DETECTION AND CLASSIFICATION OF BREAST CANCER CALCIFICATIONS USING MACHINE LEARNING WITH AUGMENTATION TECHNIQUE

NURUL SYUHAIDA BINTI BORHANUDDIN

SCHOOL OF HEALTH SCIENCES

UNIVERSITI SAINS MALAYSIA

2025

DETECTION AND CLASSIFICATION OF BREAST CANCER CALCIFICATIONS USING MACHINE LEARNING WITH AUGMENTATION TECHNIQUE

by

NURUL SYUHAIDA BINTI BORHANUDDIN

Dissertation submitted in partial fulfilment
of the requirement for the degree
of Bachelor of Medical Radiation (Honours)

JULY 2025

CERTIFICATE

This is to certify that the dissertation entitled "**DETECTION AND CLASSIFICATION OF BREAST CANCER CALCIFICATIONS USING MACHINE LEARNING WITH AUGMENTATION TECHNIQUE**" is the bona fide record of research work done by **NURUL SYUHAIDA BINTI BORHANUDDIN** during the period from December 2024 to July 2025 under my supervision. I have read this dissertation and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation to be submitted in partial fulfilment for the degree of Bachelor of Health Science Medical Radiation (Honours).

Main Supervisor:

……

Madam Siti Aishah Abd Aziz
Senior Lecturer
School of Health Sciences
Universiti Sains Malaysia
Health Campus
16150 Kubang Kerian
Kelantan, Malaysia

Co-Supervisor:

……

Dr Lau Chiew Chea
Lecturer
School of Medical Sciences
Universiti Sains Malaysia
Health Campus
16150 Kubang Kerian
Kelantan, Malaysia

Date: July 2025

DECLARARTION

I, Nurul Syuhaida Binti Borhanuddin, hereby declare that the dissertation entitled "**DETECTION AND CLASSIFICATION OF BREAST CANCER CALCIFICATIONS USING MACHINE LEARNING WITH AUGMENTATION TECHNIQUE**" is the result of my own investigations, except where otherwise stated and duly acknowledged. I also declare that it has not been previously or concurrently submitted as a whole for any other degrees at Universiti Sains Malaysia or other institutions. I grant Universiti Sains Malaysia the right to use the dissertation for teaching, research and promotional purposes.

……

Nurul Syuhaida Binti Borhanuddin

Date: July 2025

ACKNOWLEDGEMENT

In the name of Allah, the Most Gracious and the Most Merciful. All praise be to Allah, for His blessings and guidance that have enabled me to complete this final year project as part of the requirement for the completion of my undergraduate degree.

First and foremost, I would like to express my deepest appreciation to my beloved parents, Borhanuddin bin Rabipe and Saadiah binti Sata, for their unconditional love, endless prayers, and unwavering support throughout my academic journey. Their sacrifices and encouragement have been the foundation of my strength and perseverance. I would also like to extend my heartfelt gratitude to my main supervisor, Madam Siti Aishah Abd Aziz, for her valuable guidance, continuous encouragement, and insightful supervision throughout the entire research process. My sincere thanks go to my co-supervisor, Dr. Lau Chiew Chea, for her technical expertise, thoughtful suggestions, and generous support, which greatly enhanced the quality of this research.

Special appreciation is also extended to to Dr. Muhammad Akmal Bin Remli, Director of Institute for Artificial Intelligence and Big Data (AIBIG), Universiti Malaysia Kelantan, for providing access to resources and a supportive learning environment. I am equally grateful to his dedicated PhD students, Ainin Sofia Jusoh and Meor Muhammad Muaz from AiBIG, whose assistance and knowledge sharing made complex computational tasks manageable. Lastly, I would like to thank all lecturers, friends, and colleagues who have contributed directly or indirectly to the completion of this thesis. Your encouragement and support have meant a great deal to me.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF EQUATIONS

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
AUC	Area Under Curve
AEC	Automatic Exposure Control
BI-RADS	Breast Imaging Reporting and Data System
CAD	Computer-Aided Diagnosis
CC	Cranial Caudal
CLAHE	Contrast Limited Adaptive Histogram Equalization
CNN	Convolutional Neural Network
CT	Computed Tomography
DICOM	Digital Imaging and Communications in Medicine
DL	Deep Learning
DBT	Digital Breast Tomosynthesis
FN	False Negatives False Positives
FP	False Negatives False Positives
FPR	False Positive Rate
HPUSM	Hospital Pakar Universiti Sains Malaysia
KNN	K-Nearest Neighbor
LR	Logistic Regression
ML	Machine Learning
MLO	Mediolateral Oblique
MRI	Magnetic Resonance Imaging
OpComp	Optimal Compression

PACS	Picture Archiving and Communication System
RF	Random Forest
ROC	Receiver Operating Characteristic
ROI	Region Of Interest
SVM	Support Vector Machine
TiCEM	Titanium Contrast Enhanced Mammography
TN	True Negatives
TP	True Positives
TPR	True Positive Rate
2D	2-Dimensional
3D	3-Dimensional

PENGESANAN DAN PENGELASAN KALSIFIKASI KANSER PAYUDARA MENGGUNAKAN PEMBELAJARAN MESIN DENGAN TEKNIK AUGMENTASI

ABSTRAK

Kalsifikasi kanser payudara merupakan antara petunjuk awal malignan tetapi sering sukar dikesan kerana penampilannya yang halus dan bergantung kepada tafsiran radiologi yang subjektif. Kajian ini bertujuan untuk meningkatkan pengesanan dan pengelasan kalsifikasi payudara dalam imej mamogram melalui integrasi teknik pra-pemprosesan imej dan augmentasi bersama pembelajaran mesin. Sebanyak 234 imej mamogram beranotasi telah dikumpulkan daripada Sistem Pengarkiban dan Komunikasi Imej di Hospital Pakar Universiti Sains Malaysia. Imej-imej ini telah diaugmentasi menggunakan pelbagai transformasi termasuk putaran, pusingan cermin, kabur Gaussian, dan deformasi elastik, menghasilkan jumlah dataset sebanyak 2574 imej bagi meningkatkan kepelbagaian dan mengurangkan risiko overfitting. Teknik pra-pemprosesan seperti penukaran ke skala kelabu, peningkatan kontras menggunakan CLAHE, dan penapisan top hat telah digunakan untuk meningkatkan keterlihatan ciri kalsifikasi. Lima model pembelajaran mesin telah dinilai termasuk Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Random Forest (RF), Logistic Regression (LR), dan model ensemble Soft Voting. Prestasi model diukur menggunakan ketepatan, ketepatan positif (precision), kebolehpulihan (recall), spesifikiti, dan skor F1. Pengesahan dilakukan menggunakan 5-fold cross validation dan kepentingan statistik diuji dengan ujian Friedman dan ujian Wilcoxon Signed Rank. Berdasarkan keputusan, model KNN mencapai ketepatan purata tertinggi sebanyak 87.61% diikuti oleh SVM sebanyak 79.07%, RF sebanyak 78.64%, dan

LR sebanyak 69.62%. Penemuan menunjukkan bahawa model KNN sangat berkesan dalam membezakan antara kalsifikasi benign dan malignan kerana kepekaannya terhadap corak ciri tempatan. Walaupun Logistic Regression mempunyai masa latihan paling singkat, ia menunjukkan prestasi terburuk dalam semua metrik penilaian, menandakan bahawa kelajuan latihan sahaja tidak mencukupi sebagai ukuran utiliti diagnostik. Keputusan juga menegaskan bahawa augmentasi dan pra-pemprosesan yang betul bukan sahaja meningkatkan ketepatan model tetapi juga menyumbang kepada prestasi yang lebih seimbang antara kepekaan dan spesifikiti. Penggunaan pengesahan statistik mengesahkan bahawa perbezaan antara prestasi model adalah signifikan, sekali gus mengukuhkan kebolehpercayaan penemuan. Kajian ini menunjukkan bahawa model pembelajaran mesin apabila disokong oleh strategi penyediaan data yang betul, boleh menjadi alat yang berkesan dalam pembangunan sistem diagnostik bantuan komputer untuk pengesanan awal kanser payudara.

Kata kunci: Kanser payudara, kalsifikasi, mamogram, pembelajaran mesin, augmentasi, pengelasan

DETECTION AND CLASSIFICATION OF BREAST CANCER CALCIFICATIONS USING MACHINE LEARNING WITH AUGMENTATION TECHNIQUE

ABSTRACT

Breast cancer calcifications are among the earliest indicators of malignancy but are often difficult to detect due to their subtle appearance and reliance on subjective radiological interpretation. This study aimed to enhance the detection and classification of breast calcifications in mammographic images through the integration of image preprocessing and augmentation techniques with machine learning. A total of 234 annotated mammographic images were collected from the Picture Archiving and Communication System at Hospital Pakar Universiti Sains Malaysia. These images were augmented using various transformations including rotation, flipping, Gaussian blur, and elastic deformation, resulting in a total dataset of 2574 images to improve variability and reduce the risk of overfitting. Preprocessing techniques such as grayscale conversion, contrast enhancement using CLAHE, and top hat filtering were applied to improve the visibility of calcification features. Five machine learning models were evaluated including Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Random Forest (RF), Logistic Regression (LR), and a Soft Voting ensemble model. Model performance was measured using accuracy, precision, recall, specificity, and F1 score. Validation was performed using 5-fold cross validation and statistical significance was tested with the Friedman test and Wilcoxon Signed Rank test. Based on the results, the KNN model achieved the highest average accuracy of 87.61% followed by SVM at 79.07%, RF at 78.64% and LR at 69.62%. The findings suggest that the KNN model was particularly effective at

distinguishing between benign and malignant calcifications due to its sensitivity to local feature patterns. Although Logistic Regression had the shortest training time, it performed the poorest in all evaluation metrics, indicating that training speed alone is not a sufficient measure of diagnostic utility. The results also highlight that proper augmentation and preprocessing not only improve model accuracy but also contribute to more balanced performance across sensitivity and specificity. The use of statistical validation confirmed that differences among the model performances were significant, thus reinforcing the reliability of the findings. This study demonstrates that machine learning models when supported by proper data preparation strategies, can serve as effective tools in the development of computer assisted diagnostic systems for early breast cancer detection.

Keywords: Breast cancer, calcification, mammogram, machine learning, augmentation, classification

CHAPTER 1

INTRODUCTION

1.1 Background of the Study

Breast cancer remains a critical health concern for women globally, requiring continuous advancements in early detection and diagnosis (Nawaz et al., 2018). However, the interpretation of medical images, such as mammograms, often relies on manual analysis by radiologists, which can be subjective and increases the risk of diagnostic variability and human error (Madani et al., 2022). To address these limitations, artificial intelligence (AI), particularly machine learning (ML) and deep learning (DL) techniques, has appeared as a promising approach to enhance the accuracy and efficiency of breast cancer detection in medical imaging (Madani et al., 2022). The integration of artificial intelligence into biomedical technology has led to remarkable progress in various domains (Kormpos et al., 2025). The application of artificial intelligence in breast cancer screening and detection has huge significance, potentially saving radiologists time and compensating for beginners' inexperience (Gg et al., 2019).

The development and refinement of computer-aided diagnosis (CAD) systems utilising deep learning methodologies offer a pathway to automatically analyse breast images and improve diagnostic accuracy, reducing the need for manual feature extraction (Jiménez-Gaona et al., 2020). Nonetheless, the success of such models heavily depends on the availability of large, diverse, and balanced datasets which the conditions that are often unmet in medical imaging, especially in detecting subtle features like calcifications. Data augmentation has appeared as a powerful strategy to address dataset limitations by

synthetically increasing training sample variability (Shorten el al., 2019). While conventional augmentation techniques can help reduce overfitting, they often failed to capture the complex textures and subtle patterns of calcifications in mammographic images (Oza et al., 2022). Oza et al. (2022) stated that recent advances, including the use of generative adversarial networks (GANs) and medical-specific transformations, offer promising alternatives but remain underexplored in this context.

1.2 Problem Statement

Calcifications can vary significantly in size, shape, and distribution. This variability can make it difficult for radiologists to develop a consistent approach in identifying them. The presence of overlapping structures in mammograms can further complicate the detection process. Radiologists often face difficulties in accurately analysing mammograms to differentiate between benign and malignant calcifications. This can lead to unnecessary biopsies or missed diagnoses, adversely affecting patient outcomes (Mahmood et al., 2021).

The ability to accurately detect calcifications is highly dependent on the radiologist's experience and training. Less experienced radiologists may struggle more with identifying calcifications, leading to variability in diagnostic accuracy across different practitioners. The analysis of mammograms is often time-consuming and subjective, which can result in inconsistencies in diagnosis. This subjectivity in interpretation can lead to missed detections or false positives, highlighting the need for more reliable methods (Grinet et al., 2024).

The primary goal is to enhance the diagnostic accuracy of CAD systems in detecting breast cancer calcifications by leveraging machine learning models and augmentation methods to improve feature extraction and classification. The ability of models to generalise new data is crucial for their practical application in medical imaging (Prodan et al., 2023). However, achieving this capability is challenging, especially when the training dataset is small or imbalanced. The validation error should ideally decrease alongside the training error, indicating that the model is learning effectively and can generalise well. Overfitting is a common issue in CAD systems due to the limited dataset size, emphasising the need for augmentation techniques to prevent overfitting and improve model generalisation (Prodan et al., 2023).

1.3 Objective

1.3.1 General Objective

To improve detection accuracy of calcifications in mammograms using augmentation techniques.

1.3.2 Specific Objectives

1. To apply image processing to specify details of calcifications by using augmentation techniques.
2. To classify the image of breast calcifications as benign or malignant using machine learning models.
3. To evaluate the performance of machine learning models in classifying the image of breast calcifications.

1.4 Hypothesis

1.4.1 Null Hypothesis

Machine learning classifiers failed in differentiation between benign and malignant breast calcification compared to BIRADS' subjective assessment with pathology report evidence.

1.4.2 Alternative Hypothesis

Machine learning classifiers has high accuracy in differentiation between benign and malignant breast calcification compared to BIRADS' subjective assessment with pathology report evidence.

1.5 Significant of Study

While augmentation techniques are essential for improving breast cancer detection, addressing their limitations is crucial for enhancing model robustness and generalisation in clinical settings. There are still future research that should focus on developing standardised protocols for augmentation and exploring the integration of diverse datasets. Therefore, this study aims to explore and apply various data augmentation techniques to enhance the performance of machine learning models in detecting and classifying breast cancer calcifications in mammographic images. By implementing and evaluating a range of augmentation methods, the study seeks to determine their impact on model robustness, accuracy, and generalisability. The findings aim to provide practical insights into the application of augmentation strategies for developing reliable AI-based diagnostic tools, ultimately improving early detection of breast cancer calcifications.

1.6 Conceptual Framework

This study proposes a structured framework for developing and evaluating machine learning models to detect breast cancer calcifications. It begins with labeling mammographic images as malignant or benign, followed by data augmentation to improve variability and address class imbalance. The images are then preprocessed, and key features are extracted for classification. The dataset is split into training and testing sets (80:20) to ensure reliable model development. Machine learning classifiers are trained and validated, and their performance is evaluated using standard metrics. This framework provides a systematic and replicable approach to enhancing the accuracy and reliability of computer-aided breast cancer detection.

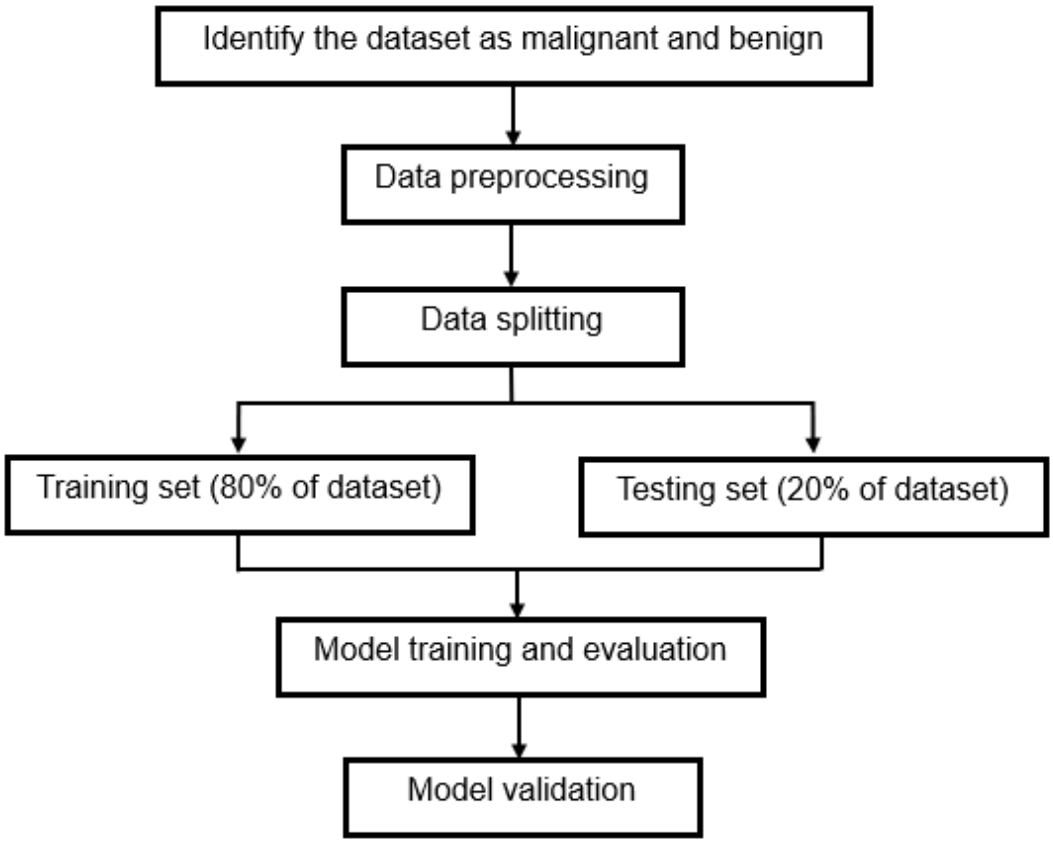

Figure 1.1 Conceptual Framework adapted from Aziz et al. (2024)

CHAPTER 2

LITERATURE REVIEW

2.1 Breast Cancer Calcifications and Imaging Technique

Breast cancer disease is one of the most recorded cancers which leads to morbidity and death among women around the world. Breast cancer occurs when the cells in the lobules or the ducts become abnormal and divide uncontrollably (Jiménez-Gaona et al., 2020). These abnormal cells begin to invade the surrounding breast tissue and may eventually spread via blood vessels and lymphatic channels to the lymph nodes, lungs, bones, brain and liver. Breast calcifications are small calcium deposits that can appear in mammograms and are often associated with breast cancer. Their presence can indicate benign or malignant lesions, making accurate diagnosis crucial for effective treatment (Udoh et al., 2024). The challenge lies in distinguishing between these types of lesions, as the automatic classification of microcalcification clusters remains complex.

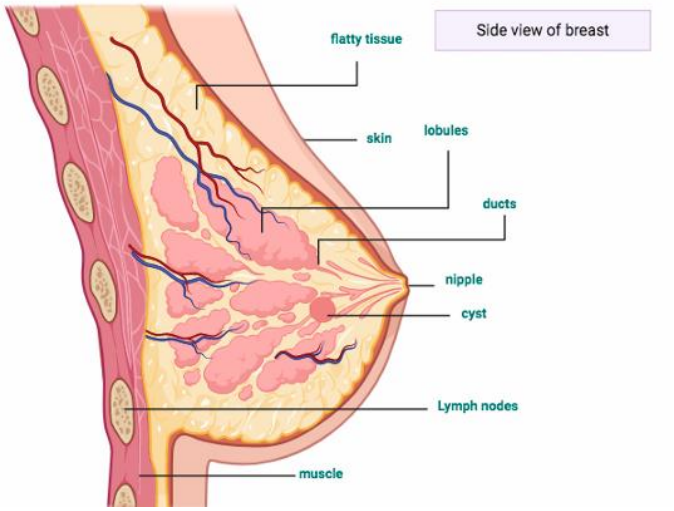

Figure 2.1 The anatomical structure of the female breast, highlighting the lobes—regions where epithelial tumors or cysts commonly develop (Jiménez-Gaona et al., 2020).

Early detection remains the most effective approach for improving breast cancer outcomes, with screening playing a crucial role in identifying the disease at an early, more treatable stage (Trimboli et al., 2020). Among the available screening methods, mammography has been widely validated as an effective tool, contributing to a reduction in breast cancer mortality rates by approximately 10% to 30% (Saadatmand et al., 2015). According to Trimboli et al. (2020), as an X-ray–based imaging technique, mammography is the most employed radiological method by healthcare providers for routine breast cancer.

The Breast Imaging Reporting and Data System (BI-RADS), developed by the American College of Radiology, serves as a standardised framework for interpreting and reporting findings from breast imaging modalities. This system aids in assessing the probability of malignancy and guides clinical management decisions. BIRADS categories range from 0 to 6, with each category indicating a different level of suspicion for cancer, thus facilitating communication among healthcare providers and ensuring consistent patient care (Barazi and Gunduru, 2023). By providing a uniform language for reporting, BI-RADS enhances diagnostic accuracy and supports informed clinical decision-making, thereby contributing to improved patient management and outcomes.

However, radiologists face significant challenges in distinguishing benign from malignant calcifications in mammograms, primarily due to the subtle differences in shape, size, and distribution of calcifications (Kim et al., 2018). Uncertain calcifications may be falsely identified as malignant, resulting in unnecessary biopsies, anxiety for patients, and increased healthcare costs. On the other hand, missed malignant calcifications may delay

critical treatment, adversely affecting patient outcomes (Mahmood et al., 2021). Kim et al. (2018) highlights that even experienced radiologists can struggle with this differentiation, especially in dense breast tissue, which can obscure calcifications.

Table 2.1 BI-RADS classification. Adopted from ARC BI-RADS Atlas.

Final Assessment Categories			
Category		Management	Likelihood of cancer
0	Need additional imaging or prior examinations	Recall for additional imaging and/or await prior examinations	n/a
1	Negative	Routine screening	Essentially 0%
2	Benign	Routine screening	Essentially 0%
3	Probably Benign	Short interval-follow-up (6 month) or continued	>0 % but ≤ 2%
4	Suspicious	Tissue diagnosis	4a. low suspicion for malignancy (>2% to ≤ 10%) 4b. moderate suspicion for malignancy (>10% to ≤ 50%) 4c. high suspicion for malignancy (>50% to <95%)
5	Highly suggestive of malignancy	Tissue diagnosis	≥95%
6	Known biopsy-proven	Surgical excision when clinical appropriate	n/a

2.2 Modern Diagnostic Tools in Medicine

The analysis of mammograms is essentially time-consuming and subjective, often leading to diagnostic inconsistencies between radiologists. Differences in training, experience, and individual judgment consequently lead to high false positive result, particularly in complex cases with subtle abnormalities. This variability introduces a risk of missed detections, both of which significantly impact patient care (Prodan et al., 2023). Subjectivity is a known challenge in radiology, especially in high stakes diagnoses like breast cancer. AI-based tools have been explored to address this issue, aiming to provide standardised interpretations that reduce the dependency on individual judgment. In a recent review, Prodan et al. (2023) claimed that AI trained on large, diverse datasets can

assist radiologists by providing consistent analyses, potentially reducing the number of missed diagnoses and unnecessary procedures.

Since the introduction of digital imaging and the establishment of the Digital Imaging and Communications in Medicine (DICOM) format as the industry standard, traditional computer-aided diagnosis (CAD) systems have primarily functioned as supportive tools in radiological diagnostics. These systems control domain knowledge and manually crafted rules to perform specific tasks, such as detecting, classifying, and segmenting medical images (Russell, 2021). In contrast, contemporary artificial intelligence (AI)-based CAD systems operate more autonomously, functioning as standalone diagnostic tools. Rather than relying on predefined rules, they utilise statistical methods, machine learning (ML) or deep learning (DL) techniques to execute tasks, eliminating the need for specific domain expertise (Russell, 2021).

AI represents the main concept of machines that own intelligence like or even exceeding human cognitive abilities. This enables them to perform complex tasks that require advanced perception and problem-solving skills, like those carried out by humans (Russell, 2021). Within the field of AI, ML and DL, equips machines with the ability to learn how to solve problems by analysing relevant data. This enables machines to develop highly refined and sophisticated representations based on the data they process.

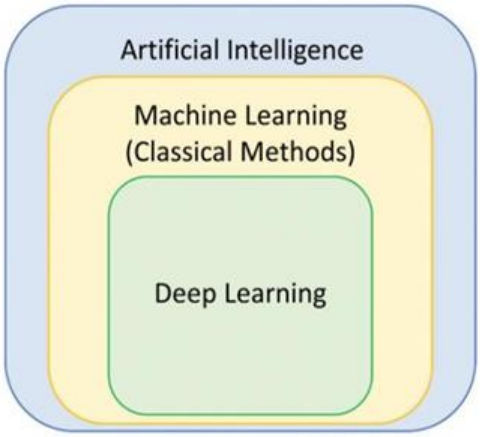

Figure 2.2 The relationship between AI, ML and DL (Prodan et al., 2023).

Binary classification has been widely adopted in the field of breast cancer detection due to its simplicity and effectiveness. By focusing on two distinct outcomes, typically benign versus malignant, it simplifies the classification task, making the results easier to interpret and the model performance more transparent (Lin et al., 2025). This approach is particularly beneficial in medical diagnostics, where clear decision-making is essential.

According to Lin et al. (2025), studies have shown that binary classification models can achieve high levels of accuracy and sensitivity, which is critical in minimising misclassification and ensuring timely treatment. For instance, feature-based machine learning algorithms have demonstrated strong performance in distinguishing between benign and malignant mammograms. The clarity provided by binary outcomes assists radiologists in making informed clinical decisions, especially in cases where early detection significantly improves patient prognosis.

In addition, binary classification helps streamline the use of resources by prioritizing patients who require immediate medical attention, contributing to more efficient healthcare delivery (Loizidou et al., 2023). Its straightforward structure also

allows for easier visualization and interpretation of decision boundaries, making it a useful starting point for both practical applications and further research (Lin et al., 2025). Furthermore, it requires relatively less computational effort compared to multi-class classification models, which adds to its practicality. Studies by Loizidou et al. (2023) mentioned that the binary classification framework also serves as a foundation for the development of more advanced systems, such as multi-class classification or longitudinal image analysis using sequential mammograms. Overall, binary classification presents a reliable, efficient, and interpretable method for supporting breast cancer diagnosis and enhancing early detection outcomes.

2.3 Image Preprocessing

Image preprocessing is a crucial step in breast cancer classification using machine learning, as medical images often contain noise, artifacts, or sensitive data that needs to be addressed before being processed by a CAD system (Grinet et al., 2024). Preprocessing enhances image quality by removing noise, which is essential for improving segmentation results (Al-Fahaidy et al., 2022). Various techniques are employed to improve the detection of microcalcifications, which are early indicators of breast cancer. Research studies have consistently demonstrated the effectiveness of these preprocessing methods in improving diagnostic accuracy and facilitating automated cancer detection (Murcia-Gómez et al., 2022)

One of the fundamental preprocessing techniques is noise reduction, which removes unwanted artifacts that can obscure essential features in medical images (Jiménez-Gaona et al., 2020). Median filtering is widely used to eliminate salt-and-pepper noise while

preserving important edges, ensuring better clarity in mammograms. Similarly, Gaussian filtering smooths textures and reduces random noise, making abnormalities such as microcalcifications more visible for analysis (Gómez-Flores and Pereira, 2023). These methods enhance image quality, allowing machine learning models to process clearer inputs for improved classification performance.

Different traditional methods such as Histogram equalization (HE), Adaptive Histogram Equalization (AHE) and Contrast limited adaptive Histogram Equalization (CLAHE) can be used to enhance the image (Jiménez-Gaona et al., 2020). Contrast enhancement techniques are crucial in medical imaging, as they improve the visibility of subtle abnormalities. HE redistributes pixel intensities, making variations in breast tissue more distinct. CLAHE further refines this process by adjusting local contrast and preventing over-amplification, which can distort image features. These contrast enhancement methods significantly aid in distinguishing between benign and malignant lesions, contributing to more precise diagnosis and classification (Abo-El-Rejal et al., 2024).

Grayscale conversion is an essential step in image preprocessing, particularly for medical imaging applications. Converting RGB images to grayscale simplifies processing and reduces computational complexity while preserving critical pixel variations that are essential for detecting tumors. The weighted sum method is commonly used to maintain intensity differences while removing unnecessary color information, ensuring that machine learning models focus on relevant features (Murcia-Gómez et al., 2022). This

transformation facilitates better segmentation and feature extraction for cancer classification.

Normalisation ensures consistency across images, allowing machine learning models to generalise effectively. Min-Max scaling normalises pixel values within a specified range, ensuring uniformity across datasets, while Z-score normalisation standardizes pixel intensities by adjusting them based on statistical measures such as mean and standard deviation. These normalisation methods mitigate discrepancies between images, reducing bias and improving classification accuracy (Gómez-Flores and Pereira, 2023).

Segmentation is the separation of region of interest (ROI) such as microcalcifications from the background of the image. For cancerous images, it is necessary to identify the lesion area and extract its relevant features for further analysis (Prodan et al., 2023). In traditional CAD systems, the tasks of specifying ROI such as initial boundary or lesions, are accomplished with the expertise of radiologists. In digital mammography, traditional segmentation methods are generally categorised into four main types which are threshold-based segmentation, region-based segmentation, pixel-based segmentation, and model-based segmentation (Jiménez-Gaona et al., 2020). Otsu's thresholding is an effective method that automatically determines an optimal threshold to differentiate foreground and background regions, making lesion identification more precise. (Abo-El-Rejal et al., 2024).

After segmentation, feature extraction and selection are the next steps to remove the irrelevant and redundant information of the data being processed. Features are

characteristics of the ROI taken from the shape and margin of lesions, masses, and calcifications. These features can be categorised into texture and morphologic features, descriptor, and model-based features, and help to discriminate benign and malignant lesions. Most of the texture features are calculated from the entire image or ROIs using the gray level value and the morphologic features focus on some local characteristics of the lesion (Jiménez-Gaona et al., 2020). Feature extraction plays a key role in breast cancer classification, enabling machine learning models to identify distinct characteristics of malignant and benign lesions. Edge detection techniques, such as Sobel and Canny filters, highlight tumor boundaries, aiding in precise classification (Murcia-Gómez et al., 2022).

Standardising image dimensions through resizing ensure compatibility with machine learning models, while cropping refining image inputs by focusing on specific regions containing potential abnormalities. These preprocessing adjustments improve classification accuracy by ensuring that models concentrate on relevant features without unnecessary distractions (Abo-El-Rejal et al., 2024). Preprocessed images undergo validation to confirm the effectiveness of applied techniques. Visualisation tools allow researchers to examine processed images, ensuring enhanced clarity and feature visibility. Comparing preprocessed images with raw images helps assess the improvements achieved through noise reduction, contrast enhancement, segmentation, and normalisation (Murcia-Gómez et al., 2022).

The study by Soliman et al. (2021) highlights that fuzzy image enhancement technique not only improves the visual quality of the images but also enhances the

performance of subsequent image analysis steps. The segmentation process, which utilizes Otsu's multiple thresholding method, achieves a high accuracy rate, with the segmented tumor region corresponding to 81% of the ground truth provided by an expert. This indicates that the proposed segmentation method is effective in accurately identifying tumor regions in mammograms. The performance of the proposed framework is evaluated using various metrics, including the Dice coefficient, Hausdorff distance, and Peak Signal-to-Noise Ratio (PSNR). These metrics confirm the framework's capability to deliver reliable and precise results in breast cancer detection.

The data preprocessing step involved addressing class imbalance using the Synthetic Minority Oversampling Technique (SMOTE**).** SMOTE is designed to tackle the issue of class imbalance, which is common in medical datasets where one class, for example malignant cases, may be significantly underrepresented compared to the other class, for example, benign cases (Kumari et al., 2020). This imbalance can lead to biased model performance. SMOTE generates synthetic samples for the minority class by interpolating between existing minority class instances. This means that instead of simply duplicating existing data points, SMOTE creates new, unique examples that are like the minority class instances. According to Kumari et al. (2020), for each instance in the minority class, SMOTE identifies its nearest neighbors and creates new instances by taking a weighted average of the features of the instance and its neighbors. This helps in enriching the dataset with more diverse examples of the minority class.

By applying SMOTE, the researchers were able to create a more balanced dataset, which is crucial for training machine learning models effectively. A balanced dataset helps

in improving the model's ability to learn from both classes equally, leading to better generalisation and performance. The study by Nuraeni et al. (2024) also combined SMOTE with Recursive Feature Elimination (RFE) for feature selection, ensuring that the models were trained on the most relevant features while also addressing class imbalance effectively.

2.4 Image Augmentation in Machine Learning

Data augmentation is a technique used to expand the size of a training dataset by adding variations to existing samples while preserving their class labels. This process helps improve a machine learning model's ability to recognize patterns and make accurate predictions (Al-Fahaidy et al., 2022). One key principle of data augmentation is state perturbation, where images are slightly altered to create new versions. By artificially increasing the diversity of training datasets, augmentation techniques can help address issues such as limited data availability and class imbalances, which are common in medical imaging (Arshad et al., 2023).

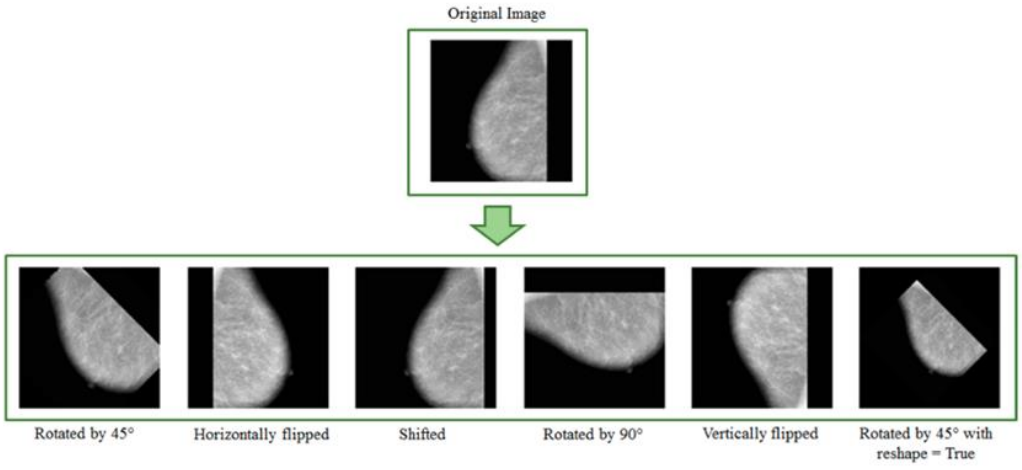

Figure 2.3 Example of images augmentation after applying geometric transformation (Oza et al., 2022).

In deep learning applications for computer vision, three common types of data augmentation exist which are dataset generation and expansion, on-the-fly data augmentation, and a combination of both methods (Oza et al., 2022). Since supervised DL models require large amounts of training data to develop strong inference capabilities, data augmentation is particularly valuable in cases where only a limited number of images are available. It involves applying random transformations such as rotation and flipping to generate new samples, which are then used during the training phase. While dataset generation and expansion can create numerous new images, these methods do not always enhance a model's ability to generalise to unseen data (Al-Fahaidy et al., 2022). On-the-fly data augmentation, also known as in-place augmentation, is another approach where image batches undergo random transformations during each training phase. This method introduces new variations to the model throughout the training process, allowing it to learn from a more diverse set of images (Oza et al., 2022).

The study by Huang et al. (2024) highlights the significant role of data augmentation in enhancing model performance, particularly in the context of calcification detection in mammograms. The proposed CalAttnMix method outperformed existing state-of-the-art (SOTA) augmentation techniques, achieving an increase in average recall by 3.40% and mean Average Precision (mAP) by 2.30% compared to the best results from other methods like Mosaic. The study emphasizes that many existing augmentation methods fail to ensure class balance, which is critical in medical imaging where certain conditions may be underrepresented. CalAttnMix addresses this issue by focusing on the minor class, thereby generating a more balanced dataset that enhances the model's ability to detect

calcifications accurately. This indicates that effective data augmentation can lead to better model performance in detecting calcifications.

This technique helps overcome the problem of limited labeled datasets, which is common in medical imaging. Similarly, another study utilized a sophisticated data augmentation process that leveraged data denoising, contrast enhancement, and GAN application, resulting in a significant improvement in classification accuracy ranging from 22.5% to 42.5% compared to traditional scans (Alawee et al., 2024). Different studies have employed various augmentation techniques. One study used rotation as a form of data augmentation to increase the size of the input data (Ragab et al., 2019). Another research incorporated advanced data preprocessing and augmentation techniques along with a cyclical learning rate strategy to enhance model performance, achieving an impressive accuracy rate of 99.68% (Al Moteri et al., 2024).

2.5 Machine Learning Models

Machine learning algorithms are automatic learning methods designed to learn from training data, identifying patterns and performing inference on novel data. These algorithms can be broadly categorised into supervised and unsupervised learning methods (Grinet et al., 2024). Supervised learning requires a labeled training dataset, while unsupervised learning can be trained without labeled data. Supervised machine learning methods can be further applied as homogeneous or heterogeneous ensemble techniques. Common machine learning algorithms used in breast cancer diagnosis include Decision Trees (DT), Naive Bayes (NB), Support Vector Machines (SVM), and Neural Networks (NN) (Yixuan et al., 2018).

Based on Yixuan and Zixuan (2018) research, it stated that although surpassed in performance by deep learning in recent years, traditional ML methods remain relevant, especially when paired with data augmentation techniques to overcome challenges like data imbalance and limited annotated datasets. Study by Loizidou et al. (2023), also mentioned that many traditional machine learning methods have been extensively studied and validated in various applications, including medical imaging. Their established nature provides a level of confidence in their performance and reliability.

SVM is a popular algorithm used to classify mammogram images into categories such as benign or malignant. It works by finding an optimal boundary that separates different classes as clearly as possible. In the scenario of nonlinearly separable data, the SVM can use a kernel function to transform the feature space into a higher-dimensional space (Islam et al., 2020).

The k-Nearest Neighbors (k-NN) algorithm is a simple, instance-based classifier that predicts the class of a new image based on the majority vote of its k closest samples in the training set. Although easy to implement, k-NN is sensitive to noise and unbalanced datasets, which makes augmentation a helpful technique to enhance its performance (ElOuassif et al., 2021).

Random Forest (RF) is an ensemble learning method composed of multiple decision trees, offering robustness against overfitting and noise. It is particularly effective when the dataset is enhanced with augmented images or features, as it can capture complex relationships between inputs (Nadarajan and Sulaiman, 2021).

Logistic Regression (LR) is a supervised learning algorithm designed primarily for binary classification tasks. It models the probability that a given input belongs to a particular class, using a logistic function to map predicted values between 0 and 1 (Grinet et al., 2024). In medical image analysis, LR is valued for its simplicity, ease of interpretation, and effectiveness on linearly separable data.

NB classifiers use probability models to predict the class of input data based on the assumption of feature independence. Though often less accurate than other methods for complex imaging tasks, its performance can be improved with balanced feature representation through augmentation (Nadarajan and Sulaiman, 2021).

Decision Trees (DT) are hierarchical models that split data based on feature thresholds to create a tree-like structure. Each internal node represents a test condition, and each terminal node assigns a class label. Due to their intuitive structure and interpretability, decision trees are widely used in medical diagnostics (ElOuassif et al., 2021).

NN are computational models inspired by the human brain's architecture. Even before the advent of deep learning, shallow neural networks were widely used in medical applications due to their capacity to model non-linear relationships between features (Grinet et al., 2024). These networks consist of input, hidden, and output layers, and use activation functions to learn complex patterns within the data.

According to Nuraeni et al. (2024) study, the SVM algorithm consistently outperformed the DT across all metrics, achieving the highest accuracy of 96.64%. This indicates that SVM is more reliable and effective for breast cancer classification compared

to DT. The findings suggest that SVM could be a promising tool for early detection and treatment of breast cancer, potentially leading to improved patient outcomes. The findings from Alobaid and Bonny (2024) highlights the critical importance of selecting appropriate AI models to enhance the accuracy of breast cancer diagnosis. Comparative analysis underscores that while many models can be effective, some significantly outperform others, which can have substantial implications for early detection and treatment.

Ensemble learning methods have become essential in improving the accuracy and reliability of breast cancer classification systems. These techniques involve combining the outputs of multiple classifiers to create a single, more accurate predictive model. By integrating the strengths of various algorithms, ensemble methods reduce errors, minimise overfitting, and enhance model generalisation (Al-masni et al., 2018). In breast cancer detection, where precision is critical, ensemble models are particularly beneficial.

One of the popular ensemble technique is called Bagging, short for Bootstrap Aggregating. This method works by creating many versions of the original dataset using random sampling. Each version is used to train a separate model, and the final prediction is made by averaging the results for regression tasks, or by taking a majority vote for classification tasks. Bagging helps make predictions more stable and less prone to overfitting, meaning the model performs better on new data (Wang et al., 2025).

Another method, Boosting, builds a strong model by training a series of smaller, weaker models one after another. Each new model focuses on correcting the mistakes made by the previous one. By doing this, the overall system becomes more accurate over time and is especially effective at handling difficult cases (Wang et al., 2025).

Lastly, Voting is a straightforward method that combines predictions from several models. In hard voting, each model gives a vote, and the most common prediction is chosen. In soft voting, each model provides a probability for each class, and the final prediction is based on the average of those probabilities. Sometimes, certain models can be given more weight than others to improve the overall outcome (Cao-Van et al., 2024).

Study by Cao-Van et al. (2024) stated that, each ensemble learning method has its own strengths and weaknesses. Bagging helps reduce prediction variance and improves accuracy by combining several models, but it does not address model bias and can be computationally expensive. Boosting focuses on reducing bias by giving more attention to difficult cases, which improves accuracy, but it can be overfit on noisy data and takes longer to train. Stacking combines different models using a meta-learner to reduce both bias and variance, offering strong performance, though it is complex and resource intensive. Voting is simple to implement and improves stability by averaging multiple models, but it cannot fix bias if the base models are already biased. Overall, ensemble learning is a powerful technique in disease prediction, helping support early diagnosis and reducing health-related risks

2.6 Evaluation Metrics and Validation Methods

Accurate assessment of machine learning models is crucial in breast cancer classification to ensure reliable detection of malignant and benign calcifications. Various evaluation metrics and validation strategies are utilized to measure model effectiveness, optimise classification accuracy, and validate the impact of augmentation techniques. Given the range of tasks, including classification, detection, and segmentation, a model's

performance can be assessed using a number of measures (Grinet et al., 2024). The most widely used metrics for evaluating the effectiveness of classification techniques are accuracy, precision, recall, specificity, false-positive rate (FPR), and F1 score (Yixuan and Zixuan, 2018).

A confusion matrix is a fundamental evaluation tool used in classification problems to assess the performance of a predictive model by comparing the actual class labels with those predicted by the model (Hernández-Del-Toro et al., 2021). These metrics are determined by quantifying the true positive (TP), true negative (TN), false positive (FP), and false negative (FN) samples in the dataset. True positive rate (TPR) and sensitivity are other names for recall. These measurements are provided by:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall\ (Sensitivity) = \frac{TP}{TP + FN} \tag{3}$$

$$Specificity = \frac{TN}{TN + FP} \tag{4}$$

$$F1\ score = \frac{2\ \times Precision\ \times Recall}{Precision + Recall} \tag{5}$$

$$FPR = \frac{FP}{FP + TN} \tag{6}$$

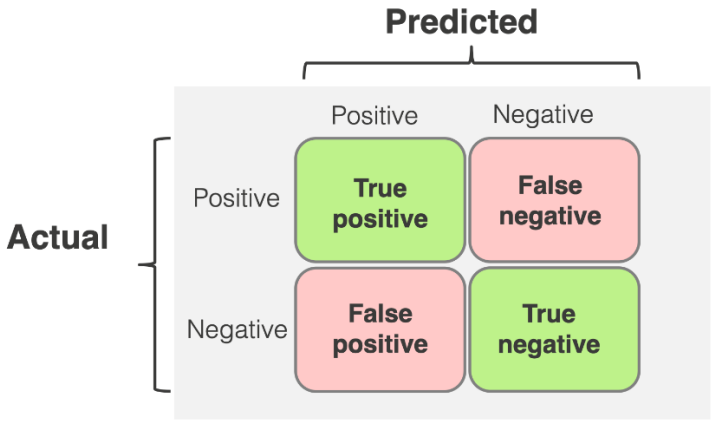

Figure 2.4 Concept of confusion matrix (Shanbehzadeh et al., 2022)

Accuracy is the most reported metric in breast cancer classification studies. It measures the proportion of true results, both true positives and true negatives, among the total number of cases examined. However, accuracy alone may not provide a complete picture of model performance, especially in imbalanced datasets (Guo et al., 2024). Area Under the Curve (AUC) is another frequently reported metric, which evaluates the model's ability to distinguish between classes. It represents the probability that a randomly chosen positive instance is ranked higher than a randomly chosen negative instance. This metric is particularly useful in binary classification tasks (Yixuan and Zixuan, 2018).

According to Grinet et al. (2024), while less than half of the studies reported precision or recall, these metrics are essential for understanding the model's performance in identifying positive cases. Precision measures the accuracy of positive predictions, while recall or sensitivity assesses the model's ability to identify all relevant instances. In addition to the Intersection over Union (IoU) of the detection model's bounding boxes, these metrics are used to evaluate the performance of detection approaches. The IoU, also known as the Jaccard similarity index (JSI), is a key statistic for evaluating segmentation algorithms based on the relationship between TP, FP, and FN. It is stated as follows: