significant issue in supervised machine learning, hindering the accurate fit of models. It occurs due to noise, limited data sets, and complex classifiers. To reduce overfitting, various methods are proposed, including early stopping, network reduction, data expansion, and regularization. These techniques help prevent overfitting, manage issues, and ensure model execution [30]. Additionally, the Neighborhood Cleaning Rule (NCR) was employed as a data preprocessing technique to remove noisy or irrelevant samples on training and testing dataset before feature extraction. This approach improves data quality, reduces overfitting, and enhances training efficiency by selecting the most relevant features for breast cancer detection and prediction.

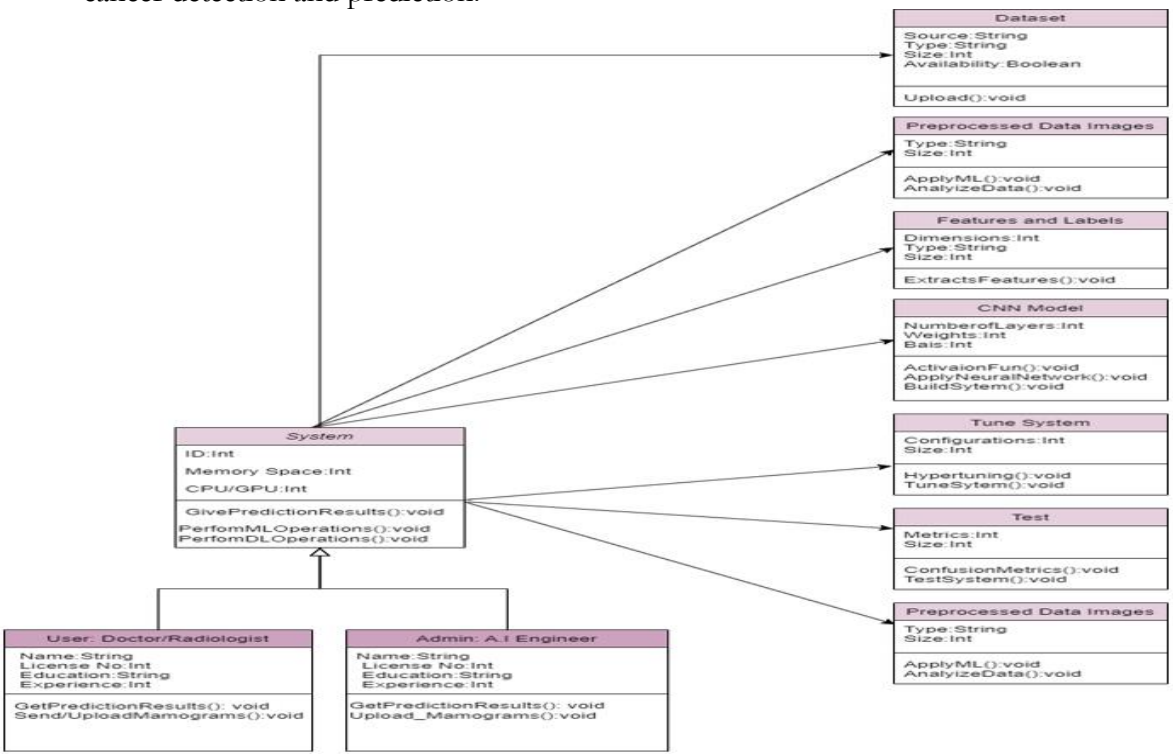

Figure 3. Model system shown through the class structure view.

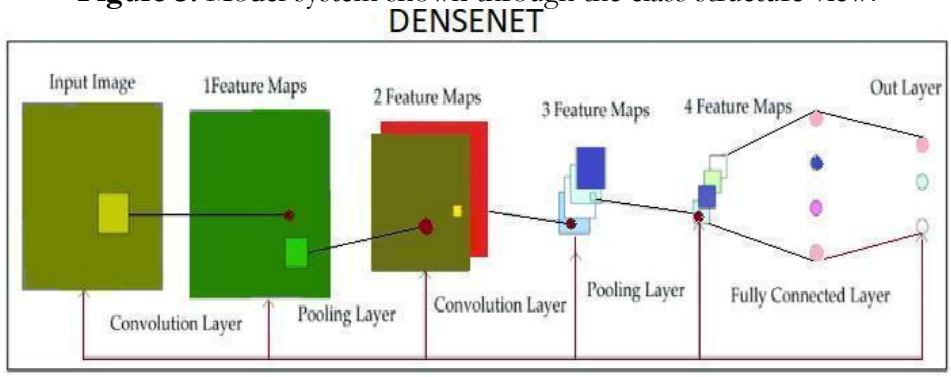

Figure 4. DenseNet Structure

Tuning System:

Hyperparameter tuning optimizes learning algorithms by finding optimal hyperparameter values for data sets, maximizing performance, and minimizing loss functions, resulting in better results with fewer errors. In this stage, tuning of the system using suitable parameters after experimenting with different values has been done. Adam optimization is a stochastic gradient descent method using adaptive estimation of first and second-order moments such as set learning rate = 0.001, epsilon=1e-07, beta-1=0.9, beta-2=0.999. These

values were carefully selected after experimenting with different parameter to find the best possible combination.

Interference

The inference phase in detecting and predicting breast cancer involves preprocessing breast mammography pictures using methods like normalization and the Neighbourhood Cleaning Rule (NCR). The modified data is then analyzed using model layers like dropout algorithms, global pooling, and dense neural networks. These layers improve the retrieved characteristics and enable accurate predictions, utilizing collective knowledge gained during training. The detection and categorization of normal, benign, or malignant issues in mammography images are the result of complex computations and transformations. The inference structure synthesizes information from data analysis, preprocessing, and model creation phases, resulting in precise predictions for breast cancer detection and prognosis. The Architecture of Breast Cancer Detection and Prediction using Deep Learning is shown in Figure 5. The implementation part is available in the appendix link with processed images. Below is the Model Structure description from Figure 5.

Model Structure:

- The model section shows the core architecture used for detecting and predicting breast cancer:

1. **DenseNet121 (Functional):** A pre-trained deep learning model (DenseNet121) is used as a feature extractor. DenseNet121 is known for its dense connectivity between layers, which improves information flow and makes the model efficient and robust.
2. **Global Average Pooling (2D):** Reduces the spatial dimensions of the feature maps output by DenseNet121 to a fixed size, making the model more computationally efficient.

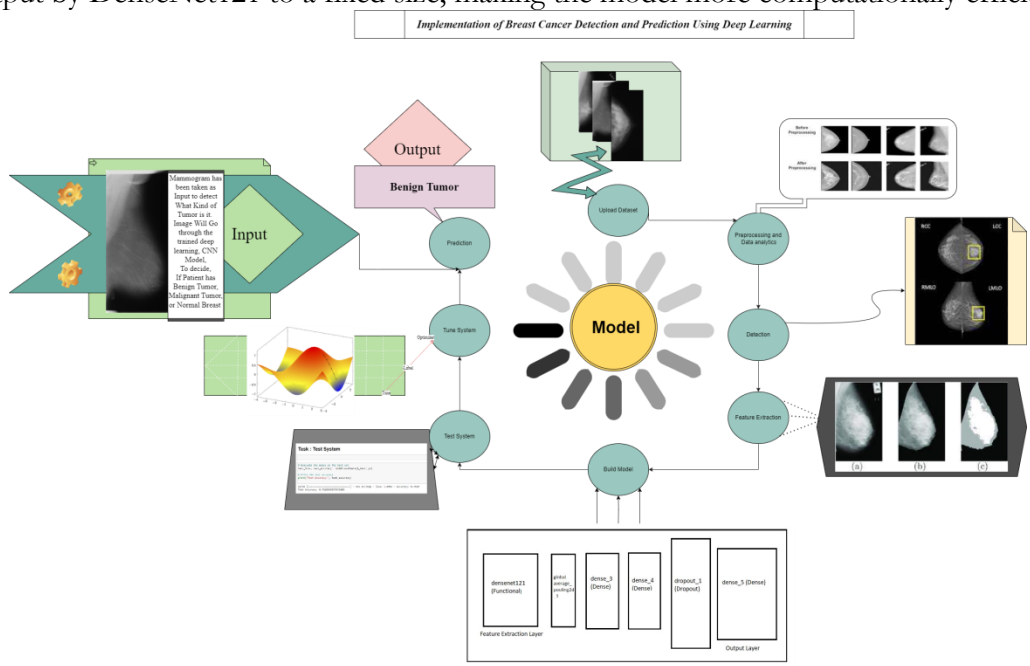

Figure 5. The Architecture of Breast Cancer Detection and Prediction Using Deep Learning

3. **Dense Layers (Dense_3, Dense_4, Dense_5):** These fully connected layers are used to learn complex patterns from the features extracted. They are responsible for making the final classification decision.
4. **Dropout Layer (Dropout_1):** This layer helps prevent overfitting by randomly dropping a fraction of neurons during training, thereby improving the model's generalization ability.
5. **Output Layer:** The final output layer processes the results from the preceding layers to provide the model's prediction.

System Tuning: The system tuning phase involves fine-tuning the model's hyperparameters, such as learning rate, batch size, and the number of epochs, to optimize overall performance.

System Compilation and Testing: Once tuning was completed, the model was compiled with a chosen loss function and optimization algorithm. It was then evaluated using a separate test set to assess its accuracy and performance.

Prediction: Finally, the model made predictions on the input mammogram images, categorizing them as "Normal," "Benign," or "Malignant." This output aided in determining the nature of the detected breast tissue and guides further medical diagnosis or treatment.

The entire architecture follows a pipeline that begins with data input, moves through preprocessing, feature extraction, model training, tuning, and testing, and culminates in a predictive output. The primary goal of this architecture is to automate the detection and classification of breast cancer in mammogram images using a deep learning model, thus supporting radiologists and clinicians in early and accurate diagnosis.

Results:

A comprehensive evaluation of the developed breast cancer detection and prediction model was conducted to assess its effectiveness. The model was tested on an independent set of images after initial validation on the provided dataset. This evaluation included a detailed analysis of the model's loss and accuracy metrics. The model's loss was calculated using the sparse categorical cross-entropy loss function, which measures the difference between predicted and true labels. Accuracy was computed as the ratio of correct predictions to the total predictions made during training and validation. In our study, we employed transfer learning with DenseNet121, a model proven effective for medical image classification, as validated by existing literature. The model was trained on the Mini-DDSM dataset, using data augmentation and the Neighbourhood Cleaning Rule (NCR) to address class imbalance and improve generalization. DenseNet121's ability to capture intricate features in medical images was critical to this success. Preprocessing steps like normalization helped stabilize the model and improve training convergence. Neighbourhood Cleaning Rule (NCR) was employed, which removed noisy samples and enhanced the model's generalization. For further, optimized performance, we applied early stopping, which halted training by call back function when validation loss stopped improving. These implementations effectively prevented overfitting and ensured a robust and reliable model. Despite the challenges of working with JPEG-compressed images we have achieved approximately 76.28% accuracy.

Model Assessment: The assessment revealed that the test accuracy, measured on a separate dataset, was approximately 76.28%, indicating the model's ability to generalize to new samples. The test accuracy of 76.28% was measured for a deep learning model trained using Keras with the Adam optimizer. The dataset used for evaluation was a test dataset (X_Test, y2), separate from the training data (X_Train, y1), the code snippet shared in Appendix google drive link. During the evaluation, the characteristics of the data were carefully considered to provide deeper insights into the model's performance. Figure 6 reinforces the critical insights gained from monitoring training dynamics, underlining the need for careful epoch selection to optimize model performance while avoiding overfitting using function early stopping.

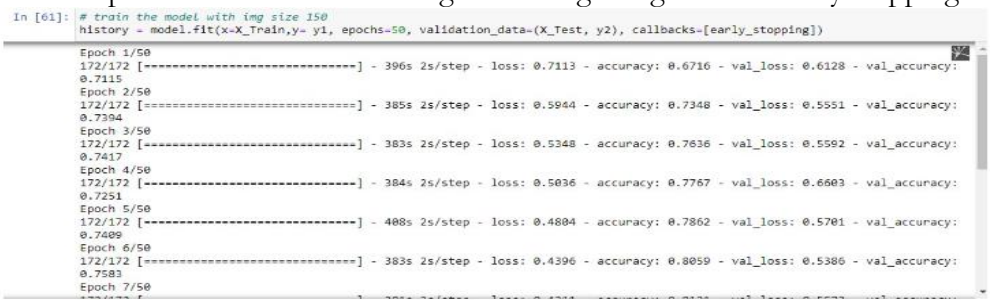

Figure 6. Early Epochs: Minimum Validation Loss and Overfitting

Pixel Statistics: Statistical analysis of the dataset's pixel values showed that the mean and median values were close, indicating a normalized distribution. The presence of 0.00 mode values aligns with expectations for image data. Additionally, the 0.05 variance and 0.21standard deviation suggest that pixel values were tightly clustered around the mean, contributing to a relatively uniform distribution. Figure 7 presents a comparison of training and validation loss over multiple epochs during the training of the model. The consistent decrease in training loss as the number of epochs increases indicates that the model is effectively learning from the training data and improving its performance by reducing error over time. The average training accuracy over the epochs is 81.0638%. A smooth and continuous decline in training loss is generally a positive sign, suggesting that the model is optimizing well.

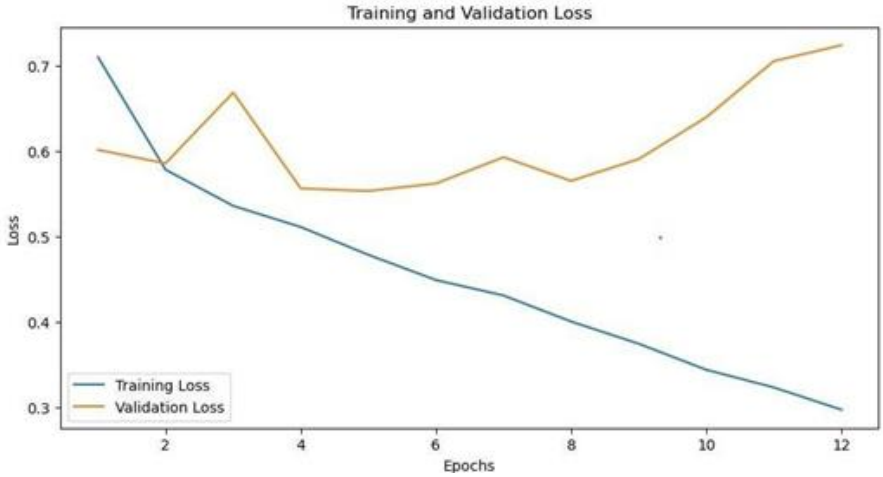

Figure 7. Comparison of Loss and epoch training dataset

The validation loss decreases by 13.3812% initially but starts to increase after a few epochs, which is a clear indication of overfitting. At this stage, the model begins to memorize the training data rather than generalize from it, leading to poorer performance on the validation set, which represents unseen data. The minimum validation loss is observed between epochs 2 and 4, marking the point where the model performed best on the validation set before overfitting set in. To mitigate overfitting, techniques such as early stopping can be applied. Early stopping involves halting the training process once the validation loss ceases to decrease, preventing the model from continuing to learn patterns specific to the training data that do not generalize well to new data. This approach underscores the importance of monitoring both training and validation losses to identify the optimal number of training epochs and avoid overfitting.

Figure 8 illustrates the training accuracy steadily increasing as the number of epochs progresses. This trend indicates that the model is becoming increasingly proficient at predicting the training data correctly, reflecting an improvement in its learning process over time. A consistent rise in training accuracy is a positive indicator, showing that the model is effectively enhancing its performance on the training set. The point at which validation accuracy peaks represents the optimal number of epochs to avoid overfitting. Beyond this point, the model's performance on new, unseen data begins to deteriorate, highlighting the importance of balancing training duration with the need to maintain generalization to new data.

Discussion:

In this paper we proposed a deep learning framework to detect breast cancer using low resolution, JPEG compressed mammogram images from the Mini-DDSM dataset. We used DenseNet121, a powerful transfer learning model and Neighbourhood Cleaning Rule (NCR) to handle imbalanced data to tackle the challenges of low quality images and imbalanced dataset. The model achieved 76.28% accuracy which shows its effectiveness under these

constraints. Our paper focuses on low resolution mammograms which are similar to [20] where MobileNet and DenseNet121 were used for breast cancer detection on ultrasound images. Although ultrasound data has better resolution, our model's performance on JPEG compressed mammograms is comparable which shows DenseNet121 can extract features well across different imaging modalities and clarify RQ1 about the framework's effectiveness in low resolution. For RQ1.1, which is about handling imbalanced data, our use of NCR is similar to [21] where data augmentation and ensemble learning was used to increase dataset diversity. However, unlike ensemble methods that combine the output of multiple CNNs, our method directly focuses on improving individual samples by removing noisy data and making the model to generalize from smaller dataset. This is more critical when there are fewer cancerous images as mentioned in [25] that highlighted the challenge of imbalanced breast cancer dataset and used GANs to artificially balance the data. Our use of NCR is a simpler yet effective alternative to more computationally expensive methods like GAN based augmentation.

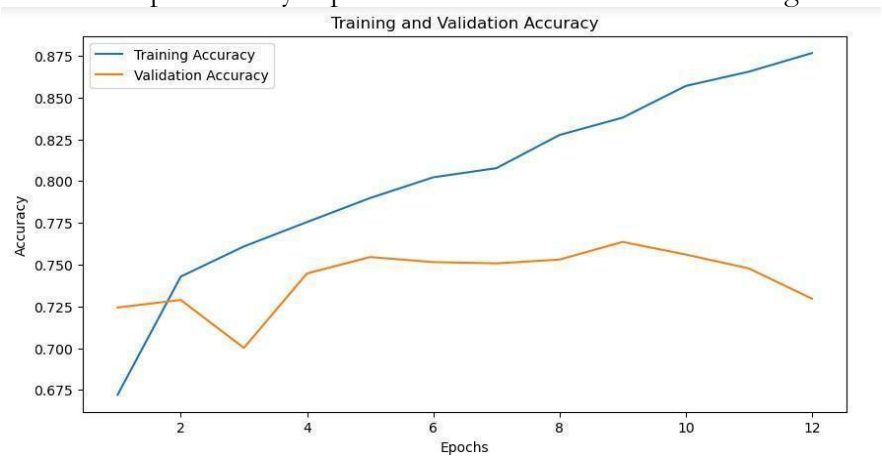

Figure 8. Comparison of accuracy and epoch training dataset

For RQ1.2, about the effect of JPEG compression and reduced pixel resolution on the model performance, our paper is the first to use low resolution images directly from the Mini-DDSM dataset. [21] used high resolution mammograms where ResNet50 achieved slightly higher accuracy but did not address the issue of low quality and small images. Our results show that even with heavy compression, DenseNet121 is still good and supports the idea that deep learning models can detect cancer patterns in resource constrained environment where high quality imaging is not possible. Besides [24] used random forest for age estimation from DDSM dataset, our CNN based approach directly targets cancer detection. Their simpler machine learning model achieved modest accuracy; our deep learning framework is a more advanced solution for breast cancer classification especially in noisy and compressed data. In summary, this paper adds to the existing research by including data imbalance, low resolution images and JPEG compression in breast cancer detection and shows that DenseNet121 with NCR and early stopping can handle these challenges to achieve good performance.

Conclusion:

In this study, we developed a breast cancer detection model using low-resolution, JPEG compressed mammogram images from the Mini-DDSM dataset. By using DenseNet121 for feature extraction and normalization and Neighbourhood Cleaning Rule (NCR) for data imbalance, we were able to build a model that achieved an accuracy of 76.28%. This shows the model can detect cancer patterns despite the challenges of image compression and low pixel resolution. The careful preprocessing, transfer learning, and early stopping in training helped in preventing overfitting so the model is reliable for breast cancer detection.

However, the model is promising there is still room for improvement. Future work can focus on using higher resolution images or other imaging modalities to overcome the

limitation of JPEG compression. Expanding the dataset to include a more diverse patient population and multi-modal data integration like genetic or patient history can increase the model's generalizability and accuracy. Refining DenseNet121 or exploring other architectures and incorporating explainable AI can further improve model transparency and performance and help in clinical adoption for early and accurate breast cancer detection.

Author's Contribution: All Authors contributed towards idea refinement, Author 1 contributed paper write-up, literature, project implementation, and reference management. Authors 2 and 3 reviewed the whole work.

Conflict of Interest: There exists no conflict of interest for publishing this manuscript in IJIST.

Appendix

Google drive link:

https://drive.google.com/file/d/1ZHEYu_Bq47atcpkjrX3dIb4LSdegNkme/view?usp=sharing

https://drive.google.com/drive/folders/1j2GEW_nwvNAI4Wx4CouMgZ24YGMqFlsG?usp=sharing

References:

[1] "Benign vs Malignant Tumors: What's the Difference?" Accessed: Sep. 29, 2024. [Online]. Available: https://www.cancercenter.com/community/blog/2023/01/whats-the-difference-benign-vs-malignant-tumors

[2] "Malignant vs. Benign Tumors: How They Differ." Accessed: Sep. 29, 2024. [Online]. Available: https://www.verywellhealth.com/what-does-malignant-and-benign-mean-514240

[3] Yadavendra and S. Chand, "A comparative study of breast cancer tumor classification by classical machine learning methods and deep learning method," Mach. Vis. Appl., vol. 31, no. 6, pp. 1–10, Sep. 2020, doi: 10.1007/S00138-020-01094-1/METRICS.

[4] "The control of breast cancer. A World Health Organization perspective - PubMed." Accessed: Sep. 29, 2024. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/2187590/

[5] "Handbook Of Research On Machine Learning Applications and Trends: Algorithms, Methods and Techniques - 2 Volumes | Guide books | ACM Digital Library." Accessed: Sep. 29, 2024. [Online]. Available: https://dl.acm.org/doi/10.5555/1803899

[6] A. O. Salau and S. Jain, "Feature Extraction: A Survey of the Types, Techniques, Applications," 2019 Int. Conf. Signal Process. Commun. ICSC 2019, pp. 158–164, Mar. 2019, doi: 10.1109/ICSC45622.2019.8938371.

[7] Z. Q. Zhao, P. Zheng, S. T. Xu, and X. Wu, "Object Detection with Deep Learning: A Review," IEEE Trans. Neural Networks Learn. Syst., vol. 30, no. 11, pp. 3212–3232, Nov. 2019, doi: 10.1109/TNNLS.2018.2876865.

[8] G. N. Sharma, R. Dave, J. Sanadya, P. Sharma, and K. K. Sharma, "Various types and management of breast cancer: An overview," J. Adv. Pharm. Technol. Res., vol. 1, no. 2, pp. 109–126, 2010, doi: 10.4103/2231-4040.72251.

[9] K. Rautela, D. Kumar, and V. Kumar, "A Systematic Review on Breast Cancer Detection Using Deep Learning Techniques," Arch. Comput. Methods Eng. 2022 297, vol. 29, no. 7, pp. 4599–4629, Apr. 2022, doi: 10.1007/S11831-022-09744-5.

[10] M. Tiwari, R. Bharuka, P. Shah, and R. Lokare, "Breast Cancer Prediction Using Deep

Learning and Machine Learning Techniques," SSRN Electron. J., Mar. 2020, doi: 10.2139/SSRN.3558786.

[11] C. Janiesch, P. Zschech, and K. Heinrich, "Machine learning and deep learning," Electron. Mark., vol. 31, no. 3, pp. 685–695, Apr. 2021, doi: 10.1007/s12525-021-00475-2.

[12] H. Asri, H. Mousannif, H. Al Moatassime, and T. Noel, "Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis," Procedia Comput. Sci., vol. 83, pp. 1064–1069, Jan. 2016, doi: 10.1016/J.PROCS.2016.04.224.

[13] S. Zhang and D. Metaxas, "Large-Scale medical image analytics: Recent methodologies, applications and Future directions," Med. Image Anal., vol. 33, pp. 98–101, Oct. 2016, doi: 10.1016/J.MEDIA.2016.06.010.

[14] R. Krithiga and P. Geetha, "Breast Cancer Detection, Segmentation and Classification on Histopathology Images Analysis: A Systematic Review," Arch. Comput. Methods Eng., vol. 28, no. 4, pp. 2607–2619, Jun. 2021, doi: 10.1007/S11831-020-09470-W/METRICS.

[15] M. F. Mridha et al., "A Comprehensive Survey on Deep-Learning-Based Breast Cancer Diagnosis," Cancers (Basel)., vol. 13, no. 23, Dec. 2021, doi: 10.3390/CANCERS13236116.

[16] B. S. Abunasser, M. R. J. AL-Hiealy, I. S. Zaqout, and S. S. Abu-Naser, "Convolution Neural Network for Breast Cancer Detection and Classification Using Deep Learning," Asian Pacific J. Cancer Prev., vol. 24, no. 2, pp. 531–544, Feb. 2023, doi: 10.31557/APJCP.2023.24.2.531.

[17] S. Ara, A. Das, and A. Dey, "Malignant and Benign Breast Cancer Classification using Machine Learning Algorithms," 2021 Int. Conf. Artif. Intell. ICAI 2021, pp. 97–101, Apr. 2021, doi: 10.1109/ICAI52203.2021.9445249.

[18] M. A. Elsadig, A. Altigani, and H. T. Elshoush, "Breast cancer detection using machine learning approaches: a comparative study," Int. J. Electr. Comput. Eng., vol. 13, no. 1, pp. 736–745, Feb. 2023, doi: 10.11591/ijece.v13i1.pp736-745.

[19] G. Hamed, M. A. E. R. Marey, S. E. S. Amin, and M. F. Tolba, "Deep Learning in Breast Cancer Detection and Classification," Adv. Intell. Syst. Comput., vol. 1153 AISC, pp. 322–333, 2020, doi: 10.1007/978-3-030-44289-7_30.

[20] L. Zhang, R. Xu, and J. Zhao, "Learning technology for detection and grading of cancer tissue using tumour ultrasound images1," J. Xray. Sci. Technol., vol. 32, no. 1, pp. 157–171, Feb. 2024, doi: 10.3233/XST-230085.

[21] M. D. Ali et al., "Breast Cancer Classification through Meta-Learning Ensemble Technique Using Convolution Neural Networks," Diagnostics 2023, Vol. 13, Page 2242, vol. 13, no. 13, p. 2242, Jun. 2023, doi: 10.3390/DIAGNOSTICS13132242.

[22] D. Kwak, J. Choi, and S. Lee, "Rethinking Breast Cancer Diagnosis through Deep Learning Based Image Recognition," Sensors 2023, Vol. 23, Page 2307, vol. 23, no. 4, p. 2307, Feb. 2023, doi: 10.3390/S23042307.

[23] Y. Yuan, "Modelling the spatial heterogeneity and molecular correlates of lymphocytic infiltration in triple-negative breast cancer," J. R. Soc. Interface, vol. 12, no. 103, Feb. 2015, doi: 10.1098/RSIF.2014.1153.

[24] C. D. Lekamlage, F. Afzal, E. Westerberg, and A. Chaddad, "Mini-DDSM: Mammography-based Automatic Age Estimation," ACM Int. Conf. Proceeding Ser., pp. 1–6, Nov. 2020, doi: 10.1145/3441369.3441370.

[25] S. Urabinahatti and D. Jayadevappa, "Breast Cancer Detection Using Deep Learning Technique," 2nd IEEE Int. Conf. Distrib. Comput. Electr. Circuits Electron. ICDCECE 2023, 2023, doi: 10.1109/ICDCECE57866.2023.10150859.

[26] K. Loizidou, R. Elia, and C. Pitris, "Computer-aided breast cancer detection and classification in mammography: A comprehensive review," Comput. Biol. Med., vol. 153, p. 106554, Feb. 2023, doi: 10.1016/J.COMPBIOMED.2023.106554.

[27] K. Clark et al., "The cancer imaging archive (TCIA): Maintaining and operating a public information repository," J. Digit. Imaging, vol. 26, no. 6, pp. 1045–1057, Dec. 2013, doi: 10.1007/S10278-013-9622-7/METRICS.

[28] "The Complete Mini-DDSM." Accessed: Sep. 29, 2024. [Online]. Available: https://www.kaggle.com/datasets/cheddad/miniddsm2

[29] N. M. ud din, R. A. Dar, M. Rasool, and A. Assad, "Breast cancer detection using deep learning: Datasets, methods, and challenges ahead," Comput. Biol. Med., vol. 149, p. 106073, Oct. 2022, doi: 10.1016/J.COMPBIOMED.2022.106073.

[30] "(PDF) Review on over-fitting and under-fitting problems in Machine Learning and solutions." Accessed: Sep. 29, 2024. [Online]. Available: https://www.researchgate.net/publication/344882855_Review_on_over-fitting_and_under-fitting_problems_in_Machine_Learning_and_solutions