UNIVERSITY OF LONDON
INTERNATIONAL PROGRAMMES

BSc Computer Science and Related Subjects

CM3070 PROJECT
PRELIMINARY PROJECT REPORT

Deep Learning Breast Cancer Detection

Author: LEE YONG HUA

Student Number: 10246454

Date of Submission: 22th September 2025

Supervisor: Dr Liu Jun Hua

Table of Contents

Table of Figures

Table of Tables

1.Introduction

This project adopts the Final Year Project Template Machine Learning and Neural Network 3.2 Deep Learning Breast Cancer Detection to establish a Deep Learning assisted X-ray mammography model that can improve the accuracy of breast cancer screening. Convolutionl Neural Network architecture will be implemented in this project to invent a deep learning model to analysis mammogram images and reduce the workloads of the radiologists.

1.1 Breast Cancer

Breast cancer is one of common diagnosed cancers among women worldwide and has emerged as a leading cause of cancer related deaths. It poses a significant global health burden, affecting woman of all ages and across all regions. According to the World Health Organization in 2022, approximately 2.3 million women diagnosed with breast cancer and resulting in over 67000 deaths globally. According to American Cancer Society(Lehman,C.D et al.,2015)[8], breast cancer accounts for nearly 19% of all cancer cases and 30% of all cancers among women, making it major concern in global oncology. Alarmingly, the incidence rate of bre

ast cancer has been increasing steadily by about 0.5% annually since the mid-2000s, highlighting the ongoing need to improved screening and early detection strategies. Breasts cancer occurs in every country of the world in women at any age after puberty, regardless of economic status or geography. Importantly, studies have consistently shown that survival rates are closely linked to the stage at which cancer is diagnosed with early-stage detection offering significantly higher chances of successful treatment and long-term survival. This emphasize the critical need for accurate and accessible diagnostic tools that can detect breast cancer at the earlier stage.

1.2 Mammography

X-ray mammography remain as the gold standard for breast cancer screening. Mammogram is an X-ray technique used for imaging breast as it can detect breast lesions before they become palpable. Screening mammograms aim to identify breast cancer at the earlier stages of the disease. Earlier detection of breast cancer increased the likelihood of a cancer being successfully being treated and allows with a better treatment option. However, interpretation of mammograms is highly dependent on radiologists' expertise and can be prone to subjectivity, fatigue and human error. Multiple studies have shown that 20-30% of diagnosed cancers could be found retrospectively on the previous negative screening exam by blinded reviewers. According to the studies done by Breast Cancer Surveillance(Radiology Business 2025)[9] shown that the screening performance of mammogram with at a sensitivity of 85% and specific of 89% respectively. These metrics describe the average performance of reviewers of mammograms, and there is substantial variance in the performance of individual, with reported false positive rates between 1-29% and sensitivities between 29-97% which is a huge gap.

1.3 Aim

Aim of this machine learning project is modelling the Digital Database of Screening Mammography(DDSM) with convolutional neural networks. Advancement of Artificial Intelligence(AI) in medical imaging, particularly design and train an AI model to process and learn from large volumes of image data. Convolutional Neural Networks(CNN) is a class of deep learning models architecture specifically designed for image recognition, have demonstrated impressive performance in numerous visual tasks. Besides that, this deep learning project also aim to develop a multi class classification system that can categorize mammograms into four different categories which are Normal, Benign, Benign Without Callback, Malignant.

According to the UK national health service (NHS) launched, in early 2025, a major research project to address that can deep learning assisted X-ray mammography can improve the accuracy of breast cancer screening and replace the manpower of radiologists in mammograms analysis. Early finding from this initiative, as well as landmark studies such as the one published in Nature by Google Health(McKinney S.M. 2020)[3], demonstrate that AI system not only match but can outperform human radiologists in specific diagnostic scenarios.

1.4 Goal

Ultimately, the primary goal of this project is to develop a robust, accurate and interpretable CNN based classification system tailored for breast cancer detection using mammography images. This system is intended to serve as a powerful clinical decision support tool for radiologists, enhancing diagnostic precision while minimizing the risk of human error that may rise from fatigue of continuous mammography analysis. By automating and optimizing the classification of mammograms into different categories, the system aim to streamline the diagnostic workflow and improve the efficiency of breast cancer detection programs.

Besides that, the project emphasizes the importance of model interpretability and clinical trust. To this end, explainability tools such as Grad-CAM will be integrated to visualize and highlight the areas of mammograms that most influenced the model's prediction. This dual focus on performance and transparency ensures that the model not only deliver reliable results but also align with the standards and expectation of real-world clinical practice. This project seeks to make a meaningful contribution to the field of AI-drive medical diagnostics. The ultimate vision is to pave the way for more scalable, trustworthy, and impactful deep learning solutions in healthcare that empower clinicians, enhance patient outcomes, and support the global effort to detect and treat breast cancer at the earliest possible stage.

1.5 Scope

This project aim to develop a deep learning model based on CNNs for multi-class classification of mammogram images utilizing the publicly available Digital Database for Screening Mammography(DDSM). The dataset comprises 2D mammogram images labelled as Normal, Benign, Benign Without Callback, and Malignant cases. This approach addresses a more nuanced classification challenge compared to traditional binary cancer detection. The methodology includes essential preprocessing steps such as image conversion, normalization, resizing, and data augmentation to enhance model robustness and generalization. This study is limited to offline training and evaluation of CNN architectures on existing datasets and does not extend to real-time clinical deployment, integration into target users workflow. Future enhancements could involve incorporating advanced imaging modalities like 3D mammography and tomosynthesis. Furthermore, the project depends on the accuracy of the annotated labels provided by the DDSM dataset and does not include independent clinical verification of the data. ***(978 words)***

2.Literature Review

2.1 What is Breast Cancer

Breast cancer is a complex and life-threatening disease characterized by the uncontrolled growth of abnormal cells in the breast tissues, which can lead to the formation or development of tumours. If the situation left undetected or untreated, these malignant cells could invade surrounding tissues and spread to other parts of the body, significantly worsening prognosis. Early detection remains critical for improving treatment outcomes and long-term survival rates.

Globally, breast cancer is one of the most frequently diagnosed cancers among women, accounting for a substantial public health burden. However, incidence rates very markedly by geographic region, ethnicity, and socioeconomic status. For example, studies have shown higher incidence rates in developed countries, such as Western Europe and United States, compared to lower rates in many developing regions, potentially reflecting differences in lifestyle factors, reproductive behaviours, genetic predispositions, and access to screening programs(Brayet al., 2018)[21]. Additionally, survival rates tend to be lower in low-income countries, where late-stage diagnosis and limited treatment availability prevail.

According to the research from World Health Organization (WHO 2024)[1], in 2022 there were 2.3 million women diagnosed with breast cancer and the disease responsible for an estimate 670000 deaths worldwide. This make breast cancer became the second most diagnosed cancer worldwide, following by lung and skin cancers.

Importance of early detection stage of breast cancer cannot be overstated. According to the study done by American Cancer Society(Hu,Y.(n.d.)2024)[17], identifying breast cancer at an early stage of breast cancer will increase the chances of survival for patients and more treatment options can be provided. Mammogram has become the most widely used breast cancer screening method for the early detection of breast cancer, owing to its effectiveness in identifying tumours before clinical symptoms become apparent. It had been proven to reduce the mortality rate of breast cancer patients by allowing healthcare providers to detect tumours at a smaller size and earlier stage.

Some critics also argue that benefits of widespread mammography screening must be carefully balanced against risks of overdiagnosis and overtreatment, which can impact patients' quality of life (Welch & Passow, 2014)[22]. These debates highlight the importance of developing more accurate, accessible, and less invasive screening technologies.

In summary, breast cancer poses a major part in the global health challenge. With a widespread awareness of early screening programs, the impact of the disease can be significantly reduced.

2.2 Screening of Breast Cancer

Screening of breast cancer typically involves two primary imaging modalities, mammograms and ultrasound. These techniques serve complementary purposes in the early detection and diagnosis of breast abnormalities.

Mammogram is a low dose x-rays screening techniques to detect changes and abnormalities of the breast tissue, such as tiny white spots or masses before they become palpable. It is widely regarded as the first line of defence in breast cancer screening due to its proven ability in early-stage tumours detection including ductal carcinoma in situ (DCIS) classified as stage 0 breast cancer. DCIS represents non-invasive cancerous cells confined within the milk ducts, which mammography detects effectively through visualization of microcalcifications. If abnormalities are detected on a mammogram, the doctor may refer the patient for additional screening, such as ultrasound, to obtain more detailed imaging for accurate diagnosis.

Breast ultrasound screening utilizes high frequency sound waves to examine the tissues in the breast and evaluating the nature of the lump detected, helping to determine whether a mass is solid or fluid filled. Ultrasound is beneficial for women with dense breast tissue and detect small tumours hiding in the breast tissue when mammogram became ineffective on the screening of breast cancer. Mammogram may be less effective due to overlapping of fibroglandular structures.

According to the article “Mammogram vs breast Ultrasound: What’s the difference?” (Ethan Cohen 2024)[2], ultrasound is not recommended as replacement of mammogram, as ultrasound is less effective at detecting stage 0 cancer which is early stage of breast cancer. This is when the breast cancer tissue is inactive. Statistic shows that mammogram detect approximately 4 to 6 cases of breast cancer per 1000 women, while ultrasound identifies only 1 to 2 additional cases beyond that. Therefore, ultrasound is best utilized as a complementary tool to mammography.

Therefore, while ultrasound enhances diagnostic precision, it is not sufficient as the most precise screening method. Instead, ultrasound plays a valuable complementary role alongside mammography, enabling a more comprehensive cancer detection. This combination of screening methods improve diagnostic accuracy and ensures that suspicious findings investigating, which contribute to earlier and effective diagnostic plan.

2.3 Emergence of Artificial Intelligence(AI) in Breast Cancer Screening

The integration of AI, deep learning into breast cancer screening has strong potential in enhancing diagnostic accuracy and improving clinical workflows. Deep learning into breast cancer screening holds significant promise for addressing some limitations inherent in traditional imaging modalities such as mammography and ultrasound. While mammography remains as gold standard for early detection, challenges such as reduced sensitivity in dense breasts and the potential for false positives and negatives persist. AI-driven systems have the potential to enhance diagnostic accuracy by improving image interpretation and reducing human error, thereby streamlining clinical workflows and alleviating radiologist workload.

Several studies have demonstrated that AI-assisted mammography can match or even surpass the performance of expert radiologists under certain conditions. According to article "International Evaluation of an AI System for Breast Cancer Screening" by Nature(McKinney S.M. 2020)[3] a deep learning model achieved an area under the receiver operating characteristics curve (AUC-ROC) that surpassed of all human readers by an absolute margin of 11.5%. This indicates a significantly higher-level of performance in distinguish cancerous from non-cancerous cases. In mammography, deep learning models have shown promise in reducing false positive and false negative. A landmark study by Nature[3] also shows that Convolutional Neutral Network based model trained on UK and US mammograms outperformed radiologists in certain cases. Despites these advancements, current commercial AI systems are generally limited to binary classification and are not designed to support multi-class diagnostic tasks. This presents a significant opportunity for further research and innovation in building a fine-grained classification models for breast cancer screening.

Despite these encouraging results, current commercial AI systems predominantly perform binary classification to differentiate between cancerous and non-cancerous findings. This limitation contrasts with the clinical reality, where different conditions of breast abnormalities. Consequently, there is a compelling need for more nuanced, multi-class classification models capable of supporting finer diagnostic distinctions, improving clinical decision-making, and potentially reducing unnecessary biopsies and anxiety.

In summary, AI has demonstrated impressive capability in augmenting breast cancer screening particularly in overcoming some limitations of mammography and ultrasound.

However, the current focus on binary classification models leaves a critical gap that this project aims to address by developing a multi-class CNN classifier tailored to distinguish between different conditions of breast abnormalities cases. Such fine-grained classification aligns more closely with clinical diagnostic needs, enhancing the potential for AI to improve screening accuracy and patient outcomes comprehensively.

2.4 Convolutional Neutral Network in Medical Image Analysis

Convolutional Neural Networks(CNN) are a type of deep learning algorithm that have become essential in the analysis of visual data, particularly tasks involving images and videos. Unlike traditional machine learning methods that rely on manual feature extraction, CNNs automatically learn to identify relevant patterns, textures and structures within images by convolutional filters. This ability to autonomously extract both low-level and high-level features makes CNNs exceptionally powerful and efficient for image understanding and classification tasks. According to the article “Convolutional Neural Network in Medical Image Understanding” by D.R Sarvamangale[5], CNN has played a key role in detecting COVID-19 using chest X-rays, helping accelerate diagnosis process and support timely treatment. This success highlights their potential for boarder applications in healthcare.

Popular CNN architecture such as VCG16(Debnath et al.,2024)[19], ResNet50(He et al., 2015)[15], and DenseNet121(Hasan, N, et al.,2021)[20] are commonly used in medical due to their strong performance in learning detailed features and capturing complex visual patterns. These models have been successfully adapted for various clinical applications, including tumour detection, organ segmentation and disease classification. Recent CNNs architectures, including DenseNet and EfficientNet have further improved features extraction capabilities and generalization across diverse medical imaging tasks. Transfer learning using pre-trained models has become a popular approach, particularly effective for limited medical datasets were training from scratch is challenging. Therefore, this project incorporates a custom-built CNN architecture designed to balance model complexity with the size of dataset and available computational resources, while comparing its performance against established pre-trained models.

2.5 Digital Database for Screening Mammography and Classification

The Digital Database for Screening Mammography(DDSM) is a resource for use by the mammography images analysis research community. DDSM is a publicly available dataset that offers labelled mammogram images, including abnormality types, pathology and lesion characteristics. According to the article "The Digital Database for Screening Mammography" (Kevin Bowyer et al., 1998) [4], the primary purpose of DDSM is to provide researchers with a standardized, large dataset of digital mammograms for developing and evaluating computer-aided algorithms used in breast cancer screening. DDSM includes a mix of normal, benign, and malignant cases that allow researchers to develop algorithms that are robust to different types of breast tissue and abnormalities.

A normal mammogram images the breast tissues typically appears in shades of grey and white. The fat appearing in the mammogram image will appear as darker grey and dense tissue like glands and ducts appearing as lighter grey and white shown in the diagram below.

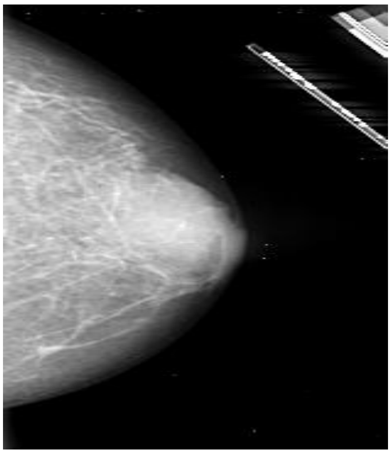


Figure 1 Normal Mammogram Case

In benign breast cases, mammogram images typically reveal well-defined, round or oval shaped masses which is a tumour. These lesions are usually circumscribed with clear, smooth borders and maintain a regular geometric shape, distinguished them from malignant growths.

These characteristics indicating that they are non-cancerous tumour that grows slowly or may stop growing entirely. However, despite their benign nature, these findings may still warrant additional imaging such as ultrasound to confirm the diagnosis and ensure accurate assessment. As shown in the diagram below, the red-circled area highlights the distinct lump with a uniform and orderly appearance.

Figure 2 Benign Mammogram Case

Benign without callback cases refer to a benign case is identified on a screening mammogram, but the features are so clearly benign that not require the patient to return for further imaging or evaluation. Characteristics such as tumour circumscribed with clear, smooth borders and maintain a regular geometric shape are well and clearly defined as shown red-highlight circle in the diagram below. The tumour in the highlighted circle are clearer and brighter compared to the circle highlighted on the benign case diagram above.

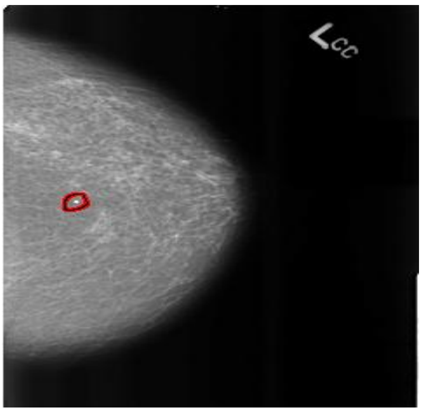


Figure 3 Benign Without Callback Mammogram Case

In malignant cases, a mammogram typically reveal a cancerous tumour characterized by abnormal and irregular cell growth, These tumours often appears as high density that represent as bright white areas on the mammogram. Additionally, a spiculated appearance that characterized by a thin line radiating outward the mass that serve as a strong indicator of malignancy. It will suggest the cancerous tissue that is invading the surrounding breast tissue. Diagram below shown a good example of malignant cases mammogram image with a white and irregular shape of lump in the red-highlighted circle.

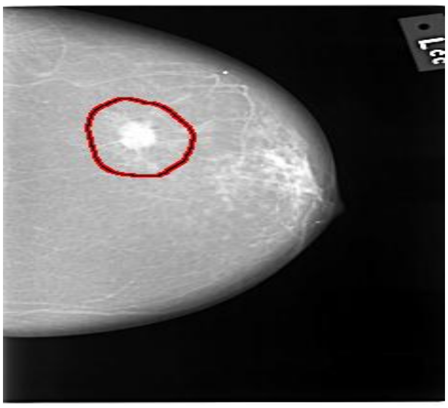


Figure 4 Malignant Mammogram Case

DDSM provide valuable resources for developing AI model, they have limitations such as relatively small sample sizes, potential annotation errors, and lack of diversity in patient demographics. These may affect model generalizability and real-world applicability.

2.6 Existing Approaches or Competitors

According to research on CNN for mammograms analysis on images analysis had focused on binary classification which are cancer and non-cancer. “Automatic Mass Detection in Mammograms using Deep CNN” (Kooi et al,.2019)[6] shows that achieved strong results using CNN for mass detection and “Automatic Breast and Fibroglandular Tissue Segmentation in Breast MRI” (Acad Radiol 2022)[7] using U-net architecture to implement and test independent validation dataset to quantify breast tissues volume in Breast MRI with binary classification. 4 class system for DDSM database is limited in the market to match the diagnosis stage, as it can help the doctors or radiologist to analyse and classify MMO in a more convincing and faster way.

2.7 Justification for Proposed Approach

This project proposed a four class CNN classifier trained on DDSM images to predict Normal, Benign, Benign without callback and Malignant cases. The goal is to move beyond traditional binary classification and offer a more clinically relevant and fined-grained diagnostic tool. To achieve high performance while maintaining computational efficiency. The mammogram analysis tool will serve as an interactive platform to process mammogram images, leveraging a CNN model to predict and classify them into their respective diagnostic categories. The project aim to modelling DDSM with CNN to improve classification system, diagnostic accuracy and reduce manpower for MMO analyse.

2.8 Summary and Research Gap

As a summary, this literature confirms the value of CNNs architecture in breast cancer detection. The existing literature underscores the significant of impact on CNNs in advancing breast cancer detection, it also reveals several important shortcomings. A majority of current CNNs deep learning system on medical detecting purpose focus solely on binary classification which distinguishes between cancerous and non-cancerous finding but without addressing the boarder clinical spectrum of the breast condition. Additionally, many of these models function as “black box” which offering little no insight into how prediction are made and hidden their acceptance in clinical environments where trust and transparency are crucial. To fill up these gaps, this deep learning project introduces a more nuanced and clinically meaningful 4-class classification that distinguishes between normal, benign, benign-without-callback and malignant cases. ***(2237 Words)***

3. Project Design

3.1 Domain and Users

Domain

The main objective of this project is to implement deep learning techniques in mammogram images analysing by learning the features of various breast cancer cases according to mammogram images data from DDSM. The focus of this project is on the application of deep learning techniques and lies at the intersection of medical imaging and AI-assisted healthcare diagnostics, addressing breast cancer screening. Specifically, it focuses on applying CNNs to enhance breast cancer through mammography.

This domain address a critical area of healthcare where accurate and timely diagnosis can significantly impact a patient’s survival rate and the effectiveness of treatment. By introducing an AI-powered classification system trained on real-world mammogram data, the project aims to support radiologist in their diagnostic tasks, reduce human error, and enhance the overall accuracy and consistency of breast cancer screening. Lastly, the project will have a critical domain where early detection can save lives.

Target Users

The primary target user of this project are radiologists, who play a vital role in interpreting mammograms images and diagnosing breast cancer within clinical environments. Radiologists possess specialized expertise in medical imaging and serve as the direct end-users of the CNN-based classification system developed in this project. These professionals face high workloads and the risk of diagnostic fatigue, which can lead to errors in interpretation The system is specially designed to assist radiologists by automating routine image classification tasks and highlighting suspicious regions within mammograms using explainability techniques such as Grad-Cam. This support aim to improve diagnostic accuracy and efficiency, reducing the cognitive burden on radiologists and allowing them to focus more complex and ambiguous cases.

In addition, radiologists, oncologists and primary care physicians who rely on diagnostic imaging to guide treatment decisions and monitor patients outcome. These healthcare professionals can leverage the system output to enhance clinical judgement and support coordinated patient care. By addressing these key user groups, the project reinforces its relevance and potential impact across various levels of the healthcare ecosystem.

3.2 Needs and Requirements

Needs

Development of this deep learning breast screening project using mammography with CNNs is driven by several critical needs within the domain of medical imaging, artificial intelligence, and healthcare diagnostics. Firstly, a growing demand for accurate classification of mammogram images and minimize the risk of human error. Traditional diagnosis relies heavily on radiologists manually interpreting scans, which can be subjective and inconsistent, especially when dealing with early stage or complex breast abnormalities. Unfortunately, a study conducted by University of Sydney (Jeffrey Gole et al., 2020)[12] found that radiologist miss between 10% to 30% of breast cancer cases. Moreover, among of the women recalled due to these diagnostic errors, 80% ultimately receive normal results, rather than a benign or malignant results. A deep learning model trained on distinguish between four different classes of breast cancer cases can offer precise, consistent and standardized classification outcome. An CNNs based classification system can act as a second reader, flagging suspicious images and reinforcing diagnostic accuracy.

Secondly, the project addresses the need to reduce the workload of radiologists. With increasing volumes of medical imaging to be reviewed daily, radiologists are often exhaust with works which can lead to diagnostic fatigue. According to the articles from Medality[10] and UChicago(Ekpo, E.U. et al., 2018)[11] there are 2700 radiologists over 108 countries, each radiologist need to handle an average of 4.7 reports in their daily practice and 62-70% of the reading are mammograms. Given these high workloads, with large volumes of mammograms to review daily, which can lead to fatigue and increased the risk of diagnostic errors. By providing a CNNs based classification system capable of accurately identifying and categorizing mammographic findings into different cases of breast cancer. The AI classification system can serve as supportive tool for radiologists to handle routine evaluations and allowing radiologists to focus on more complex cases.

Lastly, the crucial goal that supported by the project is early detection of breast cancer. Detecting the breast cancer at an early stage significantly improves survival rates and expands treatment options for the patients. According to the journal by Nature Communications[13], the survival rate of cancer increase significantly when the disease is detected at an early stage. The research highlights a five-years survival rate of approximately 91% for early-stage detection, compared to 26% when the cancer is diagnosed at late stage. By automating the analysis of mammograms and highlighting potential cases for further review, the system can help to ensure that suspicious findings are identified and acted upon promptly.

Functional Requirements

Functional requirements of this project are designed to ensure the CNNs based classification system can effectively support radiologists in mammogram analysis and classification. Firstly, the system needs to include image input capability by allowing it to accept and process mammogram images in standard forms such as DICOM which is the format from DDSM and PNG the normal format of regular mammogram images. This is to ensure the compatibility with existing clinical imaging systems and facilitates seamless integration into real-world healthcare and workflows. Secondly, the system should perform class prediction to accurately classifying each input images into one of the four diagnostic categories according to the type of cases from DDSM. This multi-class classification capability provides more granular and reliable information for the patients. Lastly, to enhance the transparency and clinical trust, the system should introduce heatmap visualization by using Grad-Cam which can highlight the specific region within the mammogram that shows the characteristics of the specific breast cancer cases. These functional requirements help us to form a robust and trustable deep learning diagnostic tool that are capable in breast cancer screening.

Non-Functional Requirements

Non-functional requirements of this project are essential to ensure the deep learning system performs reliably and efficiently in practical clinical settings. One of the key requirements is high accuracy as the system must consistently deliver correct classification across four breast cancer categories. This is crucial for minimizing false positives and negatives of the results which can lead to misdiagnosis or delay treatment. Besides that, the system must be robust to noise and variations in input images of mammograms. The reason of the images preprocess is the differ quality of images might be due to factors such as imaging equipment, patient positioning or image artifacts that is not appropriate. Therefore, the model should be trained and tested to handle such variability without significant drops in performance. Finally, the system should support fast inference, which means it must generate predictions in a fast and accurate environment. This can improve the time efficiency and workflows of the clinical or hospital and impact on patient care and diagnostic. Non-functional requirements ensure the model is not only accurate and reliable but also practical and scalable for real-world implementation.

3.3 Overall Structure

Overall structure of this deep learning breast cancer detection project using mammograms can be divided into six main phases as shown in the diagram below.

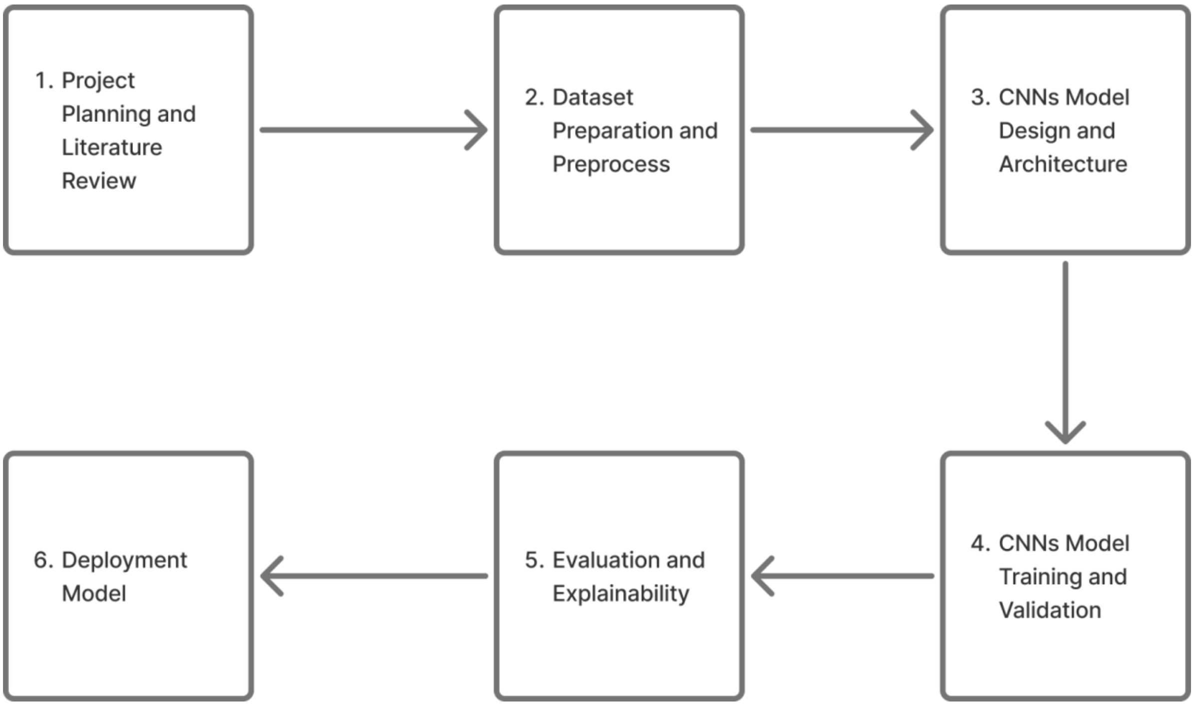


Figure 5 Project Structure Flow

The structure of the project follows a systematic and research-driven approach which divided into six key phases to ensure both academic depth with high volume of datasets and real-life applicability. The project begins with planning and literature reviews to define theory of breast cancer, screening of breast cancer, CNN application in medical imaging sector and limitations on existing models of breast cancer screening. Secondly, at dataset preparation phase ensures mammogram images from DDSM database are collected, pre-processed and labelled into four different cases. Image resizing, data normalization, and data augmentation are involved in this phase to enhance model generalization and robustness.

Third phase of the project is CNNs model architecture design which includes a custom build CNNs architecture to fit the 4-class classification to compared with the pre-trained CNNs architecture such as DenseNet121 and ResNet50. After that, training and validation phase involves feeding the data into the self-designed model, optimizing its parameters, and evaluating performance using metrics such as accuracy and sensitivity rate of the models. In the evaluation phase of the project, Grad-Cam is implemented to generate heatmap that can visualize the prediction of the specific class of breast cancer by using the custom-built CNNs model. Finally, deployment and documentation phase involves saving the trained CNNs model, design and build a simple demo interface, and completing the final report with design decisions and results. This structure workflow ensures the project achieve its goal of improving diagnostic accuracy, reducing radiologist workload, and contribution in AI tool for breast cancer screening.

3.4 Design Specification

Custom Built CNNs Architecture

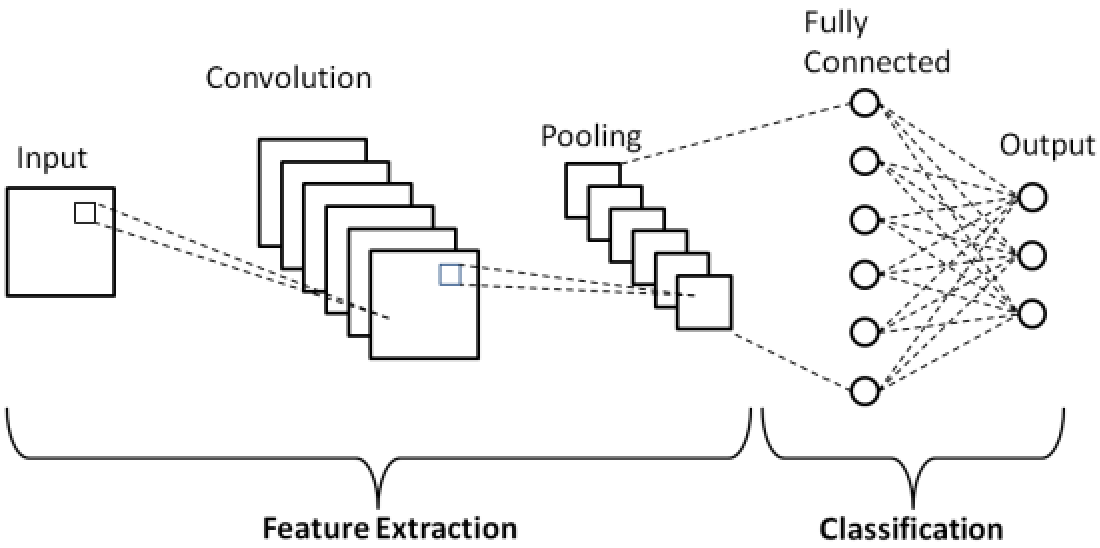


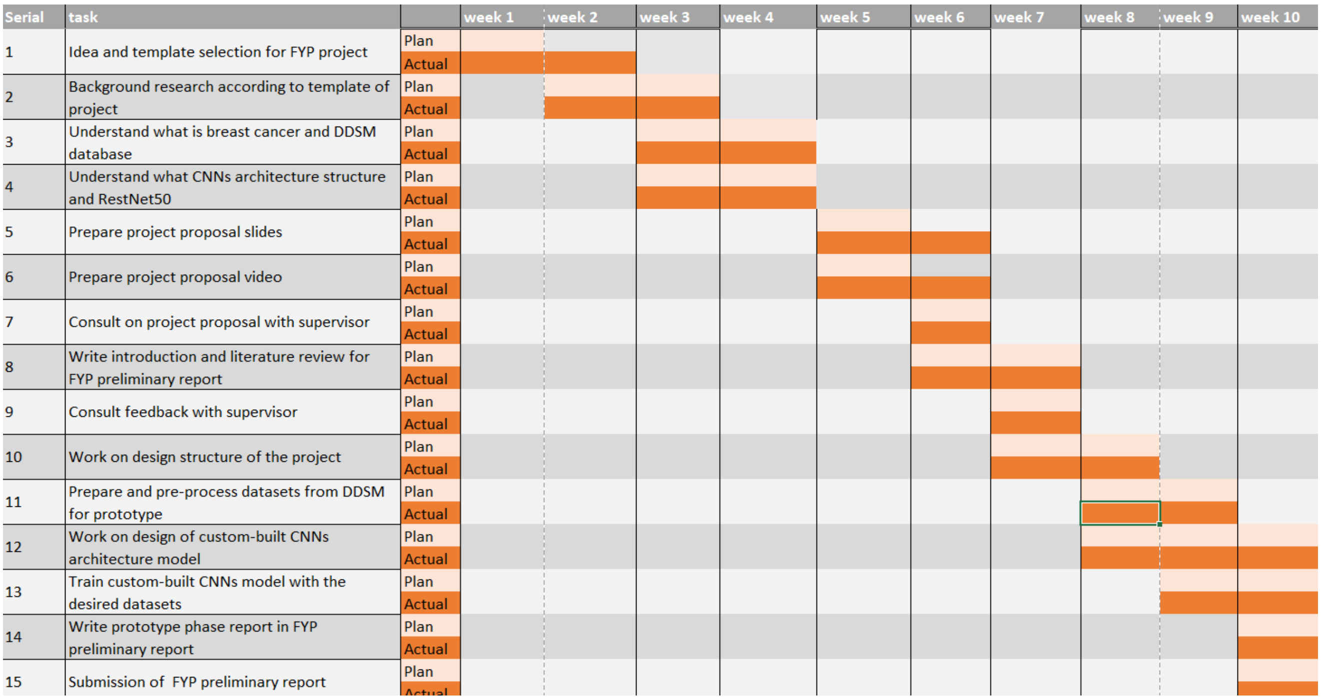
Figure 6 CNNs Architecture Structure

Figure above is a simple CNNs architecture model design with all the simple essential layers such as input, convolution, pooling, fully connected dense neural networks and final output classification. The CNNs architecture design can be divided into two main parts which are the feature extraction and classification sections. The main function of feature extraction section in CNNs architecture design is to automatically extract relevant features from raw image data, transforming it into a more useful and informative representation for classification process of images. The convolutional layer deep learning study(He,K.Zhang et al., 2015)[14] show that this layer act as a learnable filter layer that extract spatial features from images. The number of convolutional layers used typically depends on the complexity and volume of the input data which is the mammogram images. A ReLu(He,K.Zhang et al., 2015)[[14] activation function is implemented in convolutional layers as the convolutional layers use filters to detect patterns and ReLu introduces non-linearity to enable the network structure to learn more complex relationships from the mammogram images. Pooling layers are used to reduce and down samples the features produced by convolutional layers. They retain the important features from the images while reducing spatial dimensions, enhancing computational efficiency during training and help the model to prevent overfitting.

After that, a flatten layer will convert the output of feature extraction into a 1D vector that serve as the input to the fully connected dense neural layer. Classification section form by a fully connected dense layer and an output dense layer. The fully connected dense layer will gather all the learned features from the feature extraction section and combining them to make a prediction in the output dense layer of CNNs structure model. Final output dense layer will apply an activation function for four-class classification to output a probability distribution across the target classes of breast cancer.

.

3.5 Gantt Chart



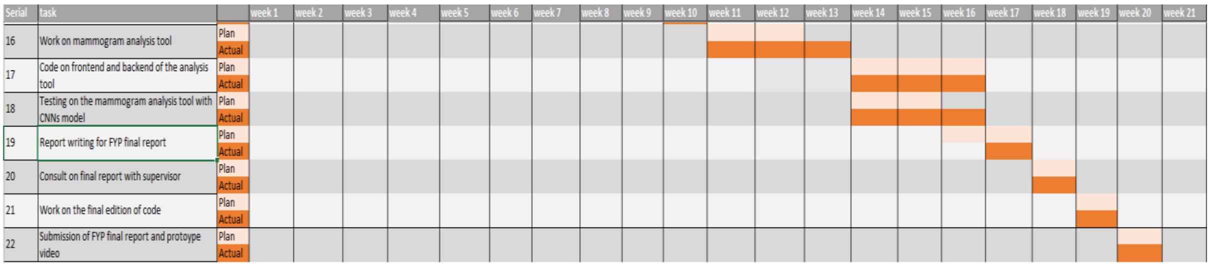


Figure 7 Gantt Chart

According to the Gantt Chart illustrated on the diagrams above, timeline of the project is structured as follows:

- Week 1 to 4 conduct a literature review and background research according to the project template
- Week 4 to 6 focus on the preparation of project proposal slides and video
- Week 7 to 10 involve drafting the FYP preliminary report and developing an initial prototype with a custom-built CNN architecture
- Week 11 to 15 implement a mammogram analysis tool to upload mammogram image, perform analysis and classify into respective class.
- Week 16 onwards, efforts will be focus on FYP final report writing, prototype code final revisions and prototype demonstration video for final submission.

3.6 Test And Evaluation

Several evaluation metrics will be applied to test and evaluate the performance of the models used in this multi-classification breast cancer project. These metrics that is being used in the test and evaluation process[18] are confusion matrix, accuracy, precision, recall, f1-score and ROC-AUC. The CNNs model will be evaluated based on their prediction and classification performance using various metrics.

Confusion Matrix

Confusion matrix for a multi-classification CNNs model is valuable tool that provides a detailed breakdown of the model's prediction across four different classes of the project.

The confusion matrix can be breakdown into four different categories which are the true positive, true negative, false positive and false negative.

Categories	Description
True Positive(TP)	Model predicted = positive Actual outcome = positive
True Negative(TN)	Model predicted = negative Actual outcome = negative
False Positive(FP)	Model predicted = positive Actual outcome = negative
False Negative(FN)	Model predicted = negative Actual outcome = positive

Table 1 Confusion Matrix Elements

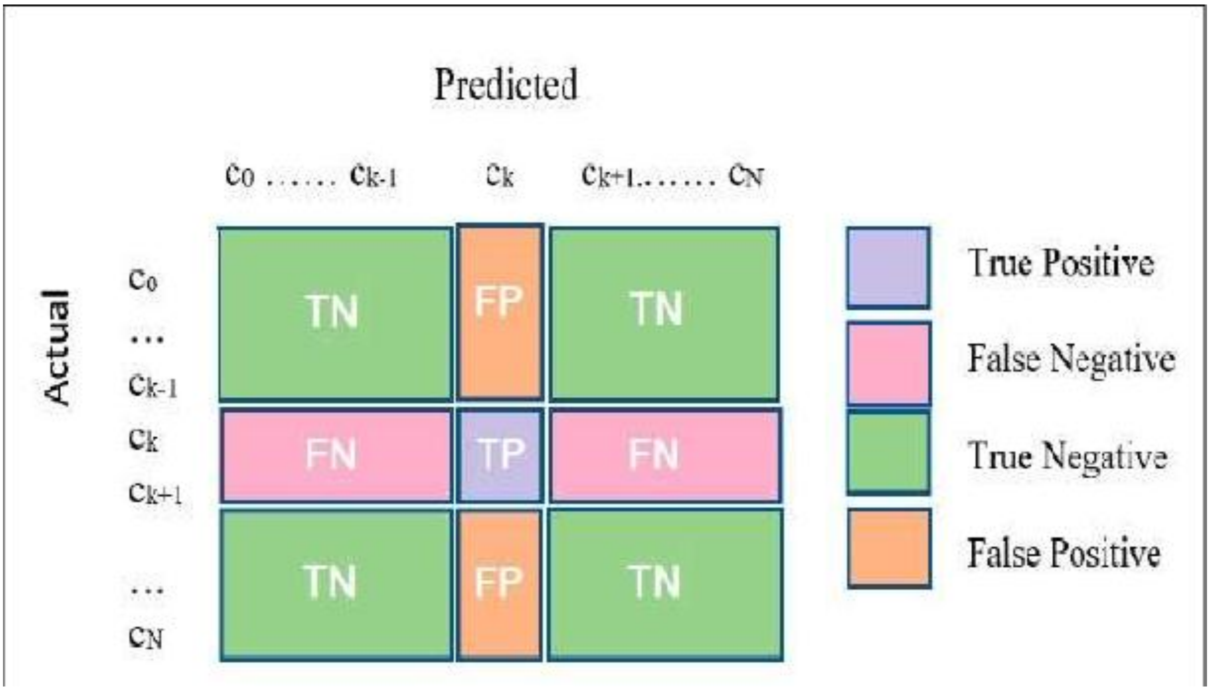


Figure 8 Confusion Matrix

The diagram above presents a sample layout of a confusion matrix, illustrating the arrangement of its four categories. The values within the four categories of the confusion matrix will serve as the foundation for calculating key evaluation metrics of accuracy, precision and recall.

Accuracy

Accuracy reflects the proportion of correct predictions made by the model out of all predictions. It serves as an overall measure of the CNNs model performance, and it is calculated by summing the TP and TN, then dividing by the total number of predictions as shown in the formula below.

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

Precision

Precision focus on the quality of the model's positive prediction. Precision measures on how well the model can correctly identify instances of a specific class out of all instances it predicted to belong that class. As an example, the model predicts a mass to be malignant among all cases, precision tells us how often the prediction is accurate. Calculation of the precision result for the performance of the model's positive prediction shown in the formula below.

$$\boldsymbol{Precision = \frac{TP}{TP + FP}}$$

Recall

Recall also known as sensitivity, measures the model's ability to correctly identify all actual positive instances of a specific class. It is calculated as the proportion of true positive (TP) among all actual positive cases. Actual positive cases can be represented by the sum of TP and FN as it show both positive cases that being detected and missed. Calculation of recall on the model's effectiveness in detecting true positive cases shown in the formula below.

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

F1-Score

F1-socre is a crucial metric for evaluating overall performance of the model. It combines both precision and recall result into a single metric to handle the imbalance class distributions when there is uneven class distribution or when both false positives and negatives matter. Formula of F1-score below represents a harmonic mean of precision and recall.

$$F1 - Score = \frac{2. Precision. Recall}{Precision + Recall}$$

F1-score will ensure a high value that is achieved when both precision and recall are reasonably high. If any of the metric is low will result in the imbalance, making the F1-score as a reliable indicator of a model's ability to correctly identify positive instances while minimizing false positives.

ROC-AUC

ROC-AUC stand for Receiver Operating Characteristic Area Under the curve is a metric used to evaluate the performance of the multi-classification model. According to the diagram below, ROC is the curve, and AUC is the shaded area under the curve.

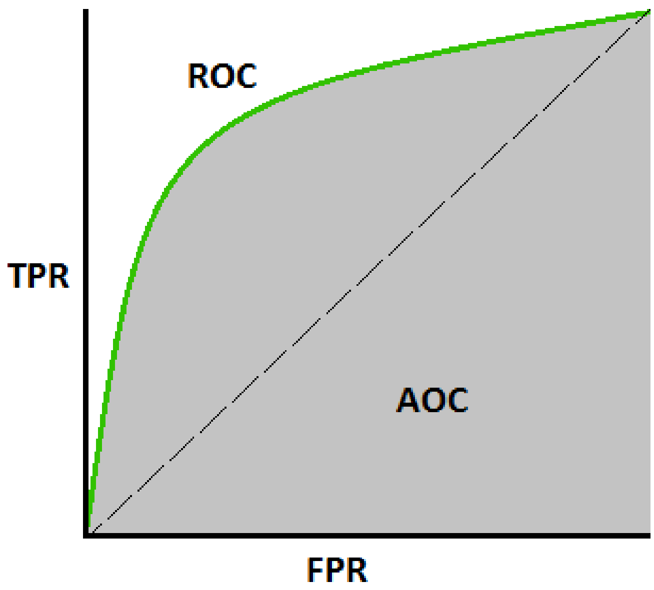


Figure 9 ROC-AUC Diagram

ROC curve is a visual representation of model performance across all thresholds. It is drawn by calculating the true positive rate(TPR) and false positive rate(FPR) and graph the curve by TPR over FPR. TPR calculation formula is same as recall which is same as measure how well the model correctly identify positive cases. FPR calculation shown in the formula below measure how often the model incorrectly identify a negative case as positive.

$$\boldsymbol{FPR = \frac{TP}{FP + TN}}$$

A perfect model of ROC graph is shown in below, which some threshold has TPR of 1 and FPR of 0 which show that the performance of model in predictions is all correct.

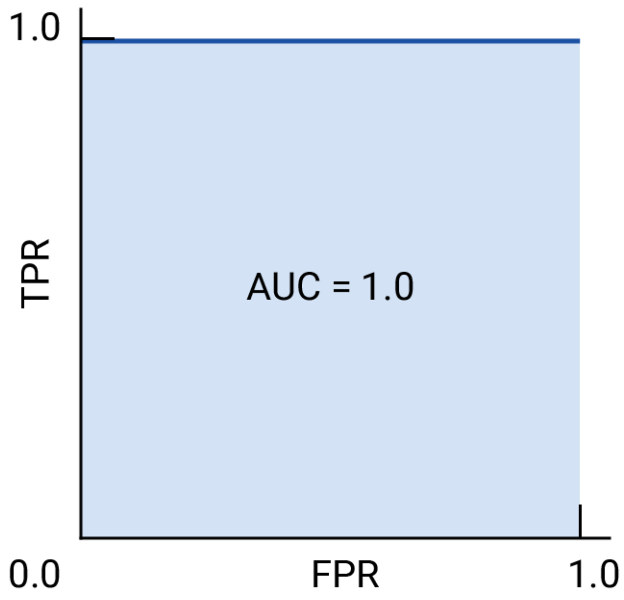


Figure 10 ROC curve with AUC =1.0

AUC represent the area under the ROC curve used to show the probability of the model. As shown in diagram above is the perfect ROC-AUC as the ROC with a TPR of 1.0 and the AUC under the ROC is 1.0 means there is 100% probability on the prediction done by the model.

(2574 Words)

4. Implementation

4.1 Overview

Prototype of the deep learning breast cancer detection project developed by demonstrating the technical feasibility of a deep learning model for multi-classification of breast cancer using mammogram images. This initial implementation serve as a proof of automating diagnostic process by accurately categorizing mammogram images into four classes Benign, Benign Without Callback, Malignant and Normal.

The prototype pipeline comprises several critical stages. It start with data preprocessing which involves image conversion, resizing, normalization, and ensuring consistent images format during the model training and validation to enhance the model performance. After that, it followed by the data augmentation which will adjust the dataset mammogram images through rotation, flipping, brightness and contrast adjustments to artificially expands the dataset and improve the model's generalizability.

Finally, the key stage of the prototype lies on a custom built CNNs architecture. This network will be carefully design and trained to extract the spatial features from mammogram images, which allowing them to learn intricate patterns that differentiate the four breast cancer classes. The prototype will be evaluated to test its model's performance, providing insight into its practical viability and paving the way for further refinements.

4.2 Data Preprocessing

Multi-Class Structuring from DDSM

Multi-class structuring of dataset involved of reformat the mammogram images of Benign, Benign Without Callback and Malignant classes from the DDSM database into the desired format in Pre_Process_DDSM_Images.ipynb. These three cases of DDSM mammogram images can be found in the CBIS-DDSM[19] dataset but conversion process need to be implemented to convert it from DICOM into PNG format. Normal class mammogram images we will convert by another path of MIAS dataset as it is different from current csv file and format.

Mammogram saved as DICOM format in DDSM database, to enhance the performance of the CNNs model we converted the mammogram images into PNG format from the DICOM files. Firstly, all the relative path of the DICOM file is stored in the csv files to perform conversion of images format we mapped out all DICOM file paths according to the diagram below by applying key value pair which key is the relative DICOM file path used to match CSV and value is the full file path used to open the image from the DICOM files.

```
[43]:
# Dictionary to store mapping of relative paths to full DICOM file paths
print("Indexing all DICOM files...")
dicom_file_map = {}

# Walk through the IMG_DIR directory and all its subdirectories
for root, _, files in os.walk(IMG_DIR):
    for file in files:
        # Only process files with .dcm extension (DICOM format)
        if file.endswith(".dcm"):
            # Get relative path to match the CSV entry
            rel_path = os.path.relpath(os.path.join(root, file), IMG_DIR)
            rel_path = rel_path.replace("\\", "/")  # Normalize slashes
            dicom_file_map[rel_path] = os.path.join(root, file)

Indexing all DICOM files...
```

Figure 11 DICOM File Selection

After that, we applied DICOM conversion function to read the DICOM images, convert the raw pixel array to float as the pixel arrays normally stored as integers and “float” help us to decimal precision during normalization. In the function we normalized the pixel values to the range of 0 to 255 as it ensure all pixel values are scaled between 0 (black) and 255(white) which is suitable for mammogram images as the PNG expects to be around 8 bits pixels. Normalization helps the mammogram images avoid value out of range to get a very dark or bright PNG images after conversion. DICOM conversion function shown in the diagram below.

Convert of .dcm images format into PNG format

[44]:

```python
# Function to convert a DICOM file (.dcm) into a PNG image
def dicom_to_png(dcm_path, png_path):
    try:
        # Read the DICOM file using pydicom
        dcm = pydicom.dcmread(dcm_path)
        # Extract the pixel array and convert it to float for normalization
        img = dcm.pixel_array.astype(float)
        # Normalize the pixel values to the 0-255 range
        img = (np.maximum(img, 0) / img.max()) * 255.0
        # Convert the pixel values to unsigned 8-bit integers
        img = np.uint8(img)
        # Create a PIL image from the numpy array
        im = Image.fromarray(img)
        # Save the image to the specified PNG path
        im.save(png_path)
    except Exception as e:
        print(f"Error converting {dcm_path}: {e}")
```

Figure 12 DICOM To PNG Conversion Function

Lastly, converted images being organized and saved in the corresponding three classes folders. We iterated over each row of the csv files in the filtered MLO data frame and found the corresponding .dcm file to determine the output folder based on the pathology class. After that we applied DICOM to PNG conversion function to convert the DICOM file into PNG format images and saved the PNG into the appropriate class folder. Sequence of the conversion will shown in the diagram below.

•[36]:

```python
print("Organizing images...")
not_found = 0

# Iterate through each row in the filtered dataframe that contains only MLO-view images
for _, row in tqdm(df_mlo.iterrows(), total=len(df_mlo)):
    pathology = row['pathology']
    # Map the pathology label to a folder name using the 'categories' dictionary
    target_class = categories.get(pathology)
    if not target_class:
        continue

    patient_id = row['patient_id']
    image_view = row['image view']

    # Try to find a DICOM file path that contains both the patient ID and view type
    # and also make sure it's a "full mammogram" image (not cropped or ROI)
    matches = [
        path for path in dicom_file_map.keys()
        if patient_id in path and f"{image_view}" in path and "full mammogram" in path
    ]
```

```
    if not matches:
        not_found += 1
        print(f"No match for: {patient_id}, {image_view}")
        continue

    # Take first match
    image_rel_path = matches[0]
    # Get the full path to the DICOM file using the mapping dictionar
    source_path = dicom_file_map[image_rel_path]

    # Construct path to save
    dest_file = f"{patient_id}_{image_view}.png"
    class_folder = os.path.join(OUT_DIR, target_class)
    dest_path = os.path.join(class_folder, dest_file)

    # Ensure class folder exists
    os.makedirs(class_folder, exist_ok=True)

    # Convert to PNG
    dicom_to_png(source_path, dest_path)

rint(f"Finished. Total not found: {not_found}")

rganizing images...
00%|██████████| 1896/1896 [1:09:41<00:00,  2.21s/it]
inished. Total not found: 0
```

Figure 13 DICOM To PNG Conversion Process

According to the diagram above, there total of 1896 images successfully converted and saved according to the different classes of breast cancer.

Normal Case Selection

Normal case mammogram images is selected from MIAS dataset which is another set from DDSM database which selecting the normal case images using metadata from mias_info.csv. These images are filtered, copied and resized to 244 x 244 pixels in the Normal_Case_Preprocess.ipynb. We copy and resize the images according to the source and destination path we defined, copies the image if it exists in according to the csv file. We opened the images and resized it to 244 X 244 pixels using LANCOSZ filter and save it into the designated folder with other three cases. Steps of filtered and extracting normal cases images is shown in the diagram below.

```
[6]:
#Define source directory where original normal case image stored in MIAS
image_dir = "normal_case_preprocess/MIAS"
normal_dir = "resize_dataset/Normal"
os.makedirs(normal_dir, exist_ok=True)

normal_df = df_info[df_info['CLASS'] == 'NORM']
normal_ids = normal_df['REFNUM'].tolist()
# Initialize a list to store paths of successfully processed files
copied_files =[]
#looop through each nomral case ID
for img_id in normal_ids:
    src_file = os.path.join(image_dir, f"{img_id}.png")
    dst_file = os.path.join(normal_dir, f"{img_id}.png")
    if os.path.exists(src_file):
        shutil.copy(src_file,dst_file)
        try:
            img = Image.open(dst_file)
            # Resize the image to 244x244 pixels using high-quality resampling
            img_resized = img.resize((244,244),Image.Resampling.LANCZOS)
            img_resized.save (dst_file)
            copied_files.append(dst_file)
        except Exception as e:
            print(f"Error processing {dst_file}: {e}")
```

Figure 14 Normal Case Selection from MIAS

4.3 Data Augmentation

Data augmentation is a technique that artificially increase the size and diversity of a training dataset by creating modified versions on the existing datasets. By applying data augmentation, it can create new, slightly altered datasets of each training images to help the CNNs model generalize the features better. It can prevent overfitting during model training as the augmentation process expose the CNNs to divers input patterns, reducing it tendency to memorize the training data. By making the model invariant to the minor changes such as rotation, translation and flipping is critical in the mammogram images. Augmentation process also act as a regularization or dropout stage as it regularizes the model by showing the different version of the same images in each epoch during training. Diagram below is the data augmentation technique we applied in our prototype of CNNs model.

```python
img_size = (244,244)
batch_size = 16

datagen = ImageDataGenerator(rescale=1./255,
                             validation_split=0.2,
                             rotation_range=10,
                             width_shift_range=0.05,
                             height_shift_range=0.05,
                             zoom_range=0.1,
                             shear_range=0.1,
                             horizontal_flip=True,
                             fill_mode='nearest')
```

```python
train_generator = datagen.flow_from_directory(
    data_dir,
    target_size = img_size,
    batch_size = batch_size,
    class_mode = 'categorical',
    color_mode = 'grayscale',
    subset = 'training',
    shuffle = True
)
```

Figure 15 Data Augmentation on Training Datasets

According to the diagram above, data augmentation being applied in "ImageDataGenerator" as it applied several steps of the augmentation technique. Firstly, it rescale the pixel values of the images from 0-255 to 0-1 for a better training stability. Random rotation images up to 10 degrees in "rotation_range" for simulating orientation variance. Besides that, it shift the images horizontally by 5% of the width and vertically 5% of the height through the "width_shift_range" and "height_shift_range" to simulate patient position variability on the mammogram images. Random zoom in and out of 10% will help the model to learn scale-invariant features of the images. Horizontal flip is to make sure the mammogram are symmetric in left and right. Fill mode used to fill in missing pixels after rotation and shifts.

"Train_generator" is the critical component in the model training pipeline, it responsible for the efficiency loading and augmenting mammogram images in real time. The generator resize the image to a uniform size of 244 x 244 which serve as a safeguard as the resizing has already performed during the initial data preprocessing stage and convert the images into grayscale. It also applies several augmentation various augmentation techniques we mentioned earlier to expand the training dataset and help the model generalize better. Images are loaded in small batch of 16 images per batch which will allow the memory of the model to train efficiently. Having a clear and separate distinction between training and validation dataset is essential for effectively evaluating the CNNs model's performance and ensuring reliable training results.

4.4 Custom Built CNNs Architecture

Custom built CNNs architecture in this project designed for multi-class classification of grayscale mammogram images into 4 classes. It follows a progressive feature extraction pipeline as extract pattern using convolutional layers, down sample to reduce spatial size of the datasets, apply normalization and regularization on the datasets, flatten high-level features and feed into the dense layer, predict class probabilities using softmax activation function. The diagram below illustrates the architecture of the custom-built CNNs used in the prototype design.

```
model= Sequential([
    Input (shape=input_shape),

    #Conv Layer 2
    Conv2D(32, (3,3), padding = 'same', activation='relu'),
    MaxPooling2D(pool_size=(2,2), strides=2),

    #Conv Layer 3
    Conv2D(64, (3,3), padding = 'same', activation='relu'),
    MaxPooling2D(pool_size=(3,3), strides=2),


    #Conv Layer 4
    Conv2D(128, (3,3), padding = 'same', activation='relu'),
    # BatchNormalization(),
    MaxPooling2D(pool_size=(3,3), strides=2),


    Conv2D(258, (3,3), padding = 'same', activation='relu'),
    # BatchNormalization(),
    MaxPooling2D(pool_size=(3,3), strides=2),
    Dropout(0.3),


    #Flatten and Dense Layer
    GlobalAveragePooling2D(),
    #Fully Connected Dense Layer
    #data regularization
    Dense(64, activation='relu', kernel_regularizer=l2(0.001)),
    # BatchNormalization(),
    Dropout(0.5),
    #four class classification
    Dense(4, activation = 'softmax'),
])
```

Figure 16 Custom Built CNNs Structure

According to the CNNs model diagram, the model's design start with a input layer that only accept images of size 244 x 244 and 1 indicates to a single colour channel which represent for grayscale images. Secondly, four convolutional layers being applied from 32→64→128→256 that progressively increases in filter depth to capture hierarchical features of the mammogram images. All four convolutional layers applied 'same' padding that keep the spatial dimensions constant and MaxPooling2D that reduce the spatial dimensions by half, retaining and preserves key feature of the datasets while reducing the less useful information.

A GlobalAveragePooling2D layer applied to replace traditional flatten layer in the CNNs model as it is easier to prevent overfitting due to flatten will generate thousands of parameters. It is simple and faster which required less memory and applied faster training and inference. After that, a fully connected dense(64) layer being applied connect all extracted features from the datasets to a 64-unit layer and a regularization step used to penalize large weight and overfitting. Randomly dropout 50% of nodes during training can improve generalization and overfitting of the model's performance. Lastly, an output layer of 4 probabilities represent four different cases of the mammogram breast cancer images with a softmax activation function used to ensure all the probabilities of four different classes add up is 1. Final prediction is the class with the highest probability which we will used for prediction in the testing and evaluation of the model.

4.5 Mammogram Analysis Tool Implementation

The mammogram analysis tool is implemented as a full-stack web application designed to classify mammogram images into four diagnostic categories. The system integrates a deep learning model trained with mammogram data with a responsive user interface to enable radiologist and healthcare professionals to easily upload images and obtain diagnostic predictions with confidence scores.

Backend Implementation

Backend development of the project is developed with Python and Flask micro web framework, serving API server that host the CNNs model for mammogram classification. Backend development is developed into four key parts which are model loading, image preprocessing, prediction API and CORS handling.

Model Loading

The CNN model is pre-trained and saved in the “custom_cnn_model” file. The Flask app loads this model at startup using Tensorflow Keras API which allow the interference without retraining.

```
app = Flask(__name__)
CORS(app)
model = tf.keras.models.load_model('custom_cnn_model.h5')
```

Image Preprocessing

Image is received through the "/predict" POST endpoint, the backend will convert it into grayscale and resize it into 244x244 pixels, matching the model input size. The pixel values are normalized to the range [0,1] and reshaped to the expected tensor shape(1,244,244,1).

```
def preprocess_image(image):
    image = image.resize((244,244))
    img_array = np.array(image) / 255.0
    img_array = img_array.reshape(1, 244, 244, 1)  # for grayscale images
    return img_array
```

Prediction API

The "/predict" route of http://127.0.0.1:5000/ accepts image files via HTTP POST request, preprocessing the image, feeds it into the CNNs model for inference, and returns a JSON response with prediction probabilities for each class.

```
@app.route('/')
def home():
    return "Welcome to the Breast Cancer Detection API. Use POST /predict to analyze images."

@app.route('/predict', methods=['POST'])
def predict():
    file = request.files['image']
    image = Image.open(file.stream).convert('L')
    input_data = preprocess_image(image)
    preds = model.predict(input_data)
    response = {'predictions': preds.tolist()}
    return jsonify(response)
```

CORS Handling

The Flask-CORS library is used to be enable Cross-Origin Resource Sharing, allowing frontend to communicate with backend API without security issues.

Frontend Implementation

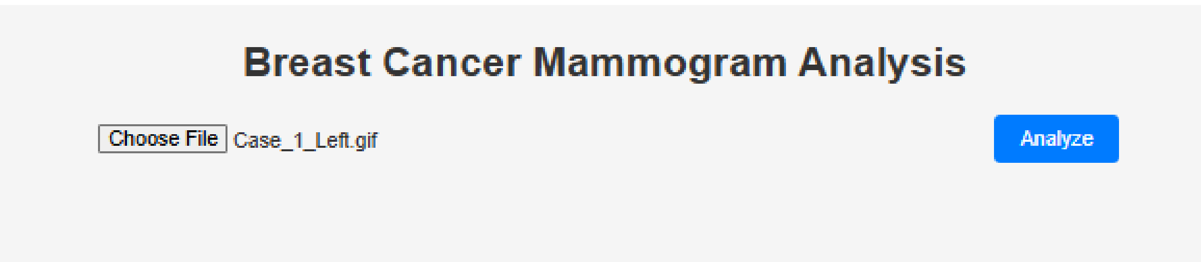
Frontend of the project developed with standard HTML, CSS, and Javascript to provide an intuitive and responsive interface for users to interact with the analysis tool. Frontend design is to ensure a smooth and accessible user experience, enabling efficient mammogram analysis with clear visual feedback.

User Interface Elements

Frontend user interface of the mammogram analysis tool is a page include a file input for users to upload mammogram images, an “Analyze” button is implemented to trigger the prediction, an image preview to display the uploaded mammogram image, and a text area to show the prediction diagnostic category and confidence score for the users.

Image Preview

Upon selecting an image, the frontend immediately displays a preview, improving user confidence that the correct file is selected as shown in the diagram below.



Asynchronous Prediction Request

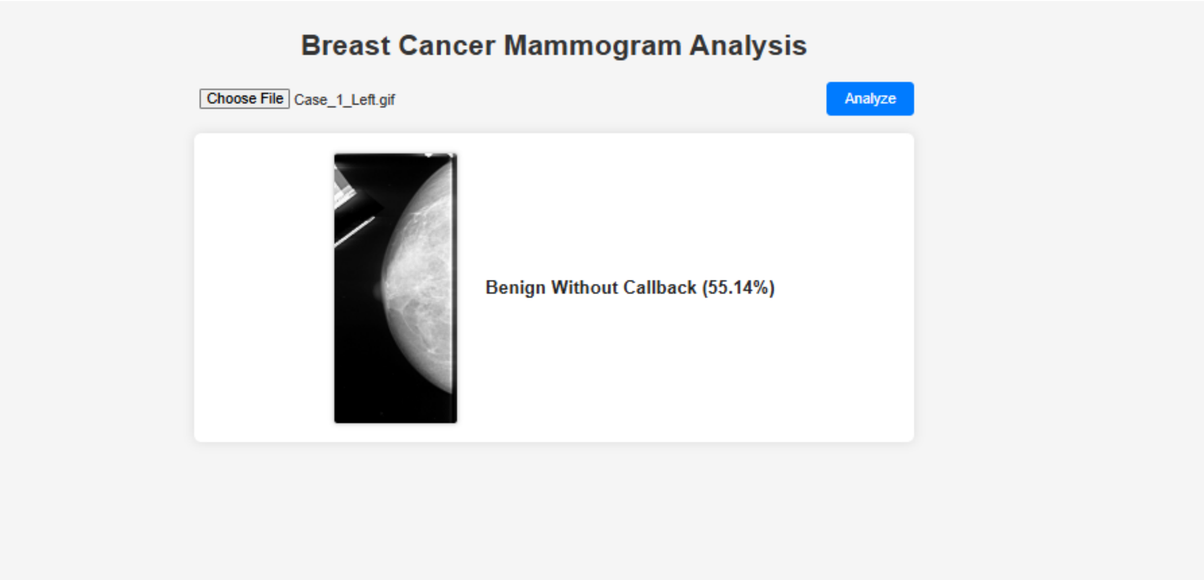
The "uploadImage" function collects the image file and sends it asynchronously to the backend "/predict" API using the fetch API with a POST request containing multipart form data.

```
// Prepare form data
const formData = new FormData();
formData.append("image", file);

try {
  const response = await fetch("http://127.0.0.1:5000/predict", {
    method: "POST",
    body: formData,
  });
```

Prediction Handling and Display

Backend response with prediction probabilities, frontend will extract the highest confidence score and maps it to the corresponding class label. This result is displayed clearly alongside the uploaded image is shown in the diagram below.



Overall, the system is being operates through a well-defined workflow:

1. The users select a mammogram image file from their device.
2. The frontend previews the image and sends it to the backend API
3. The backend preprocesses the image, runs the CNN interface, and return prediction probabilities
4. The frontend displays the most probable diagnosis along with the confidence percentage and the uploaded mammogram image side-by-side

The implementation delivers a practical tool combining a robust CNN model with an interactive web interface to support breast cancer mammogram analysis. The design prioritizes accuracy, usability, and real-time feedback, aiming to augment radiologist' diagnostic workflows and improve early detection rates through AI assisted classification. ***(2009 Words)***

5. Evaluation

Purpose of this evaluation is to critically assess the performance, reliability, and usability of both the deep learning model and overall system implementation. Evaluation is a vital phase that validates whether the project meets its stated objectives, including accurate multi-class classification of mammogram images, efficient backend processing, and an intuitive frontend user interface.

5.1 Functional Testing

Functional testing was performed to validate that the Mammogram Analysis Tool operates correctly as an integrated system. Functional testing used to ensure seamless communication between the frontend user interface and backend API and confirming that the system fulfils all specified functional requirements.

Objectives

- Confirm users can successfully upload mammogram images
- Ensure backend correctly receives, process, and classifies images
- Validate that prediction results are accurately displayed to users
- Assess the responsiveness and stability of the user interface

Test Environment

- Operating system: Window 11
- Browser: Google Chrome
- Backend: Flask API running locally on http://127.0.0.1:5000
- Frontend: Hosted on local machine, accessing backend API via HTTP request

Test case ID	Description	Steps	Expected Result	Actual Result	Status
FT-01	Upload valid mammogram images and get prediction	1. Open frontend page 2. Select a valid mammogram image 3. Click “analyse” button	Image preview display with prediction labels and confidence	As expected	Pass
FT-02	Upload invalid file	1. Select a non-image file 2. Click “analyse”	Alert message asking to select an image file No request send to backend	As expected	Pass
FT-03	Backend API receives and process image	1. Sent POST request with valid image via frontend 2. Observe backend logs and backend	Backend process image, return JSON with predictions	As expected	Pass

FT-04	Display prediction results on frontend	1. After valid prediction, verify prediction label 2. Confidence shown with 2 decimal places	Frontend correctly shows label and confidence percentage	As expected	Pass
FT-05	Responsive design check	1. Resize browser window 2. Test on mobile screen size	UI elements rearrange neatly; image and results preview remain visible and readable	As expected	Pass

Functional testing confirms that the Mammogram Analysis Tool meet its core requirements, enabling users to upload mammogram images and receive accurate multi-class predictions. The system behave as expected under typical use cases and handles invalid inputs. These tests provide a strong foundation for further model refinement, scalability, and deployment.

5.2 Model Evaluation

Model evaluation performed with the dataset split into 80% of training datasets and 20% of validation datasets. After training process of the CNNs model, the model was evaluated using various performance metrics to access it effectiveness. Firstly, we looked at the model validation accuracy as it achieved an approximately 66.1% accuracy by indicating moderate classification performance across the four different cases of mammogram breast images. The results of the model training and validation is illustrated in the diagrams below.

```
Epoch 1/30
65/65 ━━━━━━━━━━━━━━━━━━━━ 40s 598ms/step - accuracy: 0.3062 - loss: 1.4424 - val_accuracy: 0.5019 - val_loss: 1.0236
Epoch 2/30
65/65 ━━━━━━━━━━━━━━━━━━━━ 34s 520ms/step - accuracy: 0.5061 - loss: 1.0381 - val_accuracy: 0.5311 - val_loss: 0.9515
Epoch 3/30
65/65 ━━━━━━━━━━━━━━━━━━━━ 36s 550ms/step - accuracy: 0.5067 - loss: 0.9705 - val_accuracy: 0.5370 - val_loss: 0.8671
Epoch 4/30
65/65 ━━━━━━━━━━━━━━━━━━━━ 34s 527ms/step - accuracy: 0.5265 - loss: 0.9276 - val_accuracy: 0.4961 - val_loss: 0.8673
Epoch 5/30
65/65 ━━━━━━━━━━━━━━━━━━━━ 40s 611ms/step - accuracy: 0.5313 - loss: 0.9038 - val_accuracy: 0.5895 - val_loss: 0.8421
Epoch 6/30
65/65 ━━━━━━━━━━━━━━━━━━━━ 37s 568ms/step - accuracy: 0.5382 - loss: 0.8932 - val_accuracy: 0.5642 - val_loss: 0.8325
Epoch 7/30
65/65 ━━━━━━━━━━━━━━━━━━━━ 34s 519ms/step - accuracy: 0.5637 - loss: 0.8619 - val_accuracy: 0.6031 - val_loss: 0.8121
Epoch 8/30
65/65 ━━━━━━━━━━━━━━━━━━━━ 33s 507ms/step - accuracy: 0.5685 - loss: 0.8687 - val_accuracy: 0.6128 - val_loss: 0.7880
Epoch 9/30
65/65 ━━━━━━━━━━━━━━━━━━━━ 31s 480ms/step - accuracy: 0.5800 - loss: 0.8673 - val_accuracy: 0.6089 - val_loss: 0.8077
Epoch 10/30
65/65 ━━━━━━━━━━━━━━━━━━━━ 31s 482ms/step - accuracy: 0.5612 - loss: 0.8566 - val_accuracy: 0.6109 - val_loss: 0.7887
Epoch 11/30
65/65 ━━━━━━━━━━━━━━━━━━━━ 31s 476ms/step - accuracy: 0.5736 - loss: 0.8174 - val_accuracy: 0.6245 - val_loss: 0.7592
Epoch 12/30
65/65 ━━━━━━━━━━━━━━━━━━━━ 31s 475ms/step - accuracy: 0.5782 - loss: 0.8244 - val_accuracy: 0.6206 - val_loss: 0.7691
Epoch 13/30
65/65 ━━━━━━━━━━━━━━━━━━━━ 31s 483ms/step - accuracy: 0.5719 - loss: 0.8269 - val_accuracy: 0.6226 - val_loss: 0.7608
Epoch 14/30
65/65 ━━━━━━━━━━━━━━━━━━━━ 32s 493ms/step - accuracy: 0.5727 - loss: 0.8079 - val_accuracy: 0.6459 - val_loss: 0.7545
Epoch 15/30
65/65 ━━━━━━━━━━━━━━━━━━━━ 31s 483ms/step - accuracy: 0.5731 - loss: 0.8277 - val_accuracy: 0.6479 - val_loss: 0.7440
Epoch 16/30
65/65 ━━━━━━━━━━━━━━━━━━━━ 31s 477ms/step - accuracy: 0.5911 - loss: 0.7974 - val_accuracy: 0.6518 - val_loss: 0.7411
Epoch 17/30
65/65 ━━━━━━━━━━━━━━━━━━━━ 31s 480ms/step - accuracy: 0.6028 - loss: 0.8150 - val_accuracy: 0.6595 - val_loss: 0.7424
Epoch 18/30
65/65 ━━━━━━━━━━━━━━━━━━━━ 31s 479ms/step - accuracy: 0.6038 - loss: 0.7908 - val_accuracy: 0.6479 - val_loss: 0.7296
Epoch 19/30
65/65 ━━━━━━━━━━━━━━━━━━━━ 31s 478ms/step - accuracy: 0.5923 - loss: 0.8016 - val_accuracy: 0.6401 - val_loss: 0.7255
Epoch 20/30
65/65 ━━━━━━━━━━━━━━━━━━━━ 31s 479ms/step - accuracy: 0.5828 - loss: 0.7979 - val_accuracy: 0.6362 - val_loss: 0.7268
Epoch 21/30
65/65 ━━━━━━━━━━━━━━━━━━━━ 31s 476ms/step - accuracy: 0.5780 - loss: 0.8054 - val_accuracy: 0.6401 - val_loss: 0.7265
Epoch 22/30
65/65 ━━━━━━━━━━━━━━━━━━━━ 32s 496ms/step - accuracy: 0.5921 - loss: 0.7856 - val_accuracy: 0.6634 - val_loss: 0.7330
Epoch 23/30
65/65 ━━━━━━━━━━━━━━━━━━━━ 34s 519ms/step - accuracy: 0.5927 - loss: 0.8087 - val_accuracy: 0.6148 - val_loss: 0.7196
Epoch 24/30
65/65 ━━━━━━━━━━━━━━━━━━━━ 32s 487ms/step - accuracy: 0.6013 - loss: 0.7785 - val_accuracy: 0.6265 - val_loss: 0.7248
Epoch 25/30
65/65 ━━━━━━━━━━━━━━━━━━━━ 31s 480ms/step - accuracy: 0.5819 - loss: 0.7985 - val_accuracy: 0.6401 - val_loss: 0.7397
Epoch 26/30
65/65 ━━━━━━━━━━━━━━━━━━━━ 31s 480ms/step - accuracy: 0.6122 - loss: 0.8131 - val_accuracy: 0.6537 - val_loss: 0.7190
Epoch 27/30
65/65 ━━━━━━━━━━━━━━━━━━━━ 31s 480ms/step - accuracy: 0.5865 - loss: 0.7848 - val_accuracy: 0.6634 - val_loss: 0.7166
Epoch 28/30
65/65 ━━━━━━━━━━━━━━━━━━━━ 31s 479ms/step - accuracy: 0.5907 - loss: 0.7829 - val_accuracy: 0.6420 - val_loss: 0.7254
Epoch 29/30
65/65 ━━━━━━━━━━━━━━━━━━━━ 31s 476ms/step - accuracy: 0.6194 - loss: 0.7488 - val_accuracy: 0.6537 - val_loss: 0.7253
Epoch 30/30
65/65 ━━━━━━━━━━━━━━━━━━━━ 32s 484ms/step - accuracy: 0.6087 - loss: 0.7432 - val_accuracy: 0.6498 - val_loss: 0.7169
```

```
13]: val_loss, val_accuracy = model.evaluate(validation_generator)
     print(f"Validation Accuracy : {val_accuracy:.4f}")
     print(f"Validation Loss : {val_loss:.4f}")

     17/17 ━━━━━━━━━━━━━━━━━━━━ 2s 113ms/step - accuracy: 0.4236 - loss: 0.8937
     Validation Accuracy : 0.6634
     Validation Loss : 0.7166
```

Figure 17 Model Training Performance and Validation Accuracy

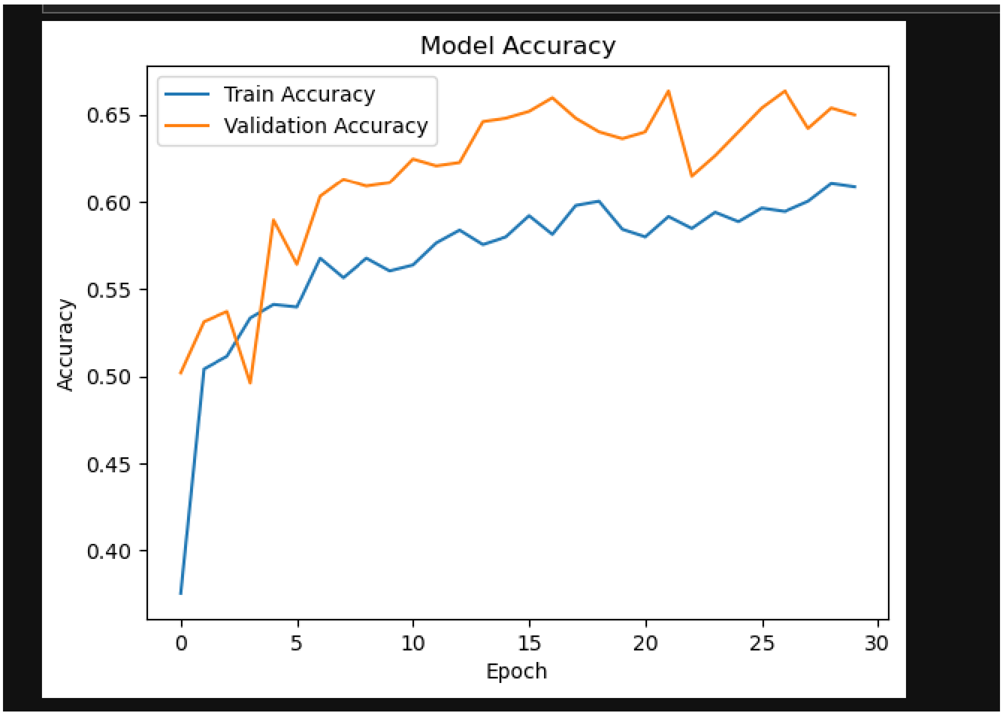


Figure 18 Model Accuracy Graph

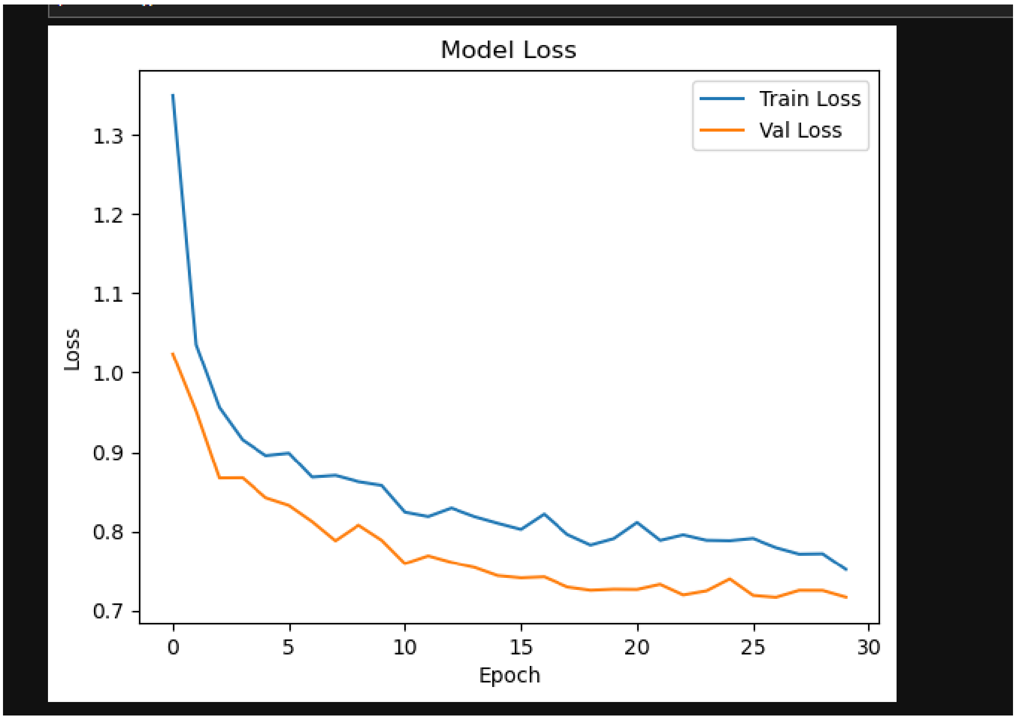


Figure 19 Model Loss Graph

According to the model accuracy graph shown above, the training accuracy starts low with around 38% increases over time reaching around 61% and it shows an upwards trend which indicates that the model is gradually learning meaningful patterns from the training data. Validation accuracy begins similar to training accuracy and overtakes it, stabilized around 65-

66% after 10 epochs which is consistently higher than the training accuracy. This is slightly unusual but a strong regularization such as dropout and augmentation of the datasets which can prevent the model from overfitting. Overall, the model shows a good learning behaviour without signs of overfitting and relatively stable and superior validation accuracy.

According to the model loss graph, it illustrates the decrease in categorical cross entropy loss for both training and validation datasets. Training and validation loss steadily decreases across epochs and shows a stable loss curve for both training and validation datasets to confirm that the model is not overfitting. The lower validation loss compared to training loss is likely due to the strong regularization and augmentation of the training datasets which restrict training performance but improve generalization of the model.

Evaluation of the model performed with the confusion matrix of different classes as shown in the confusion matrix diagram below.

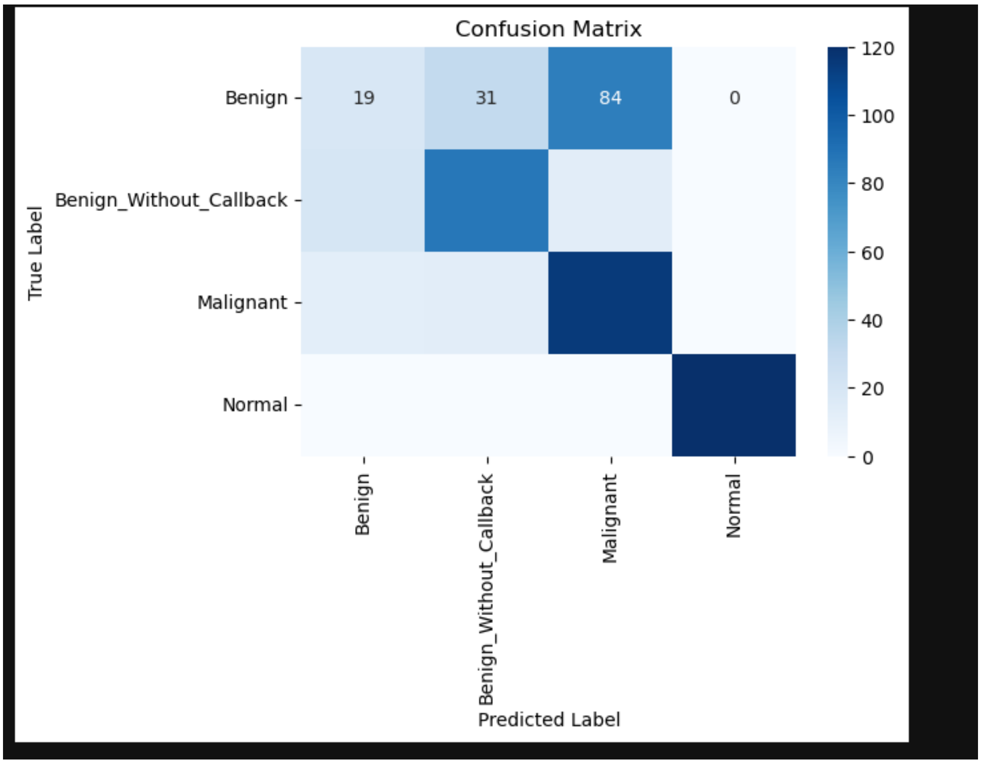


Figure 20 Multi-Class Confusion Matrix

According to the confusion matrix diagram, malignant and normal cases are well predicted, benign and benign without callback show confusion. Normal and malignant cases show a good result which represents the model is highly confident to recognise both cases mammogram images due to clearer and fewer abnormalities. Both cases are more visually distinct and evenly distributed samples. Both benign and benign without callback show a confusion result, high chance due to the similar visual features of the mass shapes of lump. ROI masking or segmentation during preprocess of data should be perform during future project development.

```
[17]: class_labels = ['Benign', 'Benign_Without_Callback', 'Malignant', 'Normal']

      report = (classification_report(y_true, y_pred, target_names=class_labels, digits=3, output_dict = True))

      # Convert to DataFrame
      df_report = pd.DataFrame(report).transpose()

      displayed_report = df_report.loc[class_labels, ['precision', 'recall', 'f1-score', 'support']].round(3)

      display(displayed_report)
```

	precision	recall	f1-score	support
Benign	0.373	0.142	0.205	134.0
Benign_Without_Callback	0.664	0.725	0.693	120.0
Malignant	0.542	0.821	0.653	140.0
Normal	1.000	1.000	1.000	120.0

Figure 21 Table Evaluation of Model

A classification report displayed on the diagram above, summarize the model performance across four mammogram classes. Each row represent precision, recall, and F1-score for a specific class. The model performs exceptionally well on the normal class which achieving a 1.00 score for precision, recall and F1-score, show that all normal case were correctly predicted without any false positive and negative. The malignant class shows a strong performance, with a recall of 0.82 and a balanced of F1-score of 0.653, suggesting the model is effectively at identifying malignant cases. Benign without callback class achieved a high recall od 0.725 and precision of 0.664, lead to a F1-score of 0.693 indicate a reliable classification. However, benign class is struggle with a low precision and recall score of 0.373 and 0.142, result in a poor F1-score 0.205. This shows that the models often misclassifies benign cases and likely confusing them between malignant and benign without callback due to similar visual features of the lump mass shape. Overall, the model shows excellent capability in identifying Normal and a reasonably good performance in detecting malignant and benign without callback cases, it needs further refinement and tuning to improve classification for benign cases.

5.3 User Testing

To evaluate the usability, effectiveness, and user satisfaction of the Mammogram Analysis Tool with real users from the target demographic. Each participant was giving access to the web-based tool and asked to perform a series of representative tasks, including uploading mammogram images from each of the four classification categories, viewing the resulting predictions, and interpreting the confidence score and classification labels displayed by the system.

Participants

A total of 10 participants were involved in the testing process which consist of:

- 5 Computer Science Bacher Degree students
- 5 general users with no medical and IT background

Testing Environment

Testing was conducted in a controlled environment using Window 11 laptops with latest Google Chrome version. All participants accessed the web-based Mammogram Analysis Tool through the same network to ensure consistent performance.

Methodology

Participants provided with a set of tasks designed to stimulate typical users interactions.

1. Upload a mammogram image with PNG format
2. View the classification output generated by the CNNs model
3. Interpret the accompanying visual explanation with the mammogram image display with the confidence score of the classification label

4. Navigate through result details and additional information
5. Complete a post-task questionnaire rating ease of use, clarity and satisfaction of the Mammogram Analysis Tool

Each session was observed by the researcher who noted user behaviour and any difficulties encountered. After task completion, participants were interviewed to gather qualitive feedback with the questionnaire form stated in the appendix section.

Results

The user testing produced the following key findings regarding task performance, usability, system accuracy, and participant feedback:

Positive feedback

- All 10 participants successfully completed the assigned tasks, including uploading mammogram images, viewing classification results, and interpreting visual explanations.
- 8 out of 10 participants rate the system is “easy to use”
- 5 of the users with IT background state that 90% of the model classification output is aligned with their expectation results
- The interface layout was considered intuitive and clean by the majority of users

Area of Improvement

- Non-expert express confusion over terminology between cases of “benign” and “benign without callback”
- A lack of visible upload progress bar during image submission led to uncertainty of some participants
- Minor UI inconsistences were noted on smaller screen sizes, impacting navigation ease

The user testing demonstrated Mammogram Analysis Tool is effective and well-received by both expert and non-expert users. While minor usability and interface issues were identified, the overall system met key user requirements and functional expectations. The limitations and recommended improvements will be explored in greater detail in the following section.

5.4 Limitation and Improvement

Several limitations had discovered while demonstrating the feasibility of the deep learning for multi-class classification. One major limitation is the class imbalance of the dataset, where certain categories of Benign Without Callback and Normal classes had significantly fewer images compared to other two classes. This class imbalance contributed to the poor model performance on the benign class as it have feature similarity with the two imbalance classes as they often share visual characteristics. Besides that, the project relies on the custom built CNNs is useful for experimentation, lacks the depth and generalize power of pretrained models. The dataset size was also relatively limited around 2500 images, which can hinder the model's ability to generalize complex patterns of the images.

5.5 Summary

Evaluation acknowledged important project limitations, including datasets class imbalance and the limited capacity of the custom CNN compared to the state-of-the-art pretrained architectures. The relatively small dataset size also constrained the model's ability to generalize complex features. These factors contributed to lower performance in certain classes and highlight directions for future research. The evaluation demonstrates that the Mammogram Analysis Tool successfully fulfils its primary goals of providing an accurate, user-friendly system for multi-class mammogram classification. The critical insights gathered offer clear pathways to enhance model accuracy, system usability, and clinical applicability, paving the way for future refinements and potential clinical deployment. ***(1611 Words)***

6. Conclusion and Summary

6.1 Main goals and Aims

The Mammogram Analysis Tool was initiated with a clear and impactful goal to build an AI-powered application that facilitates early detection of breast cancer through automated multi-class classification of mammogram images. This objective was grounded in the increasing demand for advanced diagnostic tools that assist radiologists in identifying breast abnormalities with higher accuracy and efficiency.

Throughout the project, the aim was not only to build a high-performing CNNs but also to ensure that the entire system from model training to user interaction was robust, responsive, and user-friendly. The solution was designed to reduce the manual workload of healthcare professionals, provide consist image interpretation, and enhance clinical decision-making through technology. By integrating a trained deep learning model into a responsive web-based interface, the project achieved its primary vision of merging AI with healthcare diagnostics.

6.2 Key Points and Details

- **Model Architecture Design**: A custom CNN was developed and trained to classify mammogram images into four distinct categories. The architecture was carefully chosen to strike a balance between performance and computational efficiency.
- **Backend API Development**: A Flask-based backend was implemented to handle image preprocessing, model inference, and JSON response generation. This ensured real-time interaction between the frontend and the AI model
- **System Integration and Testing**: Rigorous functional and user testing ensured tha the frontend and backend worked harmoniously. Every component from image upload to result display was tested under real-world scenarios to ensure reliability and responsiveness.

- **Model Evaluation and Metrics**: The CNN model achieved 66.1% validation cases. Detailed evaluation using accuracy graphs, loss curves, confusion matrices, and classification reports provided in-depth insight into model behaviour.

6.3 Lesson and Improvement

Lesson Learned

- Early user feedback highlighted the need for clarity in medical terminology and responsive design considerations
- Handling dataset imbalance and refining model accuracy underscored the complexities of medical image analysis
- Necessity of advanced preprocessing and augmentation techniques

Ways to Improve

- Model and Data Enhancement
  - Addressing class imbalance with advanced data augmentation
  - Incorporating region-of-interest masking during preprocessing
  - Exploring deeper CNN architecture to boost classification accuracy
- Deployment and Scalability
  - Preparing backend for production-level deployment with secure authentication and scalable infrastructure to handle multiple concurrent users
  - Implementing comprehensive logging and monitoring to support clinical reliability and maintenance

6.4 Overall

Development of the Mammogram Analysis Tool represents a meaningful contribution to the intersection of artificial intelligence and medical imaging. By combining deep learning methodologies with an intuitive web-based interface, the project has shown that it is feasible to create an AI-powered support system capable of assisting radiologist in the multi-class classification of mammogram images. This accomplishment goes beyond technical experimentation as it reflects a practical step towards integrating AI into clinical workflows in a way that prioritizes both accuracy and usability.

From a technical perspective, the project successfully demonstrated the end-to-end pipeline of data preprocessing, augmentation, model design, and deployment. The custom-built CNNs architecture achieved moderate but encouraging performance in detecting normal and malignant cases of the mammogram breast cancer detection. There is a challenge of distinguishing between visually similar classes of benign and benign without callback cases. These results validate the potential of CNN-based models in healthcare applications but also emphasize the importance of addressing class imbalance, dataset diversity, and preprocessing refinements in future iterations. Through functional and user testing, the prototype proved accessible not only to technical users but also to non-experts which indicate that AI tools can be designed to support clinical decision-making while remaining approachable for a broad range of users. The project design includes image preview, prediction confidence scores, and a clean interface, meeting user needs and providing a base for future improvements like clearer medical terms, progress indicators, and mobile support.

In conclusion, this project underscores the transformative potential of AI in healthcare. It demonstrates how carefully designed deep learning systems can support earlier, more reliable diagnoses, ultimately improving patient outcomes. Through challenges remain, the Mammogram Analysis Tool establishes a solid foundation for continued research, refinement and eventual clinical adoption which marking a step forward in the ongoing fight against breast cancer. ***(676 Words)***

Chapter 7 Appendices

Code Repository

https://github.com/yhlee026/deep-learing-breast-cancer-detection.git

References

1. World Health Organization. (2024). Breast cancer. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/breast-cancer
2. Cohen, E. (2024). Mammogram vs. breast ultrasound: What’s the difference? MD Anderson Cancer Center. https://www.mdanderson.org/cancerwise/mammogram-vs--breast-ultrasound--what-s-the-difference.h00-159702279.html
3. McKinney, S. M. (2020). International evaluation of an AI system for breast cancer screening. Nature, 577(7788), 89–94. https://doi.org/10.1038/s41586-019-1799-6
4. Heath, M., Bowyer, K., Kopans, D., Kegelmeyer, P., Moore, R., Chang, K., & Munishkumaran, S. (1998). Current status of the Digital Database for Screening Mammography. In Computational Imaging and Vision (pp. 457–460). https://doi.org/10.1007/978-94-011-5318-8_75
5. Sarvamangala, D. R., & Kulkarni, R. V. (2021). Convolutional neural networks in medical image understanding: A survey. Evolutionary Intelligence, 15(1). https://doi.org/10.1007/s12065-020-00540-3
6. Agarwal, R., Diaz, O., Lladó, X., Yap, M. H., & Martí, R. (2019). Automatic mass detection in mammograms using deep convolutional neural networks. Journal of Medical Imaging, 6(3), 1–11. https://doi.org/10.1117/1.jmi.6.3.031409
7. International Agency for Research on Cancer. (2022). Global cancer observatory: Cancer tomorrow. https://gco.iarc.fr/tomorrow
8. Lehman, C. D., Wellman, R. D., Buist, D. S. M., Kerlikowske, K., Tosteson, A. N. A., & Miglioretti, D. L. (2015). Diagnostic accuracy of digital screening mammography with and without computer-aided detection. JAMA Internal Medicine, 175(11), 1828–1837. https://pubmed.ncbi.nlm.nih.gov/26414882/
9. Radiology Business. (2023). The 2023 radiology practice development report: Radiologists read an average of 4.7 subspecialties in daily practice; less than half of radiologists feel ‘very confident’ in the studies they read.

https://radiologybusiness.com/press-release/2023-radiology-practice-development-report-radiologists-read-average-47-subspecialties-daily

10. UChicago Medicine. (2025). Ability to read mammograms varies widely. https://www.uchicagomedicine.org/forefront/news/ability-to-read-mammograms-varies-widely
11. Ekpo, E. U., Alakhras, M., & Brennan, P. (2018). Errors in mammography cannot be solved through technology alone. Asian Pacific Journal of Cancer Prevention, 19(2), 291–301. https://doi.org/10.22034/APJCP.2018.19.2.291
12. Nature Communications. (2020). Non-invasive early detection of cancer four years before conventional diagnosis using a blood test. Nature Communications, 11, 3475. https://www.nature.com/articles/s41467-020-17316-z
13. Keita, Z. (2023). An introduction to convolutional neural networks (CNNs). Datacamp. https://www.datacamp.com/tutorial/introduction-to-convolutional-neural-networks-cnns
14. He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep residual learning for image recognition. arXiv. https://arxiv.org/abs/1512.03385
15. The Breast. (2022). Current and future burden of breast cancer: Global statistics for 2020 and 2040. The Breast, 66, 15–23. https://doi.org/10.1016/j.breast.2022.08.01
16. Shockney, L. (2023). Breast cancer facts & statistics. National Breast Cancer Foundation. https://www.nationalbreastcancer.org/breast-cancer-facts/
17. Hu, Y. (n.d.). Convolutional neural networks in multi-class classification of medical data. arXiv. https://arxiv.org/pdf/2012.14059
18. The Cancer Imaging Archive. (n.d.). CBIS-DDSM. https://www.cancerimagingarchive.net/collection/cbis-ddsm/
19. Debnath, S., Roy, R., & Changder, S. (2024). A novel approach using deep convolutional neural network to classify the photographs based on leading-line by fine-tuning the pre-trained VGG16 neural network. Multimedia Tools and Applications, 83(3), 3189–3214. https://doi.org/10.1007/s11042-022-13338-5

20. Hasan, N., Bao, Y., Shawon, A., Hossain, M. S., Chakraborty, S., & Mahmud, M. (2021). DenseNet convolutional neural networks application for predicting COVID-19 using CT image. SN Computer Science, 2, 389. https://doi.org/10.1007/s42979-021-00782-7
21. Arbyn, M., Weiderpass, E., Bruni, L., de Sanjosé, S., Saraiya, M., Ferlay, J., ... & Bray, F. (2020). Estimates of incidence and mortality of cervical cancer in 2018: A worldwide analysis. *The Lancet Global Health*, *8*(2), e191–e203. https://doi.org/10.1016/S2214-109X(19)30482-6
22. Welch, H. G. (2014). Quantifying the benefits and harms of mammography. *JAMA Internal Medicine*, *174*(3), 448–449. https://jamanetwork.com/journals/jamainternalmedicine/article-abstract/1792915