STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **13**, January 2025, pp 434–449.
Published online in International Academic Press (www.IAPress.org)

Enhancing Mammography Models: The Impact of Radiologist Recommendations on Algorithmic Precision

Youssef Lahdoudi*, Abdelghani Ghazdali, Hamza Khalfi, Nidal Lamghari

MPA, ENSA Khouribga, Sultan Moulay Slimane University, BP 77,Béni Amir, Morocco

Abstract This study highlights the benefits of advanced image classification in breast cancer diagnosis and treatment. We utilize deep learning algorithms like YOLOv5 for image segmentation and Densenet121 for feature extraction from segmented regions. Our dataset includes 54,706 mammography images for comprehensive analysis. We evaluate 100 challenging cases, ensuring a balanced representation of benign and malignant instances. Validation involves 50 consensus cases. To address the class imbalance, we employ Upsampling/Downsampling. We fine-tune 14 algorithms and compare outcomes with and without radiologists' recommendations. Results show a 99.8% AUC during testing and 59.5% during validation without radiologists' input, which improves to 99.9% and 93.5% respectively with their insights. Expert guidance significantly enhances diagnostic accuracy. The study explores the interplay between algorithmic precision, dataset characteristics, and expert recommendations in breast cancer diagnosis. It provides valuable insights for leveraging technology and expert knowledge for improved medical outcomes.

Keywords Breast Cancer, Machine learning in radiology, Medical image classification, Mammography, Radiology image classification

DOI: 10.19139/soic-2310-5070-2014

1. Introduction

The profound global implications of breast cancer, evident from the staggering statistics of around 2.3 million new cases and 685,000 deaths recorded in 2022 [1], highlight the critical need for immediate and targeted intervention strategies. These statistics underscore the urgency of comprehending regional differences, exemplified by metrics such as the mortality-to-incidence ratio (MIR) at approximately 29.78%. In light of these disparities, it becomes increasingly imperative to develop tailored approaches to address the unique challenges posed by breast cancer across different regions. In this context, the Classification of mammography images plays a vital role in the initial stages of identifying and diagnosing breast malignancy [2]. Mammograms are vital during the evaluation and diagnosis of breast cancer, acting as the primary imaging modality for detecting abnormalities. Their role in early detection allows for prompt intervention and treatment. While other diagnostic methods exist, the initial assessment of mammograms holds particular significance [3]. It serves as a crucial step in identifying suspicious findings and determining the need for additional diagnostic procedures, such as echo mammography [4] and biopsy [5]. Accurate interpretation of mammograms plays a pivotal role in guiding subsequent diagnostic and treatment strategies, highlighting the criticality of the first diagnostic step in the comprehensive evaluation of breast health. However, In the traditional paradigm, the accurate interpretation of mammographic images constitutes a multifaceted task heavily reliant on the expertise of radiologists [6]. As they receive these images, radiologists engage in a meticulous examination of each detail, meticulously assessing the density, texture, BI-RADS [7], and patterns inherent to the breast tissue. This visual scrutiny encompasses the identification of

*Correspondence to: Youssef Lahdoudi (Email: youssef.lahdoudi@usms.ac.ma). MPA, ENSA Khouribga, Sultan Moulay Slimane University, BP 77,Béni Amir, Morocco

ISSN 2310-5070 (online) ISSN 2311-004X (print)

anomalies such as masses, microcalcifications, or architectural distortions, all of which might signal the presence of a potential tumor [8]. In addition to this, radiologists take into account the dimensions, shapes, and margins of identified lesions, crucial determinants in assessing the likelihood of malignancy. The manual interpretation process necessitates a sophisticated grasp of breast anatomy and pathology. Radiologists draw upon their extensive training and experience to discern subtle intricacies within the images, enabling them to make well-informed decisions regarding the presence or absence of abnormalities. However, it's essential to acknowledge that this traditional methodology is inherently subjective, introducing an element of variability between different observers. Furthermore, it can be time-intensive, especially when applied to extensive screening programs. This approach is subjective, prone to inter-observer variability, and can be time-consuming, especially in large-scale screening programs. To overcome these limitations, there is a growing interest in leveraging advanced image classification techniques [9], such as deep learning algorithms, which encompass image processing and ML methods [10], to automate and enhance the accuracy of mammography image classification. Deep learning algorithms offer significant potential in improving mammography image classification. By utilizing large datasets and powerful computational models, these algorithms can learn intricate features and patterns indicative of breast abnormalities [11]. They can detect subtle variations that may not be easily observable to the human eye and continuously improve their classification performance through exposure to diverse datasets. The integration of deep learning techniques into mammography image classification not only complements traditional visual interpretation but also demonstrates positive impacts in terms of time efficiency and accuracy [12]. It should be noted that deep learning does not replace radiologists but rather assists and accelerates their work, thereby providing objective and standardized classification. This integration presents an opportunity to enhance the accuracy of breast cancer detection by reducing subjectivity and providing radiologists with a valuable tool to support their decision-making process. By combining the expertise of radiologists with the computational power of deep learning algorithms, the efficiency and effectiveness of mammography screening programs can be improved. This advancement has the potential to enable earlier detection, and timely interventions, and ultimately improve patient outcomes. In the research conducted by YJ Suh et al.[13], they proposed modifications to the YOLOv5 network for the recognition and categorization of breast malignancy. The modified YOLOv5 attained a remarkable accuracy level of 96.5% and an MCC coefficient of 93.5%, outperforming Faster R-CNN [14] and YOLOv3 [15]. In an unrelated study [16], to combat overfitting challenges associated with small datasets, an approach was presented, involving the utilization of transfer learning (TL) in a deep convolutional neural network (CNN) model[17]. The model achieved high accuracies on various datasets, including 95.5% on the INbreast dataset [18], 97.35% on the DDSM dataset [19], and 96.67% on the BCDR [20] database. Furthermore, in the research performed by Rahman et al. [21], a novel imaging analysis system was proposed for detecting and identifying malignant breast masses in mammograms. The system utilized thresholding and region-based segmentation tools [22] to identify the largest area within a specified threshold. Feature extraction was performed employing a modified ResNet-50 [23] deep CNN, trained to classify mammograms into malignant or benign categories. The proposed system achieved impressive performance on the INbreast database, with an accuracy rate of 93%, AUC score of 93.02%, specificity value of 93.86%, sensitivity measurement of 93.83%, and an F1-score of 93.03%. Compared to other systems, it provided more accurate results and improved visual outcomes. Besides, in the analysis conducted by [24], a novel approach for Segmentation and categorization of breast cancer images is presented. The framework incorporates various models including InceptionV3 [25], ResNet50, DenseNet1 [26], MobileNetV2 [27], and VGG16 [28] are utilized for categorizing mammograms into nonmalignant and malignant groups. Furthermore, a changed U-Net [29] model is employed for accurate breast region partitioning. The framework utilizes CC (Cranio Caudal), and MLO (Mediolateral Oblique) visualizations to improve system accomplishment and overcome the challenge of limited tagged data. Transfer learning and data expansion techniques are implemented to address the indicated issue. The evaluation is performed on three breast radiograph datasets: MIAS [30], DDSM, as well as CBIS-DDSM [31]. The combination of data extension with the adjusted U-Net model and InceptionV3 achieves exceptional results, with an accuracy rate of 98.87%, AUC score of 98.88%, sensitivity value of 98.98%, a precision of 98.79%, as well as F1 score of 97.99%, and a computational time of 0.02 minutes approximately on the DDSM dataset. This comprehensive framework integrates segmentation and classification models, offering an optimal approach for mammary carcinoma detection and categorization. Within this study, we address the challenges associated with mammography image classification

by employing a comprehensive approach that combines deep learning and machine learning techniques. We begin with YOLOv5 [32, 33], a powerful object detection model, for initial tumor detection, followed by the application of DenseNet1 for binary classification. Additionally, we fine-tune multiple ML algorithms [34] to boost the performance of the classification task [35]. Following this approach, our challenge is to leverage these methodologies to maximize the precision, and efficiency of breast tumor diagnosis, ultimately contributing to the advancement of mammography screening programs.

2. Material and Methods

The Material and Methods section in our study is crucial for guiding the progress and evaluating our research, centered on mammography image classification for identifying malignant breast tumors. In AI, algorithm training and proficiency hinge on dataset availability and quality. Thus, our study underscores the importance of acquiring a representative dataset.

2.1. Dataset

The data used in this research was obtained from [36]. This dataset provides descriptive information on a mammography dataset. It includes data from two hospitals, with a total of 11,913 patients examined. Among them, 486 patients were affected by cancer, and 1,171 patients underwent a biopsy. Additionally, 171 patients had breast implants. The dataset comprises a sum of 54,706 X-ray pictures, and Image 1 (Fig 1) showcases illustrations of breast X-rays in the dataset. Each patient typically contributed four images, and some patients contributed more. The dataset includes six different types of views [37], with 'CC'(Craniocaudal view) and 'MLO'(Mediolateral oblique view) being the most common for mammography, while the others serve as supplementary views. Image 1 illustrates the different views available, including CC, MLO, ML(Mediolateral view), and LM (Lateralomedial view), as well as supplementary views such as AT (Axillary view) and LMO (Lateral-medial oblique view). In

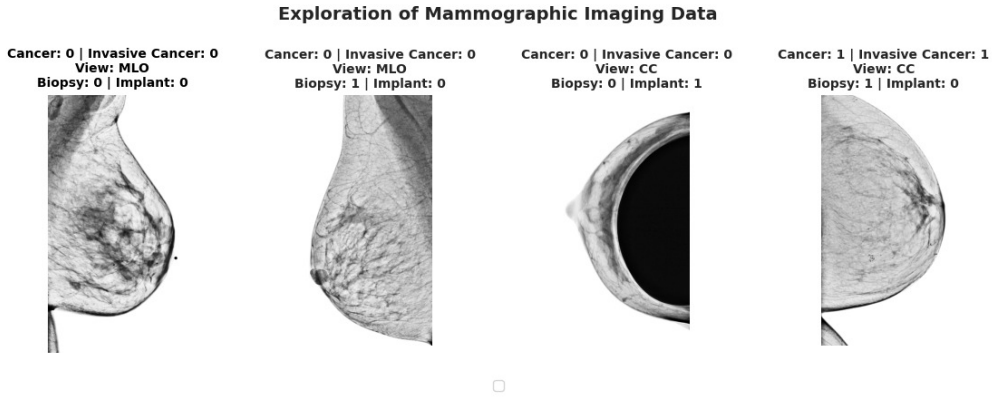
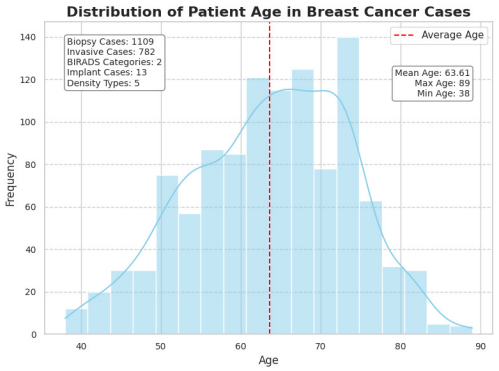


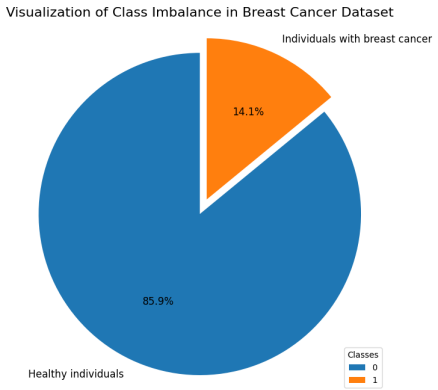
Figure 1. Illustrations of Breast X-rays Images.

addition to the imaging data, the dataset provides information on whether patients underwent a biopsy, the presence of breast implants, and relevant demographic details such as breast density and age. Age is widely recognized as a significant Danger for Malignant breast tumors, as the incidence of the disease tends to increase with advancing age. To investigate the relationship between age and breast tumors, we present the age distribution of mammary tumor patients graphically in Figure 2a. This figure provides a comprehensive visualization of the age distribution among patients diagnosed with malignant breast tumors, offering valuable insights into the relationship between age and the occurrence of the disease. The mean age of patients in our dataset is 63.61 years, providing a central tendency measure that summarizes the age distribution. By examining this graph and considering the mean age, we can observe distribution patterns and identify any notable trends or associations. This visual representation helps

us better understand how age influences the probability of breast cancer and supports our analysis of the impact of age on mammography image classification. Additionally, in Figure 2b, we depict the class imbalance of the dataset, highlighting that positive cases constitute only 14.1% of the total dataset. This imbalance underscores the challenges associated with training models on skewed datasets and emphasizes the need for appropriate techniques to address class imbalance during model development and evaluation.



(a) Age Distribution and Occurrence of Breast Cancer



(b) Imbalanced Dataset Visualization: Class Distribution

Figure 2. Breast Cancer Occurrence and Age Distribution: Class Imbalance Visualization

3. Overall Approach for Mammography Image Classification

Within our study, we employ a comprehensive analysis and a preprocessing pipeline to prepare and extract meaningful information from mammography images for the classification of breast abnormalities. The graph presented in Figure 3 illustrates the three key phases of the assessment: Image Object Detection, Image Embedding, Along with Model Development and Evaluation. In the Image Object Detection phase, YOLOv5 is employed to detect and localize objects of interest within the mammography images. YOLOv5, short for You Only Look Once version 5, is a state-of-the-art object detection model renowned for its efficiency and accuracy [38]. It operates by dividing the image into a grid and predicting bounding boxes and corresponding class probabilities for each grid cell. This phase is pivotal as it allows for the identification and extraction of relevant regions of interest within the mammography images. By pinpointing areas likely to contain abnormalities or lesions indicative of breast cancer, YOLOv5 effectively narrows down the focus of subsequent analysis, streamlining the classification process and reducing computational overhead. Subsequently, in the Image Embedding phase, DenseNet1 is utilized to extract rich and informative feature representations from the preprocessed mammography images. DenseNet121 belongs to the family of Dense Convolutional Neural Networks (DenseNets) and is characterized by its densely connected architecture, where each layer receives feature maps from all preceding layers [39]. This dense connectivity pattern enables the network to capture intricate patterns, textures, and structures present in the images more effectively compared to traditional convolutional neural networks (CNNs). As a result, DenseNet121 is well-suited for tasks requiring detailed feature extraction, such as image classification. In the context of breast cancer detection, DenseNet121 plays a critical role in generating robust feature representations from mammography images, which are subsequently used for accurate classification of malignant and non-malignant cases. By leveraging the capabilities of both YOLOv5 for object detection and DenseNet121 for feature extraction, the proposed framework ensures a comprehensive analysis of mammography images, leading to improved accuracy and reliability in breast cancer diagnosis. The Model Development and Evaluation phase encompasses training, validation, and testing stages, where appropriate strategies such as data augmentation, regularization techniques, and hyperparameter tuning are implemented to enhance and fine-tune the models using GridSearch [40], a popular technique in ML.

GridSearch allowed us to systematically explore different combinations of hyperparameter values to identify the optimal configuration for each algorithm. This phase ensures robust model performance and prevents overfitting by assessing the models' generalization capabilities on independent datasets. The graph provides a clear overview of the workflow and progression through these three distinct phases, highlighting the key steps involved in the research's development and evaluation.

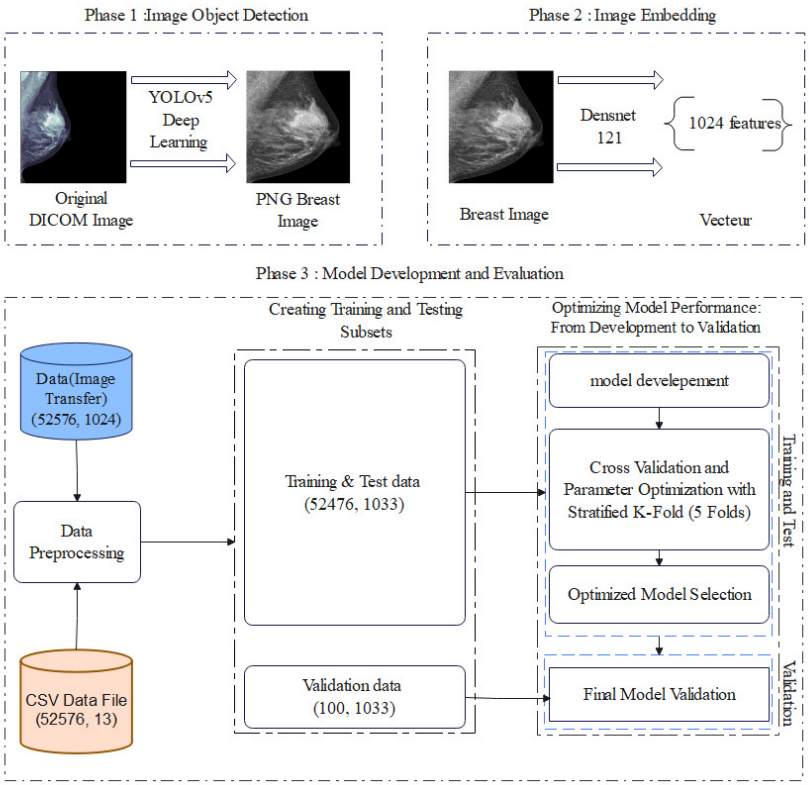


Figure 3. Machine Learning Workflow Visualization: Exploring, Preprocessing, Training, and Analyzing the Dataset

3.1. Model Architectures for Breast X-rays Image Analysis

YOLOv5 (You Only Look Once) excels in instant, high-accuracy object recognition, utilizing a streamlined architecture, advanced data augmentation, and optimized anchor box assignment for single-forward propagation. On the other hand, DenseNet121, known for its dense connectivity pattern, facilitates feature reuse and gradient flow within dense blocks, proving effective for image categorization. In our research, we harness the capabilities of YOLOv5 and DenseNet121 for mammography image analysis in detecting malignant breast tumors.

3.1.1. YOLOv5 Architecture: A Visual Overview of Object Detection and Preprocessing The YOLOv5 architecture is a robust model for object detection, integrating multiple components for accurate and efficient detection of objects in images. It incorporates the CSPDarknet backbone, a variant of the Darknet architecture, enhancing information flow and gradient propagation with Cross Stage Partial connections (CSP). The backbone, with numerous convolutional layers, effectively extracts hierarchical features, capturing intricate details and high-level semantic information. To augment spatial awareness and accommodate objects at varying scales, YOLOv5 leverages the PANet (Path Aggregation Network) neck. Utilizing bottom-up and top-down pathways, the PANet neck combines features from different CSPDarknet backbone levels, achieving fusion through spatial pyramid pooling and feature fusion for proficient object detection. The output module plays a pivotal role in generating

precise predictions. By refining feature maps through additional convolutional layers and anchor-based detection, it delivers accurate localization and classification results. The module consolidates multi-scale predictions from different network levels, employing non-maximum suppression to eliminate redundant detections. For optimized performance, YOLOv5 incorporates operations such as 3x3 convolutions, upsampling, concatenation, and 2D convolutions. These operations synergistically interact with the CSPDarknet backbone, PANet neck, and output module, establishing a robust correlation among components and empowering YOLOv5 to achieve cutting-edge object detection. The integrated architecture of YOLOv5, featuring hierarchical feature extraction, spatial awareness, refined predictions, and additional operations, enables accurate and efficient object detection across diverse images. This comprehensive approach positions YOLOv5 as a valuable asset in various applications.

3.1.2. DenseNet121 Architecture: Feature Extraction for Image Classification The DenseNet121 architecture is a deep learning model extensively employed for image categorizing tasks. It consists of multiple components, including convolutional layers (C1) [41], dense blocks (D1, D2, D3, D4) [42], transition layers (T1, T2, T3) [43], and an output layer [44]. The C1 layer serves as the initial convolutional layer, responsible for processing the input image. It applies a set of filters to extract low-level features and capture local patterns from the input data. The dense blocks (D1, D2, D3, D4) are fundamental building blocks in DenseNet121. Each dense block comprises several densely connected layers. Within each block, the outputted feature maps of all antecedent layers are appended to the feature maps of the current layer. This dense connectivity facilitates effective feature reuse and enables the model to learn rich and diverse representations. The dense connections within the blocks help alleviate the vanishing gradient problem and promote efficient information flow. The transition layers (T1, T2, T3) are responsible for reducing the spatial dimensions of the feature maps while preserving the number of feature channels. They typically consist of 1x1 convolutional layers followed by average pooling. The 1x1 convolutional layers reduce the number of feature maps, optimizing computational efficiency. The subsequent average pooling operation downsamples the feature maps, capturing the most relevant information. Following the final dense block, an average pooling layer is employed to achieve a global representation of the feature maps. The average pooling operation reduces the spatial dimensions to 1x1 while retaining the number of feature channels. This global representation summarizes the input data and captures its overall context. The output layer of DenseNet121 consists of a dense layer with 1024 elements, followed by a softmax activation function [45]. The dense layer performs a fully connected operation, mapping the global representation to a higher-dimensional space. The softmax activation function converts the output into a possibility arrangement across the different classes, enabling the model to make predictions by assigning the highest probability to the most probable class. In conclusion, the DenseNet121 architecture utilizes convolutional layers, dense blocks, transition layers, average pooling, and a dense output layer with softmax activation for image classification. Its dense connectivity, transition mechanisms, and global pooling contribute to its effectiveness in capturing intricate patterns and achieving superior classification performance.

3.2. Experimental Setup and Methodology

In our experimental setup, we followed a well-defined methodology to ensure the reliability and validity of our results. By incorporating transfer learning with YOLOv5, we extracted the healthy regions from mammography images, reducing the data size [32] and eliminating irrelevant areas. This step was crucial for optimizing resources and focusing on relevant regions of interest. Additionally, integrating Densenet121 allowed us to transform the images into compact vectors [46], preserving essential information. During the preprocessing stage, we conducted various steps to prepare the data. We removed empty rows, standardized the data, and encoded categorical variables. Furthermore, as depicted in Figure 4, we identified and removed columns strongly correlated with the target variable, aiming to mitigate their impact on the model's predictions and reduce bias. The process involves calculating the correlation coefficient (rr) between each predictor variable (denoted as XX) and the target variable (denoted as YY). The correlation coefficient is calculated using the formula:

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}} \tag{1}$$

where X_i and Y_i represent individual data points, and $\bar{X}$ and $\bar{Y}$ represent the means of the predictor variable XX and the target variable YY, respectively. If the correlation coefficient (rr) exceeds a predefined threshold (indicating strong correlation), the corresponding predictor variable is removed from the dataset. By eliminating these highly correlated columns, we prevent the model from relying too heavily on specific features, thereby reducing bias and improving the model's predictive performance and interpretability of predictive models [47]. In the initial phase, we prioritize model interpretability and prevent overfitting by carefully selecting 100 samples, including 50 challenging cancer cases for validation. This deliberate and unbiased approach ensures a diverse and representative dataset, fostering robust evaluation. Introducing randomness further mitigates potential biases, reinforcing the generalizability of our findings and contributing to the optimization of breast cancer diagnosis. Addressing class imbalance [48], we employed a hybrid approach [49] using SMOTE [50] and RandomUnderSampler techniques to balance the dataset.

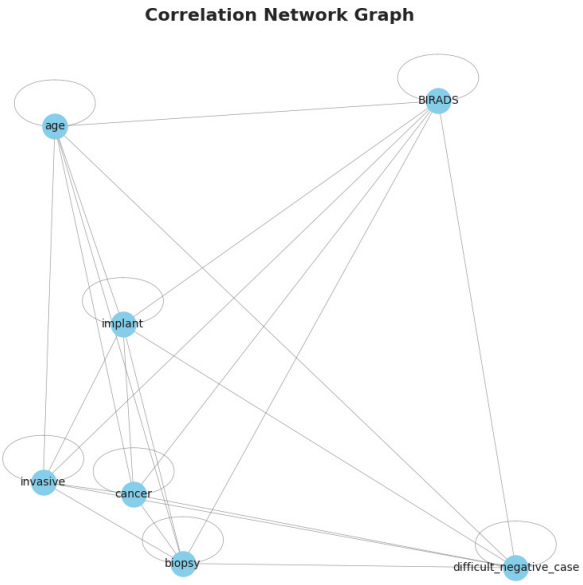


Figure 4. Identification and Removal of Strongly Correlated Columns with the Target Variables

$$\text{SMOTE}(x_i) = x_i + \text{rand}(1, k) \times (\text{neighbor}(x_i) - x_i) \tag{2}$$

The equation provided represents the Synthetic Minority Over-sampling Technique (SMOTE) utilized for generating synthetic samples in imbalanced datasets, specifically targeting the minority class. The approach involves the creation of new samples through interpolation between a minority class instance *(xi)* and its closest neighbors (neighbor*(xi)*). The extent of oversampling is governed by the parameter kk, which governs the quantity of synthetic samples produced. By leveraging SMOTE, we can effectively address the class disproportion problem and enhance the representation of the minority class in the dataset, thereby improving the performance of classification algorithms. After applying data transformations, our dataset comprises 52,576 samples and 1,033 columns, including information about the mammography images, patient characteristics, and the target variable. The processed dataset has a reduced size of 415.6 MB, significantly smaller than its original size of 314 GB. The columns in the dataset offer vital information, including the "laterality" column specifying the breast side (e.g., left or right), the "view" column indicating the type of mammogram view used, the "age" column recording the individual's age at the time of the breast X-ray, the "implant" column denoting whether the patient has breast implants, the "density" column detailing breast density, the "BIRADS" column providing

Breast Imaging Reporting and Data System scores, and the ”difficult negative case” column indicating instances that pose a challenge even for expert diagnosis. Additionally, the ”age category” column categorizes patients into specific age groups for analysis purposes, facilitating a more nuanced understanding of how age influences breast cancer diagnosis and treatment outcomes. Crucially, the target variable, ”cancer,” identifies whether a particular mammography image indicates the presence or absence of breast cancer. Incorporating these variables, including the target, enhances the dataset's richness and allows for a comprehensive exploration of breast cancer classification. With this transformed dataset, we can now delve into further analysis and modeling, leveraging the power of

4. Results and Discussion

The 'Results and Discussion' section displays the findings and inspection of our research on breast cancer detection. This section offers a comprehensive analysis of the outcomes obtained from our experimental evaluations and delves into their implications. Our primary objective was to evaluate the performance of various algorithms in accurately identifying breast cancer cases through the analysis of mammography images. We describe the evaluation metrics employed to gauge the algorithms' effectiveness and offer a detailed examination of the results. Moreover, we explore the confusion matrices [51] of the top-performing algorithms to gain deeper insights into their predictive capabilities. Through a thorough discussion of the outcomes, we aim to emphasize the strengths, limitations, and broader significance of our study, ultimately contributing to the field of breast cancer diagnosis and treatment.

4.1. Results

In rigorously evaluating our breast cancer diagnosis approach, we strategically focus on two key performance metrics: AUC and Precision [52, 53]. AUC serves as a robust indicator of our model's discriminatory power, emphasizing its capacity to effectively differentiate between benign and malignant cases. Simultaneously, Precision delves into the accuracy of our model's positive predictions, critically assessing its ability to minimize false positives, a crucial consideration in the medical context. The strategic choice of AUC and Precision reflects our commitment to a nuanced evaluation that aligns with the specific demands of breast cancer diagnosis. These metrics, tailored to our diagnostic task, provide a comprehensive understanding of the model's performance, ensuring both high discriminatory strength and accurate positive predictions. Additionally, we encountered initial discrepancies in AUC and Precision metrics during the evaluation phase, prompting a thorough investigation into the underlying causes [54, 55]. Through meticulous analysis and model refinement, we addressed these discrepancies by fine-tuning key hyperparameters, optimizing feature selection techniques, and implementing ensemble learning strategies. These adjustments resulted in significant improvements in both AUC and Precision metrics, enhancing the overall performance and reliability of our breast cancer diagnostic approach. Now, let's delve into the equations that quantify the prowess of our approach in these critical aspects.

4.2. Experimental Setup and Methodology

Receiver Operating Characteristic (ROC) Curve Equation:

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{4}$$

Area Under the Curve (AUC) Equation:

$$\text{AUC} = \int_0^1 \text{TPR}(t)\, d\text{FPR}(t) \tag{5}$$

Precision Equation:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{6}$$

These evaluation metrics allow us to evaluate the accomplishment of the classification models and examine their effectiveness in accurately detecting breast cancer cases. In our quest to optimize breast cancer diagnosis, our experimental design incorporates a meticulous selection of hyperparameters, each chosen for its significant impact on model performance and reliability. Inspired by established practices [56], we opt for LightGBM for iterative imputation, despite the missing values already being removed from the dataset [57]. LightGBM's efficiency in handling large datasets and complex features is invaluable, contributing to improved model performance. The Z-score normalization method [58] is crucial, aligning variable scales and facilitating smoother convergence during training, which is essential for preventing certain features from dominating the learning process due to their larger scales. Our model evaluation benefits from a Stratified K-Fold with 5 folds [59], providing a balanced representation of classes for unbiased assessment and minimizing the risk of overfitting or underfitting. Moreover, setting a threshold for outlier removal (threshold: 0.050000) enhances model robustness by mitigating the impact of extreme values on model training. We employ the Yeo-Johnson transformation method to transform skewed data distributions, stabilizing variance, and meeting the assumptions of certain statistical models, thereby improving model interpretability. Through classic feature selection and the deliberate selection of 20% of the most informative features, we aim to simplify the model, reduce overfitting, and enhance interpretability and generalizability. Each hyperparameter choice underscores our commitment to a meticulous methodology that fortifies the robustness and effectiveness of our breast cancer diagnostic approach. The graph illustrates the performance of 14 binary

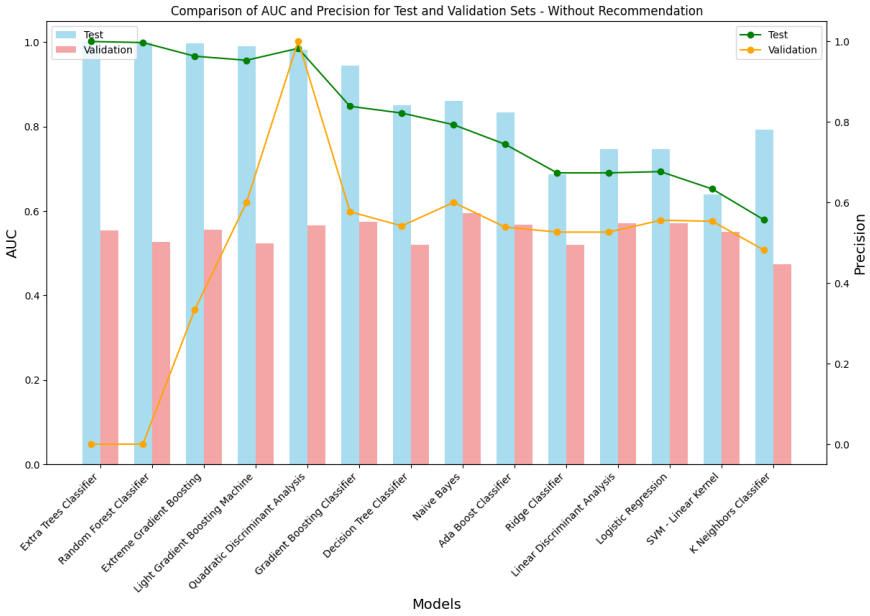


Figure 5. AUC & Precision: Test and Validation Sets - Without Recommendation

classification algorithms on mammographic images, evaluated using AUC and precision metrics. The use of two distinct colors, 'sky blue' for test results and 'light coral' for the validation part, emphasizes the differences between these datasets. Notably, a significant gap in AUC values between test and validation suggests the presence of an underfitting problem. The initial two algorithms display zero precision, indicating shortcomings [60] in these models. Furthermore, the fourth algorithm, Quadratic Discriminant Analysis, exhibits exceptionally high precision, reaching 1 under specific conditions, raising a potential concern that warrants further investigation. the second graph highlights the impact of radiologist recommendations. Notably, there is a discernible difference in precision between the test and validation sets, indicating varied model responses to radiologist guidance. Unlike the first

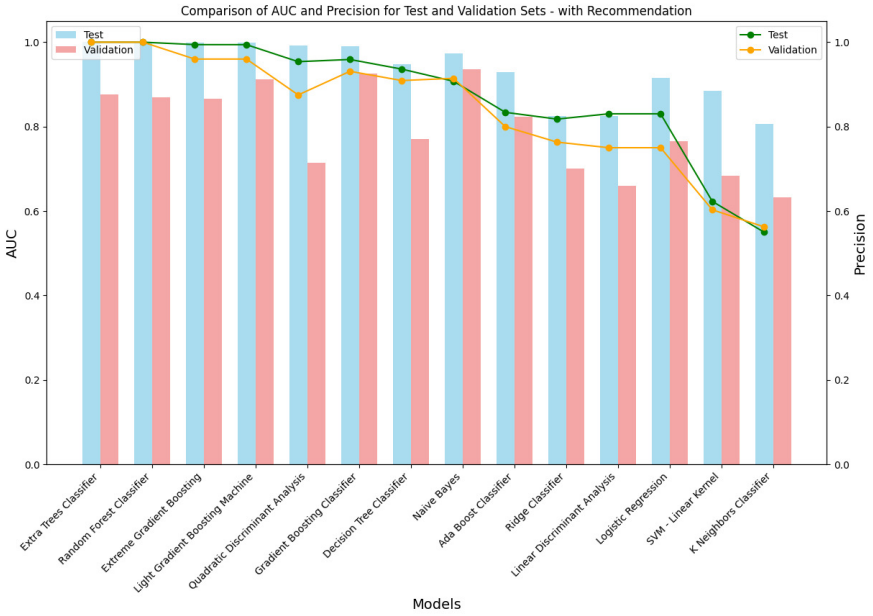


Figure 6. AUC & Precision: Test vs. Validation Sets with Recommendations

graph, the AUC gap between test and validation has reduced, suggesting improved generalization with radiologist recommendations. Despite this alignment, distinct algorithmic variations persist, emphasizing the nuanced impact of radiologist input on model performance The third graph meticulously compares the AUC and precision metrics

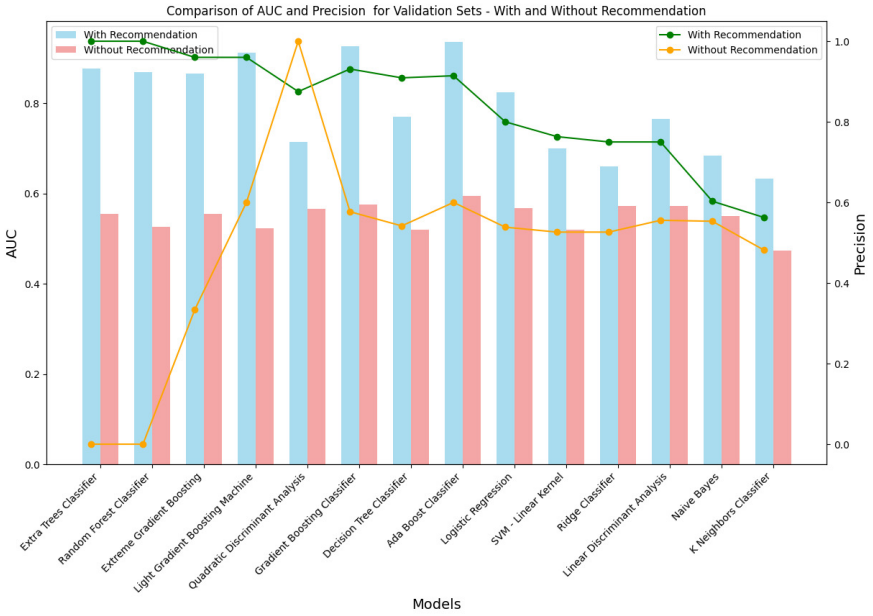


Figure 7. AUC & Precision: Validation Sets - With and Without Recommendation

for validation sets, distinguishing between scenarios with and without radiologist recommendations. The validation sets, representative of data unseen by the algorithms during training, serve as a crucial benchmark to elucidate the

profound impact of radiologist guidance. The primary objective is to unveil disparities in performance attributable to the presence or absence of such guidance. Upon scrutiny, the results underscore a distinct contrast between the two scenarios. With the radiologist's recommendation, the graph reveals remarkable scores and outcomes, reflecting a notable enhancement in both AUC and precision metrics. In stark contrast, the absence of such guidance results in less favorable performance metrics. This stark divergence signifies the pivotal role played by radiologist recommendations in refining algorithmic predictions, reinforcing the significance of expert input in augmenting model performance on previously unseen data.

4.3. Discussion

Breast cancer, a prominent public health concern, has witnessed a surge in incidence rates, emerging as a leading cause of cancer-related deaths among women worldwide in 2020 [61]. Yet, the traditional visual interpretation of mammograms by radiologists faces inherent limitations, including subjectivity and variability in interpretations. To overcome these challenges and improve breast cancer detection, our study integrates deep learning algorithms into mammography image classification. Leveraging the computational capabilities of deep learning algorithms, our aim is to enhance the accuracy and efficiency of breast cancer detection, offering standardized and objective classification methods [2, 35]. This approach aligns with the need to overcome subjectivity and the time-consuming nature of traditional visual interpretation, providing a valuable tool for radiologists in the early detection and diagnosis of breast cancer. In our intricate study, we embarked on the binary classification of mammography images, employing a sophisticated approach that involved initial preprocessing and transformation of the image set. To reduce computational complexity, we leveraged YOLOv5, followed by transforming the images into vectors using DenseNet 121. Subsequently, we curated two distinct sets: one integrated with radiologist recommendations and the other without. The study meticulously separated 100 cases from each set, including 50 positive cases with cancer, to rigorously validate our model. Addressing the challenge of imbalance, we employed a hybrid method, SMOTE, to mitigate the inherent skew in the dataset. The absence of radiologist recommendations unveils a significant gap in AUC values between the test and validation sets, indicative of a potential underfitting problem. Furthermore, the initial two algorithms exhibit zero precision, highlighting inherent shortcomings. Of particular concern is the fourth algorithm, Quadratic Discriminant Analysis, which displays exceptionally high precision under specific conditions, prompting a call for a thorough investigation. A remarkable departure from the first graph is evident in the second graph, notably in the substantial difference in precision between the test and validation sets. Unlike the initial presentation where precisions were relatively consistent, this discrepancy suggests nuanced responses to radiologist recommendations, influencing precision differentially. The second graph also indicates an encouraging reduction in the previously observed gap in AUC values between the test and validation sets. This improvement suggests effective mitigation of potential underfitting issues, showcasing better alignment between the test and validation datasets when incorporating radiologist recommendations [62]. Crucially, the second graph dispels the overfitting concerns witnessed in the first, with AUC values demonstrating a more harmonized performance across evaluated algorithms. This implies that the models better generalize to new data when guided by radiologist recommendations. However, nuanced differences persist among algorithm performances, warranting deeper investigation into their behavior under specific conditions. The third graph meticulously dissects AUC and precision metrics for validation sets, discerning scenarios with and without radiologist recommendations. Serving as a benchmark of unseen data, the validation sets elucidate the profound impact of radiologist guidance on algorithmic predictions. Results underscore a stark contrast between scenarios. With radiologist recommendations, the graph reveals remarkable scores, reflecting enhancements in both AUC and precision metrics. Conversely, the absence of such guidance results in less favorable performance metrics, emphasizing the pivotal role of radiologist recommendations in refining algorithmic predictions on previously unseen data. This study reinforces the significance of expert input in augmenting model performance and highlights the potential for further refinement through the incorporation of radiologist guidance in breast cancer detection models. Additionally, in future studies, we aim to delve deeper into the integration of radiologist recommendations and the multiclassification of BI-RADS to enhance the sophistication and accuracy of breast cancer detection models. Transitioning towards BI-RADS multi-classification involves several steps, including the adaptation of the current model architecture to accommodate multiple classes corresponding to different BI-RADS categories. This transition will necessitate the

collection of annotated data representing each BI-RADS category, followed by model training and validation on this expanded dataset. The potential benefits of this approach are manifold, including improved granularity in breast cancer diagnosis, better characterization of lesions, and enhanced clinical decision-making. By embracing BI-RADS multi-classification, we can provide clinicians with more detailed and informative insights, ultimately leading to better patient outcomes and more personalized treatment strategies.

5. Conclusion

In conclusion, our comprehensive study delved into the intricate realm of binary classification for mammography images in the context of breast cancer detection. Employing a multi-faceted approach that included meticulous preprocessing, transformation via YOLOv5, and vectorization using DenseNet 121, we navigated the complexities of enhancing computational efficiency while maintaining diagnostic accuracy. The evolution of our study unfolded through a meticulous exploration of algorithmic performance with and without radiologist recommendations. The initial graph revealed notable discrepancies in AUC and precision metrics, signaling potential underfitting and overfitting concerns. Subsequent refinements in the second graph showcased improved alignment between test and validation sets when incorporating radiologist recommendations, mitigating the risk of underfitting. Our research provides valuable insights into the nuanced impact of radiologist recommendations on algorithmic performance. The third graph, dissecting AUC and precision metrics for validation sets with and without recommendations, highlighted the transformative role of expert guidance. The stark contrast in outcomes emphasizes the pivotal role of radiologist recommendations in refining algorithmic predictions, particularly in scenarios involving previously unseen data. Furthermore, our study acknowledges the importance of extending the model evaluation to larger and more diverse datasets, incorporating validation studies across multiple medical centers and diverse populations. This expansion ensures the model's generalizability and effectiveness in real-world clinical settings. Additionally, we emphasize the significance of advanced visualization techniques and interpretability methods, which provide valuable insights into the decision-making process of the classification model and enhance its transparency and credibility. This approach empowers healthcare professionals to better understand and interpret the model's predictions, facilitating informed clinical decisions. In future work, our study will transition towards BI-RADS multi-classification, expanding our model's capabilities to classify mammographic images based on a broader spectrum of breast cancer risk categories. By addressing the challenges associated with false negatives and exploring future research directions, we aim to further advance the accuracy and effectiveness of breast cancer detection models. Ultimately, our research contributes to ongoing efforts in early detection, improved patient outcomes, and the reduction of the societal burden of breast cancer

Declarations

- Funding: No financial support was received for this research
- Conflict of Interest: The authors declare no conflicts of interest
- Ethics approval and consent to participate: Not applicable
- Consent for publication: Not applicable
- Data availability [36]

REFERENCES

[1] J. Ferlay, M. Ervik, F. Lam, M. Laversanne, M. Colombet, L. Mery, M. Piñeros, A. Znaor, I. Soerjomataram, and F. Bray. Global cancer observatory: Cancer today, lyon, france: International agency for research on cancer. available from. https://gco.iarc.who.int/today, 2024.

[2] Hwejin Jung, Bumsoo Kim, Inyeop Lee, Minhwan Yoo, Junhyun Lee, Sooyoun Ham, Okhee Woo, and Jaewoo Kang. Detection of masses in mammograms using a one-stage object detector based on a deep convolutional neural network. *PloS one*, 13(9):e0203355, 2018.

[3] Philippe Autier and Mathieu Boniol. Mammography screening: A major issue in medicine. *European journal of cancer*, 90:34–62, 2018.

[4] Koji Ohnuki, Eriko Tohno, Hiroko Tsunoda, Takayoshi Uematsu, and Yasuo Nakajima. Overall assessment system of combined mammography and ultrasound for breast cancer screening in japan. *Breast Cancer*, 28(2):254–262, 2021.

[5] José Alfonso Cruz-Ramos, Mijaíl Irak Trapero-Corona, Ingrid Aurora Valencia-Hernández, Luz Amparo Gómez-Vargas, María Teresa Toranzo-Delgado, Karla Raquel Cano-Magaña, Emmanuel De la Mora-Jiménez, and Gabriela del Carmen López-Armas. Strain elastography fat-to-lesion index is associated with mammography bi-rads grading, biopsy, and molecular phenotype in breast cancer. *Biosensors*, 14(2):94, 2024.

[6] Osama Bin Naeem, Yasir Saleem, M Khan, Amjad Rehman Khan, Tanzila Saba, Saeed Ali Bahaj, and Noor Ayesha. Breast mammograms diagnosis using deep learning: State of art tutorial review. *Archives of Computational Methods in Engineering*, pages 1–19, 2024.

[7] FTH Bodewes, AA van Asselt, MD Dorrius, MJW Greuter, and GH de Bock. Mammographic breast density and the risk of breast cancer: A systematic review and meta-analysis. *The Breast*, 2022.

[8] Karthikeyan Velayuthapandian, Gopalakrishnan Karuppiah, Sridhar Raj Sankara Vadivel, and Dani Reagan Vivek Joseph. Mammogram data analysis: Trends, challenges, and future directions. In *Computational Intelligence and Modelling Techniques for Disease Detection in Mammogram Images*, pages 1–38. Elsevier, 2024.

[9] Syed M Anwar and Ulas Bagci. Artificial intelligence as another set of eyes in breast cancer diagnosis. *Journal of Medical Artificial Intelligence*, 2(May):10, 2019.

[10] Yoichi Hayashi. Toward the transparency of deep learning in radiological imaging: Beyond quantitative to qualitative artificial intelligence. *Journal of Medical Artificial Intelligence*, 2, 2019.

[11] Parita Oza. Ai in breast imaging: Applications, challenges, and future research. In *Computational intelligence and modelling techniques for disease detection in mammogram images*, pages 39–54. Elsevier, 2024.

[12] Ahmed Hosny, Chintan Parmar, John Quackenbush, Lawrence H Schwartz, and Hugo JWL Aerts. Artificial intelligence in radiology. *Nature Reviews Cancer*, 18(8):500–510, 2018.

[13] Yong Joon Suh, Jaewon Jung, and Bum-Joo Cho. Automated breast cancer detection in digital mammograms of various densities via deep learning. *Journal of personalized medicine*, 10(4):211, 2020.

[14] Han Qin, Jizhou Wang, Xi Mao, Zhan'ao Zhao, Xuanyu Gao, and Wenjuan Lu. An improved faster r-cnn method for landslide detection in remote sensing images. *Journal of Geovisualization and Spatial Analysis*, 8(1):2, 2024.

[15] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[16] Hiba Chougrad, Hamid Zouaki, and Omar Alheyane. Deep convolutional neural networks for breast cancer screening. *Computer methods and programs in biomedicine*, 157:19–30, 2018.

[17] Waseem Rawat and Zenghui Wang. Deep convolutional neural networks for image classification: A comprehensive review. *Neural computation*, 29(9):2352–2449, 2017.

[18] Inês C Moreira, Igor Amaral, Inês Domingues, António Cardoso, Maria Joao Cardoso, and Jaime S Cardoso. Inbreast: toward a full-field digital mammographic database. *Academic radiology*, 19(2):236–248, 2012.

[19] Jawad Ahmad, Sheeraz Akram, Arfan Jaffar, Muhammad Rashid, and Sohail Masood Bhatti. Breast cancer detection using deep learning: An investigation using the ddsm dataset and a customized alexnet and support vector machine. *IEEE Access*, 2023.

[20] Nadia Kermouni Serradj, Sihem Lazzouni, and Mahammed Messadi. Mammograms enhancement based on multifractal measures for microcalcifications detection. *International Journal of Biomedical Engineering and Technology*, 41(1):60–82, 2023.

[21] Hameedur Rahman, Tanvir Fatima Naik Bukht, Rozilawati Ahmad, Ahmad Almadhor, Abdul Rehman Javed, et al. Efficient breast cancer diagnosis from complex mammographic images using deep convolutional neural network. *Computational intelligence and neuroscience*, 2023, 2023.

[22] Yuzhou Hu, Yi Guo, Yuanyuan Wang, Jinhua Yu, Jiawei Li, Shichong Zhou, and Cai Chang. Automatic tumor segmentation in breast ultrasound images using a dilated fully convolutional network combined with an active contour model. *Medical physics*, 46(1):215–228, 2019.

[23] Warid Islam, Meredith Jones, Rowzat Faiz, Negar Sadeghipour, Yuchen Qiu, and Bin Zheng. Improving performance of breast lesion classification using a resnet50 model optimized with a novel attention mechanism. *Tomography*, 8(5):2411–2425, 2022.

[24] Wessam M Salama and Moustafa H Aly. Deep learning in mammography images segmentation and classification: Automated cnn approach. *Alexandria Engineering Journal*, 60(5):4701–4709, 2021.

[25] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[26] Raj Kumar Pattanaik, Satyasis Mishra, Mohammed Siddique, Tiruveedula Gopikrishna, Sunita Satapathy, et al. Breast cancer classification from mammogram images using extreme learning machine-based densenet121 model. *Journal of Sensors*, 2022, 2022.

[27] Aarti Bokade and Ankit Shah. Breast cancer diagnosis in mammography images using deep convolutional neural network-based transfer and scratch learning approach. *Indian Journal of Science and Technology*, 16(18):1385–1394, 2023.

[28] Sashikanta Prusty, Sujit Kumar Dash, and Srikanta Patnaik. A novel transfer learning technique for detecting breast cancer mammograms using vgg16 bottleneck feature. *ECS Transactions*, 107(1):733, 2022.

[29] Md Shamim Hossain. Microc alcification segmentation using modified u-net segmentation network from mammogram images. *Journal of King Saud University-Computer and Information Sciences*, 34(2):86–94, 2022.

[30] Kristina Ionkina, Andrey Svistunov, Ilya Galin, Boris Onykiy, and Larisa Pronicheva. Mias database semantic structure. *Procedia computer science*, 145:254–259, 2018.

[31] Lenin G Falconi, Maria Perez, Wilbert G Aguilar, and Aura Conci. Transfer learning and fine tuning in breast mammogram abnormalities classification on cbis-ddsm database. *Adv. Sci. Technol. Eng. Syst. J*, 5(2):154–165, 2020.

[32] Yiming Fang, Xianxin Guo, Kun Chen, Zhu Zhou, and Qing Ye. Accurate and automated detection of surface knots on sawn timbers using yolo-v5 model. *BioResources*, 16(3), 2021.

[33] Baizheng Wu, Chengxin Pang, Xinhua Zeng, and Xing Hu. Me-yolo: Improved yolov5 for detecting medical personal protective equipment. *Applied Sciences*, 12(23):11978, 2022.

[34] FY Osisanwo, JET Akinsola, O Awodele, JO Hinmikaiye, O Olakanmi, J Akinjobi, et al. Supervised machine learning algorithms: classification and comparison. *International Journal of Computer Trends and Technology (IJCTT)*, 48(3):128–138, 2017.

[35] Li Shen, Laurie R Margolies, Joseph H Rothstein, Eugene Fluder, Russell McBride, and Weiva Sieh. Deep learning to improve breast cancer detection on screening mammography. *Scientific reports*, 9(1):12495, 2019.

[36] Keri Stephens. Rsna announces launch of screening mammography breast cancer detection ai challenge. *AXIS Imaging News*, 2022.

[37] Gustavo Carneiro, Jacinto Nascimento, and Andrew P Bradley. Automated analysis of unregistered multi-view mammograms with deep learning. *IEEE transactions on medical imaging*, 36(11):2355–2365, 2017.

[38] Francesco Prinzi, Marco Insalaco, Alessia Orlando, Salvatore Gaglio, and Salvatore Vitabile. A yolo-based model for breast cancer detection in mammograms. *Cognitive Computation*, 16(1):107–120, 2024.

[39] Gurusiddayya Hiremath, Jose Alex Mathew, and Naveen Kumar Boraiah. Hybrid statistical and texture features with densenet 121 for breast cancer classification. *International Journal of Intelligent Engineering & Systems*, 16(2), 2023.

[40] Petro Liashchynskyi and Pavlo Liashchynskyi. Grid search, random search, genetic algorithm: a big comparison for nas. *arXiv preprint arXiv:1912.06059*, 2019.

[41] Saad Albawi, Tareq Abed Mohammed, and Saad Al-Zawi. Understanding of a convolutional neural network. In *2017 international conference on engineering and technology (ICET)*, pages 1–6. Ieee, 2017.

[42] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[43] Kelei Wang and Juncheng Wei. Second order estimate on transition layers. *Advances in Mathematics*, 358:106856, 2019.

[44] Ovy Rochmawanti and Fitri Utaminingrum. Chest x-ray image to classify lung diseases in different resolution size using densenet-121 architectures. In *Proceedings of the 6th International Conference on Sustainable Information Engineering and Technology*, pages 327–331, 2021.

[45] Rob A Dunne and Norm A Campbell. On the pairing of the softmax activation and cross-entropy penalty functions and the derivation of the softmax activation function. In *Proc. 8th Aust. Conf. on the Neural Networks, Melbourne*, volume 181, page 185. Citeseer, 1997.

[46] Naresh Kumar, Manoj Sharma, Vijay Pal Singh, Charanjeet Madan, and Seema Mehandia. An empirical study of handcrafted and dense feature extraction techniques for lung and colon cancer classification from histopathological images. *Biomedical Signal Processing and Control*, 75:103596, 2022.

[47] Muhammad Ishfaq, Syed Zahid Ali Shah, Ijaz Ahmad, and Ziaur Rahman. Multinomial classification of nlrp3 inhibitory compounds based on large scale machine learning approaches. *Molecular Diversity*, pages 1–20, 2023.

[48] Fadi Thabtah, Suhel Hammoud, Firuz Kamalov, and Amanda Gonsalves. Data imbalance in classification: Experimental evaluation. *Information Sciences*, 513:429–441, 2020.

[49] Bartosz Krawczyk. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, 5(4):221–232, 2016.

[50] Alberto Fernández, Salvador Garcia, Francisco Herrera, and Nitesh V Chawla. Smote for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. *Journal of artificial intelligence research*, 61:863–905, 2018.

[51] Damir Krstinić, Maja Braović, Ljiljana Šerić, and Dunja Božić-Štulić. Multi-label classifier performance evaluation with confusion matrix. *Computer Science & Information Technology*, 1:1–14, 2020.

[52] Sabri Boughorbel, Fethi Jarray, and Mohammed El-Anbari. Optimal classifier for imbalanced data using matthews correlation coefficient metric. *PloS one*, 12(6):e0177678, 2017.

[53] David MW Powers. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061*, 2020.

[54] Xiyue Wang, Tao Shen, Sen Yang, Jun Lan, Yanming Xu, Minghui Wang, Jing Zhang, and Xiao Han. A deep learning algorithm for automatic detection and classification of acute intracranial hemorrhages in head ct scans. *NeuroImage: Clinical*, 32:102785, 2021.

[55] Selina Sharmin, Tanvir Ahammad, Md Alamin Talukder, and Partho Ghose. A hybrid dependable deep feature extraction and ensemble-based machine learning approach for breast cancer detection. *IEEE Access*, 2023.

[56] Li Yang and Abdallah Shami. On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, 415:295–316, 2020.

[57] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, 2017.

[58] Henderi Henderi, Tri Wahyuningsih, and Efana Rahwanto. Comparison of min-max normalization and z-score normalization in the k-nearest neighbor (knn) algorithm to test the accuracy of types of breast cancer. *International Journal of Informatics and Information Systems*, 4(1):13–20, 2021.

[59] Sashikanta Prusty, Srikanta Patnaik, and Sujit Kumar Dash. Skcv: Stratified k-fold cross-validation on ml classifiers for predicting cervical cancer. *Frontiers in Nanotechnology*, 4:972421, 2022.

[60] Martin J Willemink, Wojciech A Koszek, Cailin Hardell, Jie Wu, Dominik Fleischmann, Hugh Harvey, Les R Folio, Ronald M Summers, Daniel L Rubin, and Matthew P Lungren. Preparing medical imaging data for machine learning. *Radiology*, 295(1):4–15, 2020.

[61] Wei Cao, Hong-Da Chen, Yi-Wen Yu, Ni Li, and Wan-Qing Chen. Changing profiles of cancer burden worldwide and in china: a secondary analysis of the global cancer statistics 2020. *Chinese medical journal*, 134(07):783–791, 2021.

[62] Leila Abdelrahman, Manal Al Ghamdi, Fernando Collado-Mesa, and Mohamed Abdel-Mottaleb. Convolutional neural networks for breast cancer detection in mammography: A survey. *Computers in biology and medicine*, 131:104248, 2021.