Organizing Ultrasound Imaging Data for Enhanced Breast Cancer Diagnosis with Deep Learning Models

Pranali Ravindra Shendekar

MSDS, University of Wisconsin – La Crosse

DS785: Capstone Project

Dr. Alexander Korogodsky

Date: April 28, 2023

Abstract

This capstone project aimed to develop a software tool to manage and process ultrasound imaging data to train deep-learning models for predicting the presence of cancerous lesions in breast tissue. The project involved collecting, managing, updating, and summarizing study and annotation data in batches as they become available. The study data included ultrasound images, patient information, biopsy results, BI-RADS scores, and metadata about the images and ultrasound equipment. In addition, the annotation data had additional labels and visual outlines of the lesions, which were provided retrospectively by trained radiologists.

The software tool was designed to extract information from the metadata and text annotations accompanying each image and add it to the master index file. Additionally, the tool had to be able to copy the images to the correct locations in the collection and ensure that the image and patient IDs were distinct from those of previously added data. The tool also allowed for the possibility of overwriting existing data, adding additional columns to the index file, and incorporating corrections. Furthermore, the tool included additional columns that may be collected later from the annotation data.

Python code was also developed to enable simple filtering of the image data and to write data downloaders allowing retrieving the image data from the collection. Finally, to show the efficiency of the database, the dataset was tested on a simple deep learning ResNet50 pre-trained model, and the PyTorch data loaders were used to pass data to the deep learning model.

Overall, this project contributed to developing a useful tool that can aid in the early detection of breast cancer, leading to better patient outcomes. The software tool developed in this project can be used in further research studies in medical imaging, deep learning, and clinical settings to support radiologists in diagnosing breast cancer.

Keywords: Ultrasound imaging, deep-learning, decision support software, breast cancer diagnosis, radiology

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

1.1 Background:

Breast cancer is caused by the abnormal growth of cells in the breast, particularly in the milk-producing ducts, known as invasive ductal carcinoma. This occurs when breast cells grow abnormally. Breast cancer is a prevalent cancer type in the United States, with approximately 287,850 women being diagnosed and 43,250 dying from the disease each year, according to the American Cancer Society (2022). Mcdowell S. (2015) says that breast cancer mainly affects after the age of 50 in women. However, half of the United States women dying from breast cancer are age 70 and older. Having many pregnancies and giving birth before age 30 is associated with a lower risk of developing breast cancer in the future. In contrast, women who have never given birth or given birth after 30 have a higher risk of developing breast cancer later in life.

According to Saenz J. (2022), breast cancer was more likely to cause death in Black women than in White women, with a 41% higher mortality rate, even though the former had a lower risk of being diagnosed with the disease. This was primarily due to the fact that Black women tended to be diagnosed with breast cancer at a later stage, which made treatment more challenging. Anastasiadi et al. (2017) analyzed that the breast cancer detection rates were found to have increased among African American women, decreased among Hispanic women, and remained stable among whites, Asian Americans, and American Indians. We need increased quality screenings, treatments, and early diagnosis to control breast cancer. Breast cancer screenings are developed differently in different countries. Mamograms are commonly used for the early detection of breast cancer. They are an x-ray picture of the breast. In European countries, women age 50 to 70 receive an invitation from the government for mammographic screening exams periodically (every two years). When radiologists illuminate the screening mammograms, they mainly look for lesions with different characteristics.

1.2 Problem statement:

The study aimed to develop a software system that can efficiently collect, manage, update, and summarize breast ultrasound imaging data to train deep-learning models to predict potentially cancerous lesions in ultrasound images. This system should be able to handle both the patient study data and ultrasound annotation data, which arrive in batches and consist of various types of data, including images, metadata, text labels, and visual outlines. The system should be able to extract relevant information from the data, add it to a master index file, and organize the images and data in a structured manner that allows for easy collection expansion. Additionally, the system should be able to correct errors and allow for simple data filtering to select relevant subsets for further analysis.

1.3 Inspirations:

I have always been fascinated by using machines to analyze medical data. So I was delighted with the opportunity to work with Mayo Clinic on their real-time data. In a screening of mammography programs, artificial intelligence seemed to be a reliable tool for breast cancer detection. As per the new research, radiologists, when assisted by an artificial intelligence screen, can perform their work more successfully than alone. Kiros H. (2022) said that when the doctor and artificial intelligence worked together, they did 2.6% faster than the doctor working alone. It also raised fewer alarms. The earlier detection of breast cancer will help improve survival rates. Artificial intelligence analyzes the images, detects the breast density and masses, and generates the cancer risk assessment. Artificial intelligence is helping automate processes, increasing work efficiency, and improving decision-making. It all inspired me a lot to undertake this project.

1.4 Objectives

Healthcare data received for this project was fairly unstructured. The information did not have specific or predefined data models or schema. The main goal was to convert the unstructured or raw data into an organized format. The organized data is easier to work with and analyze, leading to more informed and effective business practices. The high data quality will also increase overall productivity and allow the highest quality of information for decision-making.

The study aimed to develop decision support software that used deep learning models to predict the presence of potentially cancerous lesions in breast ultrasound images. The project aimed to collect, manage, update, and summarize ultrasound imaging data, including annotations added by radiologists, to train deep learning models. The study also aimed to create a structured master image collection that can be easily expanded with additional information or annotations as needed for the decision support software.

1.5 Research Questions:

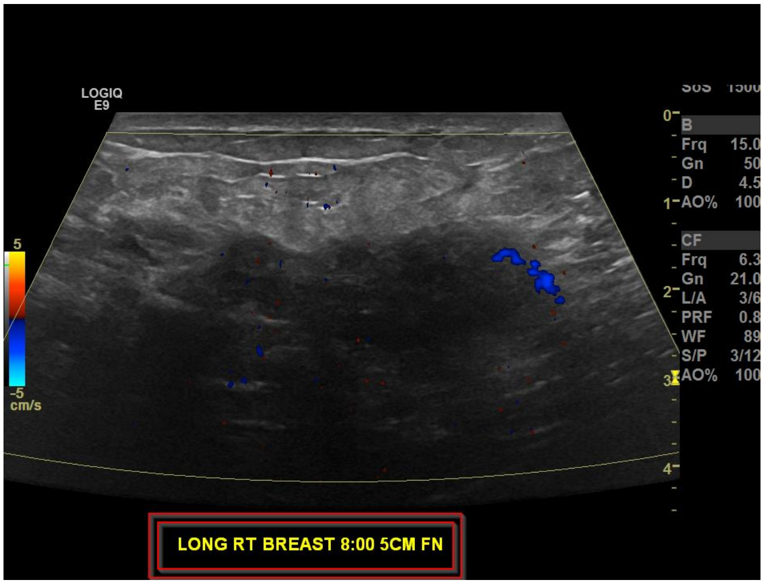
- Can deep learning models be trained using the collected breast ultrasound imaging data and annotation data to accurately predict the type of lesions (benign and malignant) in breast tissue?
- How can collecting, managing, updating, and summarizing ultrasound imaging data be optimized to ensure the resulting data is high quality and suitable for training deep learning models for breast cancer diagnosis?

1.6 Proposed approach:

The project started with the data collection process. Kryzhanivska (2017) emphasized that data was essential to measuring progress, making informed decisions, and achieving desired outcomes. Without baseline data, it would be challenging to determine if interventions are effective or to identify areas that require improvement. The data for this project was received from the client in JSON format. The ultrasound images from the patient's examination were in the .png format. Data standardization involves converting the client files into a suitable working format such as CSV files, excel spreadsheets, etc. Next, data analysis involves reviewing the data for any errors or inconsistencies—the process of data cleansing help with fixing incorrect, incomplete, inaccurately formatted, and duplicated dataset. It was the first step in converting improper data into a meaningful format. As everyone is aware, machine learning is data-driven. Therefore, it was important to process data finely before use. As shown in Figure 1, the cropped ultrasound image data was provided by the client and had comments written by the radiologists.

Figure 1

Breast ultrasound image



It was an important requirement by the client to collect all the comments in a new column and store it along with the additional patient information. The extraction of the comments from the ultrasound image was performed using the python *pytesseract* library. It allowed me to specify the image coordinates while programming and then extract the comment lines.

Python has a pre-programmed toolset that allows efficient data cleansing, named Pandas and Numpy. These libraries offer a diverse range of built-in functions. In prior data cleansing approaches, programmers used to write functions manually. For example, to calculate the average of columns in a dataset, previously, programmers would write the lines of code where they first add up all the available values and then divide them by a total number of values. For the larger mathematical equations, it seemed troublesome. Numpy provides the built-in function to calculate the mean of an entire column by just calling the *np.mean()* function. Once the dataset got ready after the cleansing process, I developed the dynamic report for the client's data. The report is generated in the Jupyter Notebook, an open-source integrated development environment. It provides a web-based interactive computational environment. I used Python libraries to generate descriptive statistics about the data. Descriptive statistics were created in the form of graphs and tables.

Additionally, the PyTorch downloaders have developed the test, train, and validate dataset. They keep the data manageable and help to simplify the machine learning pipeline. I have also defined PyTorch neural network module. After training the model, I evaluated the performance on the test set or validation set. Then, accuracy, precision, and recall were generated to evaluate the model's performance.

1.7 Organization of the Project:

The project is divided into the following phases:

1. Explore and study the breast ultrasound data.
2. Data standardization: Gathering and converting given data into the required format.
3. Data cleansing: Removing inaccurate, duplicate, and incomplete entries and dealing with missing values.
4. Data extraction: Extracting the selected data from the breast ultrasound images and storing them in the dataset.
5. Descriptive statistics: The dynamic Jupyter notebook takes the organized data set from the above steps as inputs and generates the visualizations and the summary tables for the uploaded dataset.
6. PyTorch downloaders: Creating the downloaders for the train, test, and validation sets which keep the data manageable and help to simplify the machine learning pipeline.
7. Neural network performance evaluation.

All mentioned project phases are discussed further in detail in upcoming chapters.

1.8 Significance of the study:

By leveraging a large and diverse dataset of breast ultrasound images and associated metadata, the study has the potential to develop more accurate and effective tools for identifying potentially cancerous lesions in breast tissue. These tools could ultimately improve patient outcomes by enabling earlier and more accurate detection of breast cancer, leading to earlier intervention and treatment. Additionally, the study's development of tools for managing and expanding the image collection could have broader implications for other areas of medical imaging research and other fields that rely on large-scale image analysis, such as computer vision and machine learning.

1.9 Limitations:

- Data availability: The success of this project was highly dependent on having access to a large and diverse dataset of ultrasound images, which may be limited or difficult to obtain. The trained deep learning models may not generalize to new data if the dataset is small or biased.
- Data quality: The quality of the ultrasound images can vary based on many factors, including the skill of the technician, the quality of the equipment, and the patient's body size and composition of poor-quality images can be difficult or impossible to interpret accurately, and may lead to incorrect diagnoses.
- Limited annotation data: Collecting annotation data for each study requires the input of a trained radiologist, which can be time-consuming and expensive. Additionally, not all studies may have annotation data available, which could limit the ability to train deep learning models to predict important image features for cancer diagnosis.

- OCR accuracy: The text on ultrasound images can be difficult to extract accurately using OCR algorithms, especially if the text is in different locations, colors, or fonts. This could lead to errors in the indexing and analysis of the images.
- Legal and ethical considerations: The collection and use of medical data raise important legal and ethical considerations, including patient privacy and consent, data security, and compliance with regulations such as HIPAA. These considerations could limit the scope or feasibility of the project.

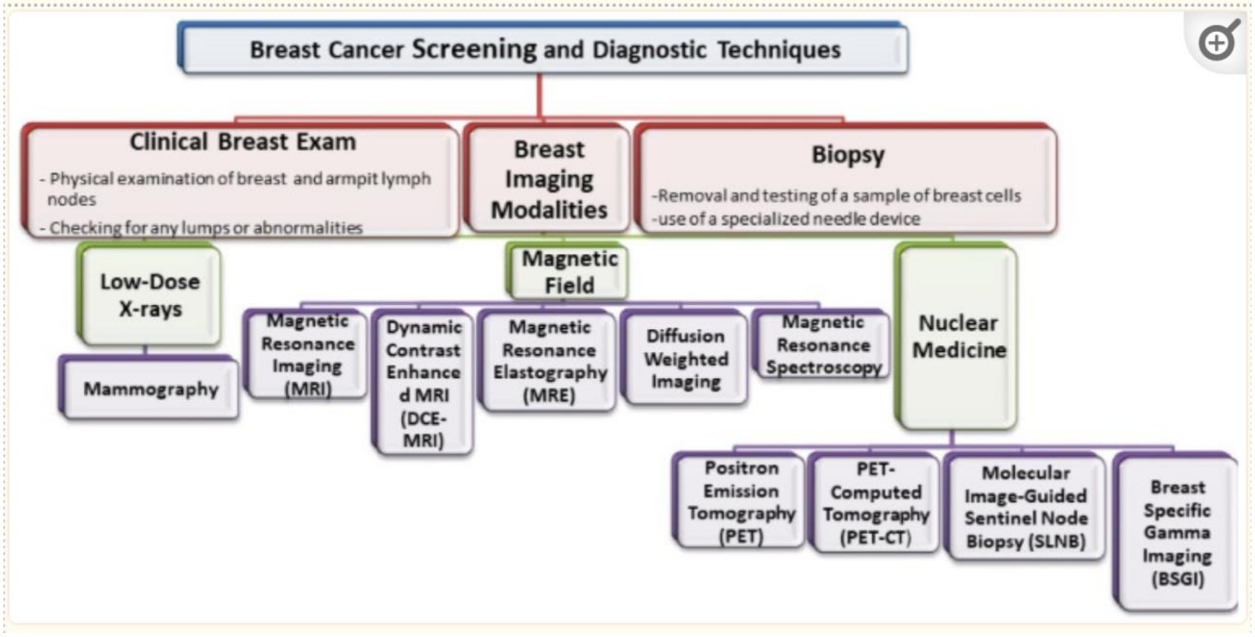
Chapter 2: Literature Survey

2.1 Introduction:

As discussed in the previous chapter, breast cancer severely threatens women's health. The early detection of breast cancer may help with better treatment and cure of the disease. It will also result in increased chances of survival. If the detected tumor is relatively smaller, it is easier to remove with surgery. Sometimes smaller tumors are less aggressive and slower to spread in other body parts (Zheng et al., 2021). Also, if breast cancer is detected early, there are more options for treatment too, such as surgery, chemotherapy, radiation surgery, etc. The study by Bhushan et al. (2021) focuses on the different methods employed in breast cancer diagnosis, which include mammography, ultrasound, MRI, and biopsy. They also cover the different types of breast cancer, including their subtypes and genetic mutations.

Figure 2

Types of imaging techniques for breast cancer diagnosis


Breast Cancer Screening and Diagnostic Techniques
Clinical Breast Exam
- Physical examination of breast and armpit lymph nodes
- Checking for any lumps or abnormalities
Breast Imaging Modalities
Biopsy
-Removal and testing of a sample of breast cells
-use of a specialized needle device
Low-Dose X-rays
Mammography
Magnetic Field
Magnetic Resonance Imaging (MRI)
Dynamic Contrast Enhanced MRI (DCE-MRI)
Magnetic Resonance Elastography (MRE)
Diffusion Weighted Imaging
Magnetic Resonance Spectroscopy
Nuclear Medicine
Positron Emission Tomography (PET)
PET-Computed Tomography (PET-CT)
Molecular Image-Guided Sentinel Node Biopsy (SLNB)
Breast Specific Gamma Imaging (BSGI)

Note: This image is reprinted from Bhushan, A., Gonsalves, A., & Menon, J. U. (2021). Current State of Breast Cancer Diagnosis, Treatment, and Theranostics. Pharmaceutics, 13(5), 723. https://doi.org/10.3390/pharmaceutics13050723

Mayo Clinic uses advanced artificial intelligence (AI) technology and large case studies from Mayo Clinic Enterprise to create cutting-edge software that will aid radiologists in interpreting breast ultrasound lesions. The AI technology will combine deep learning models based on convolutional neural networks, machine learning models, and automated human rule-based models used by expert radiologists. Artificial intelligence (AI) can be very useful in detecting breast cancer. It can analyze mammograms and detect tumors or other areas of concern (Lang et al., 2021). AI can also analyze the patient's medical history. If an individual is identified as at risk, the information can help doctors make more informed decisions and work closely on prevention strategies. AI can provide recommendations for detecting, diagnosing, and treating breast cancer. However, AI needs to be used in conjunction with human expertise. It cannot replace the skills and knowledge of medical professionals (Wu et al., 2019).

The project provides an organized breast ultrasound dataset, which can be further useful in deep learning studies for developing lesion interpretation software for a client. Breast ultrasound datasets can be sensitive as they contain health information about the patients, family history, treatment plans, diagnostic tests, etc. A study by Al-Dhabyani et al. (2020) says that breast ultrasound datasets are required to train machine learning models that can be helpful in the classification, detection, and segmenting of the micro-calcification or prior signs of masses in breast cancer. In addition, researchers who are interested in preprocessing the stages of breast cancer can combine breast ultrasound data with combining other datasets. It can then be used for analysis and further insights generation. The availability of large, diverse, and large datasets allows AI algorithms to learn from more examples and improve the accuracy and reliability in detecting and classifying breast cancer.

2.2 Trends & gaps:

Manual checking of mammography images by a radiologist, also known as visual interpretation, has been the traditional approach to breast cancer detection for many years. This approach involves the radiologist examining the mammogram for any suspicious masses or areas of abnormality that may indicate the presence of breast cancer.

The study by Shah et al. (2022) stated that traditionally, radiologists manually review breast images with the naked eye to diagnose breast cancer. Then, after communicating with medical experts, they will finalize the diagnosis. Although manual inspection is the predominant approach, certain unavoidable factors associated with image inspection can result in erroneous diagnoses and a protracted treatment procedure.

While mammography was an effective screening tool, visual interpretation was not infallible, and there was potential for human error. For example, radiologists miss some breast cancers on mammography, leading to false negative results. Additionally, some abnormalities detected in mammography may be benign, leading to unnecessary follow-up procedures and anxiety for the patient (Kolata, 2023). Next, Reig et al. (2019) discussed one such trend—the increasing use of deep learning models to aid diagnosis. The models in their study have demonstrated impressive results in detecting cancerous lesions in mammography images, and their use is likely to become more widespread as the technology continues to improve.

Another trend in breast cancer diagnosis research was the integration of multiple imaging modalities. For example, Zhu et al.(2020) combined different imaging techniques, such as ultrasound and MRI, to improve diagnostic accuracy. By using multiple modalities, clinicians can better detect subtle differences in tissue structure and identify cancerous lesions that a single imaging modality may have missed.

Thinking about the gaps, Healy (2022) highlighted the lack of standardization in imaging protocols and reporting in her article. The author said this could make it difficult to compare results across studies and hinder the development of reliable and consistent deep-learning models. Standardization could help ensure that imaging data is consistent and can be used to develop accurate models that can be applied across multiple centers and populations.

Fenton et al. (2013) studied the need for a better understanding of effectively integrating deep learning models into clinical practice. While these models can improve diagnostic accuracy, ensuring they do not result in unnecessary biopsies or missed diagnoses is important. The research was needed to understand how to implement these models and integrate them into clinical workflows effectively.

2.3 Research opportunities:

The article by Colangelo (2023) discussed the use of deep learning to improve the accuracy of breast cancer diagnoses for different subtypes of the disease. The author notes that breast cancer is a complex disease with many subtypes and that deep learning algorithms could be used to distinguish between these subtypes and provide more accurate diagnoses.

New opportunities for research in breast cancer diagnosis using deep learning models include developing novel methods to collect and annotate ultrasound images to improve the quality and availability of training data. Standardizing imaging protocols and reporting can also enhance the consistency of results across studies and facilitate comparing outcomes. The ethical and social implications of using deep learning models for breast cancer diagnosis must be explored, focusing on privacy, equity, and patient autonomy. Shared decision-making tools and decision aids may be promising approaches to achieving this goal.

2.4 Methodologies:

Several methodologies were used for the preparation of the breast ultrasound dataset. First, Richey et al. (2020) use optical character recognition (OCR) to enable the detection and localization of textual markers (fiducials) that are placed on the surgical site to guide the surgeon during the procedure. These fiducials are often small and difficult to see with the naked eye, especially once covered by surgical drapes or obscured by blood or tissue. OCR technology can read the fiducial markers and translate them into machine-readable data, which can then be used to register the surgical site with preoperative images and guide the surgeon during the procedure. This can enhance the accuracy and precision of surgical navigation and help ensure complete tumor removal while minimizing damage to healthy tissue.

Optical character reorganization (OCR) is essential for analyzing breast cancer images. It identifies and converts texts that appear on the images by using pixels into more machine-friendly representations. OCR can be used to extract data from the breast ultrasound image. As a result, researchers can more adequately identify patterns, trends, and features/abnormalities that may indicate cancer. OCR mainly extracts the text and converts the input text into the machine-encoded format.

Researchers documented the earliest development of the OCR system in the 1940s. With technological advancements, the system became proficient in processing handwritten and printed characters. At the 1965 World Fair in New York, "IBM 1287" was unveiled, which was the first optical reader capable of converting handwritten text into machine-encoded data (Mori et al., 1992). During the 1970s, researchers focused on enhancing the performance and response time of the OCR system.

Fast forward to 2015, OCR is openly accessible. It captures the text from pictures and further provides the facility of indexing and editing. Genzel et al. (2015) determined that OCR still has some limitations while dealing with noise in text recognization and if the lower quality images are provided as input. Nevertheless, the current OCR works great for high-resolution images, clear scanned materials, and regularly used typefaces.

Chen et al. (2022) use PyTesseract, a python wrapper for Google's Tesseract-OCR engine, to extract relevant information from unstructured clinical text data. The reason for using PyTesseract is to convert images or scanned documents containing clinical notes into machine-readable text that can be analyzed using natural language processing techniques. This is particularly useful in breast cancer outcomes research, where large volumes of clinical data must be processed and analyzed to identify trends, outcomes, and predictors of treatment success.

PyTesseract allows researchers to extract relevant data quickly and accurately from clinical notes, such as patient demographics, tumor characteristics, treatment details, and follow-up information, without manual chart review or data entry. This can save time and resources, reduce errors, and improve the efficiency and reproducibility of outcomes research.

PyTesseract is an open-source platform for OCR. As explained in the previous paragraphs, OCR extracts characters or text data from images. However, the tesseract is extremely flexible and can be adapted to work with various languages, including multilingual texts. Adjetey et al. (2021) used the tesseract OCR engine for their research. According to them, the main objective of PyTesseract OCR is to identify characters within an image, which are then extracted and stored. The extracted characters are grouped to form words and sequences of characters without any white space. Each word is associated with a confidence level that indicates the recognized word's accuracy.

In the context of breast cancer classification, Sothivelr (2020) uses PyTorch to train a deep convolutional neural network (CNN) on a dataset of mammography images. The purpose of training the convolutional neural network is to use the learned features from the breast ultrasound images to classify them as either benign or malignant. PyTorch allows the author to easily define and optimize the neural network architecture, preprocess the image data, split the data into training and testing sets, and monitor the model's performance during training.

Understanding PyTorch workflow is important to create and training neural networks. Novac et al. (2022) describe in their study that PyTorch follows three key principles, the first of which is the method used to define its functions. Its components are defined in a pythonic manner, making it easy for users familiar with Python to use the framework. The second principle aims to simplify defining components by using interfaces. This leads to a better understanding of essential neural networking concepts, making the learning curve less steep. Lastly, the framework values simplicity over complexity, even if it means sacrificing some performance. Maintaining a simple design enables faster resolution of potential issues.

Pytorch is one of the more efficient deep learning frameworks currently available. A study by Jiang et al. (2021) performs image classification and found that the Pytorch graph structure is simple to grasp and, most significantly, it is convenient for researchers to debug. Researchers can utilize the Pytorch framework for constructing convolutional neural network models quickly and effortlessly and subsequently training them on vast data sets.

Ultrasound image classification can be performed using various pre-trained models. Chandrakesan (2021) used the ResNet50, a pre-trained deep learning model trained on a large dataset of images, and achieved state-of-the-art performance on the ImageNet dataset. The model was capable of accurately classifying images into one of 1,000 different categories, such as "dog,” "cat,” "car,” "flower,” and so on.

2.5 Summary:

The early detection of breast cancer is crucial for better treatment and higher chances of survival. The utilization of AI to examine mammograms and the medical background of patients has the capacity to assist in identifying, diagnosing, and treating breast cancer. However, it is important to use AI in conjunction with human expertise as it cannot replace the skills and knowledge of medical professionals. Large, diverse, and organized breast ultrasound datasets are required to train machine-learning models for accurate and reliable breast cancer detection.

The current state of knowledge around breast cancer diagnosis is rapidly evolving, with a growing emphasis on using deep learning models and integrating multiple imaging modalities.

However, there are still significant gaps in our understanding of effectively integrating these models into clinical practice and ensuring that they do not result in unnecessary biopsies or missed diagnoses. There is also a need for more research on how to effectively incorporate patient preferences and values into decision-making around breast cancer screening and diagnosis. Nevertheless, the opportunities for new research in this area are vast, with the potential to significantly improve the accuracy and effectiveness of breast cancer diagnosis and ultimately save lives. This field of study is important and warrants ongoing attention and investment.

The methods used to deliver the organized breast cancer dataset, which can be used for deep learning studies and to understand the patterns of ultrasound image labeling, include OCR, PyTesseract, and PyTorch. Optical character recognition (OCR) plays a significant role in analyzing breast cancer images and extracting text data from them. PyTesseract OCR, an open-source platform, is flexible and can be adapted to work with various languages, including multilingual texts. In addition, PyTorch, a deep learning framework, is efficient and convenient for researchers to construct convolutional neural network models and train them on vast datasets.

Chapter 3: Methodologies

3.1 Introduction

Ultrasound imaging is a commonly used diagnostic tool for detecting breast cancer. Radiologists typically examine a series of ultrasound images focused on a specific area of breast tissue, referred to as a lesion. Based on their analysis of the ultrasound images, radiologists need to determine whether it is necessary to recommend a biopsy of the tissue to detect the presence of cancer. Making a recommendation for a biopsy is a complicated process that considers multiple factors, such as the size and properties of the lesion, the patient's medical background, and other clinical observations. Ultimately, ultrasound imaging aims to improve breast cancer detection accuracy and provide patients with the best possible care (Pan, 2016; Saul, 2010).

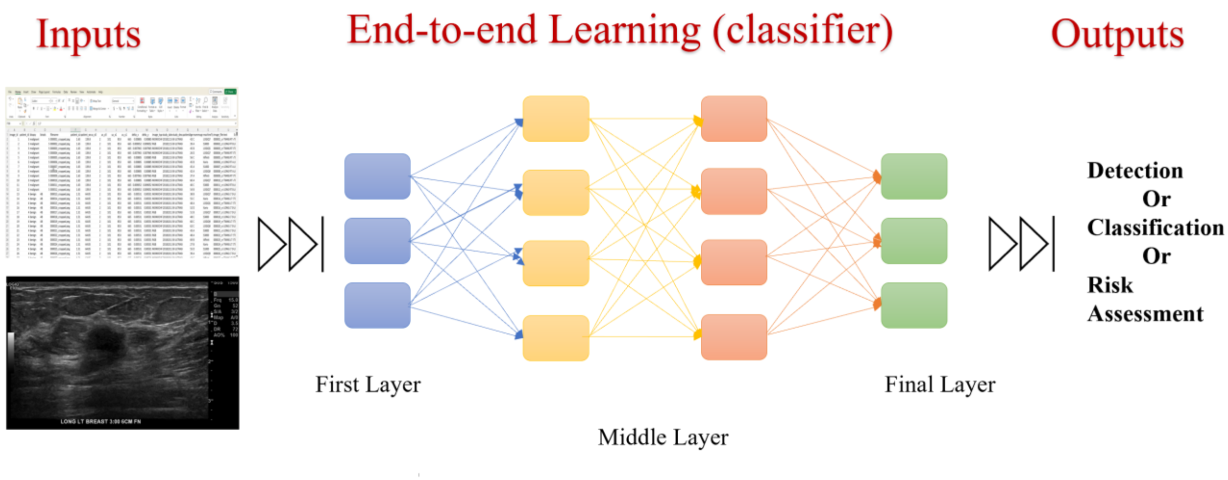
The primary goal of this project was to support a broader initiative focused on creating decision-making software that employs deep learning models for forecasting the existence of cancerous growths in ultrasound images. In addition, this project focused on building the tools for gathering, organizing, updating, and summarizing ultrasound imaging data that will be utilized to train deep-learning models. The dataset used for the project includes breast ultrasound imaging studies and supplementary annotations that radiologists have added to ultrasound images to facilitate software development.

The project structure started by developing software for an ultrasound image database that could efficiently process incoming batches of data and integrate them into the image collection database. Next, the PyTorch downloaders were implemented, which served as the interface between the deep learning models and master image collection. Lastly, the neural network classification model was trained to predict the malignant or benign for each image. It represented a robust, organized database that was configured well, and the downloaders could interface with the collection correctly. Geras (2019) discovered that combining AI and human radiologists can lead to more accurate breast cancer detection, potentially reducing the number of unnecessary biopsies and improving patient outcomes.

In the classification model, the neural networks function similarly to the neurons in the human brain, as computing systems composed of interconnected nodes. Using algorithms, they can identify concealed patterns and correlations within raw data, group and categorize them, and gradually enhance and refine their learning ability. Figure 3 shows the overall structure of the deep learning classification model.

Figure 3

Schematic design of end-to-end deep learning network



Alanazi et al. (2021) applied a Convolutional Neural Network (CNN) to train the classifier for distinguishing between benign and malignant breast tumors using mammogram images. The CNN model was trained through transfer learning techniques. A pre-trained CNN architecture (ResNet-50) was used as the base model. The final classification layer was replaced with a new one for the binary classification of benign and malignant tumors.

3.2 Research design

The research methods used in this project involve the combination of data collection, data processing, data management, software engineering, and deep learning. These methods are used to build tools to collect, manage, update, and summarize ultrasound imaging data. In addition, the resulting organized database will be used to train deep-learning models to detect cancerous lesions in breast ultrasound images. These methods were well-suited to the project goals. Collecting, processing, and managing data are essential for building any decision-making software. Software engineering is crucial for designing and implementing effective tools to manage and process data efficiently. Additionally, deep learning is a promising technology that has shown remarkable success in image recognition tasks, making it a suitable choice for detecting cancerous lesions in ultrasound images. The combination of these research methods provides a comprehensive approach that can help achieve the project's objectives efficiently and effectively.

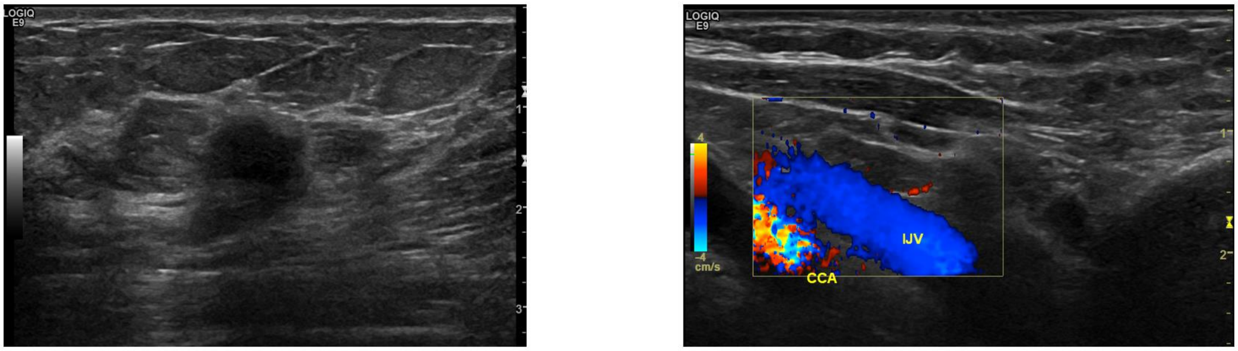
The specifics of research methods involved in the project are broken down into the following steps:

Data Survey: The breast ultrasound images consisted of two formats. The first was a plain ultrasound, and the second was an ultrasound with a doppler. Figure 4 shows the example of different types of images, i.e., a plain ultrasound image and the ultrasound image captured by the doppler. High-frequency sound waves are utilized in plain ultrasound imaging to generate images of the internal organs and structures in the body. On the other hand, a doppler ultrasound is used to evaluate blood flow in the body. It measures the frequency shift of sound waves bouncing off moving red blood cells to create images of blood vessels and blood flow patterns. This allows for detecting blockages or narrowing of blood vessels and abnormal blood flow patterns such as turbulence or reversal.

The article by Bard et al. (2021) highlights the use of doppler imaging as a technique that can help overcome this issue by providing additional information about blood flow in the breast tissue. This can help identify areas of increased blood flow, which may indicate a tumor.

Figure 4

Types of ultrasound images



Plain Ultrasound

Ultrasound with Doppler

Data manipulation: The image database software developed in this project extracted the annotated text from each image using OCR. First, annotations were parsed by the system and added to the master index. In the next step, I processed the batches of data as they became available and added them to the image collection. The new information was added to the corresponding rows in the collection for each new batch of annotation data. Each ultrasound image row contained information about patient details such as age, size, biopsy results, study descriptions, etc., along with the annotations collected from ultrasound images. Crucial information, such as scanning region of the tumor, the position of the lesion from nipple, and whether a tumor was located on the left or right side of the breast, were included in the annotations.

Data management: The master image collection was structured to allow additional information or annotations to be added to the existing collection later, as determined by the needs of the decision support software. The tools were designed to allow for the inclusion of additional annotations that could be collected later.

Software engineering: The programming in this project allowed for simple filtering of the image data to select relevant subsets of the collection. Next, the PyTorch downloaders were written to retrieve new batches of study and annotation data as they became available. Lastly, the code was developed to allow for corrections of entries in the collection.

Deep learning: In the last stage of the project, I trained deep learning models on the collected ultrasound imaging data to predict whether there was a potentially cancerous lesion shown in the ultrasound images. The software was designed to predict whether a patient's study contained a cancerous lesion and explain how it reached this decision.

3.3 Population and sample:

Target population: The target population for this project was female breast cancer patients aged 16 years and above diagnosed with lesions in their breast tissue. Patients were excluded whose breast ultrasound images met the following exclusion criteria:

- Patients younger than 16 years old
- Male patients

- Patients with empty breast ultrasound images (no lesion visible)
- BI-RADS 6: If a lesion is classified as BI-RADS 6, it implies that the imaging characteristics highly indicate malignancy and a biopsy is necessary for confirmation.

Sample: For this study, a total of 4099 ultrasound images were collected from a sample of 362 female patients, with an average of 11-15 images per patient. The sample was selected using a convenience sampling method. Patients who met the inclusion criteria, such as patients with a single breast ultrasound exam per day, patients with definitive breast biopsy pathology reports, etc., were recruited from a single hospital.

Data source: This project only contained the ultrasound images and patient data provided by the Mayo Clinic. The ultrasound images were taken as part of routine clinical care and for the patients who found suspicious lumps or abnormalities during a physical breast exam, a mammogram, or a breast MRI. They were provided by the hospital where the patients were treated.

Criteria for data selection: The ultrasound images were selected based on the presence of lesions in the breast tissue. Only images that clearly showed the lesions were included in the study. Images of poor quality or did not clearly show the lesions were excluded from the study. Additionally, only patients who had undergone a biopsy to confirm the presence of cancerous cells in the lesions were included in the study.

3.4 Bias handling:

The criteria for data selection included the presence of lesions in the breast tissue and the exclusion of images of poor quality or did not clearly show the lesions. Patients who had undergone a biopsy to confirm the presence of cancerous cells in the lesions were included in the study. The exclusion criteria included patients with ages below 16 years, male patients, empty breast ultrasound images, multiple ultrasound exams per patient per day, a number of breast ultrasound images per exam that were either too small or too large, discordant cases, BI-RADS 6 exams, images with burnt-in pixel information, and non-definitive breast biopsy pathology report.

One potential type of bias that could arise in the study is selection bias, which occurs when the sample is not representative of the target population. I used a convenience sampling method to recruit patients from a single hospital to mitigate this bias. While this method may have limitations in terms of generalizability, it ensured that the sample was drawn from a homogenous population of breast cancer patients who received treatment at the same hospital. Additionally, the criteria for data selection were clearly defined and followed consistently, which reduced the risk of bias in the selection of images. Moreover, I consulted with radiologists to ensure that the ultrasound images were of high quality and accurately identified the lesions, which helped minimize bias in the data collection process.

3.5 Data Collection, analysis, and tools:

The client provided the unstructured data in the form of a JSON file. It collected all patient information, such as patient ID, age, size, lesion information, lesion type, machine type, which is used to capture the lesion information, and many more. After discussing with the client, the required data were collected separately to process further.

Patient study data:

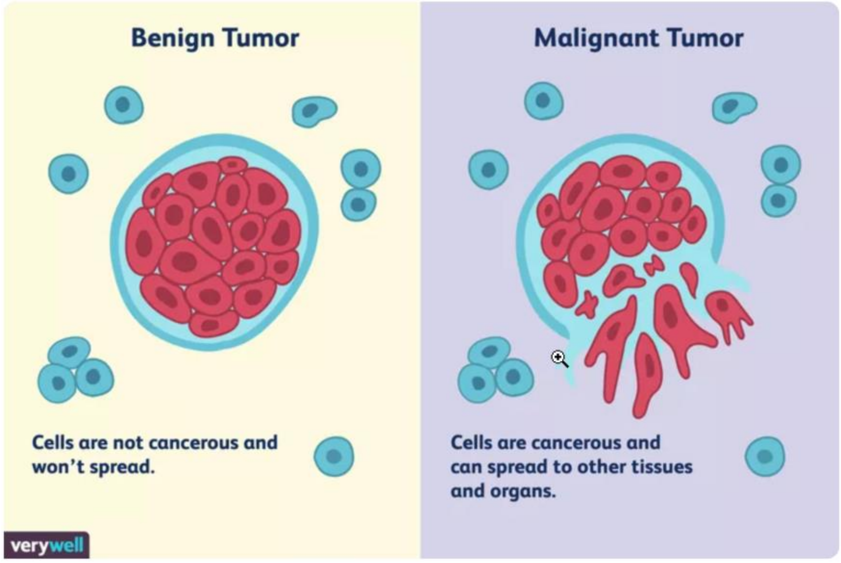
There were 2 types of unstructured data provided by the Mayo Clinic. As mentioned above, the JSON file contained the patient study data. The other folder had all patient's ultrasound image data. The following steps were taken place in order to output the organized database.

- Processed the JSON file to extract labels and metadata. Added it to the master index file.
- Copied the breast ultrasound images to the correct places in the collection.
- Checked for the distinct image and patient IDs from those of previously added data.
- Allowed to overwrite the existing data in case of correction for ultrasound images and patient data in an index file.
- Enabled the possibility of adding additional columns to the index file for additional data that might be extracted from the studies later.
- The relevant text from each image was extracted using OCR. Next, I parsed it and added it to the master index.
- Each ultrasound image had its row in a master index. The row contained patient ID, patient age, patient size, biopsy labels, Breast Imaging Reporting and Data System known as BI-RADS, lesion regional axis locations, image type (doppler image or plain ultrasound image), mammographic breast density labels, machine types, and extracted annotated data from an ultrasound image.

Biopsy labels were noted as benign or malignant. The article by Splane B. (2022) explains the key differences between benign and malignant tumors. The visual representation presented in Figure 5 showcases the distinct types of cancer tumors. Benign tumors are non-cancerous growths that neither infringe upon the adjacent tissues nor spread to other body parts. Conversely, malignant tumors denote cancerous growths that can infiltrate and impair the neighboring tissues and organs and propagate to other body parts via metastasis. Comprehending the contrasting features of benign and malignant tumors enables individuals to make informed decisions regarding their healthcare and treatment alternatives.

Figure 5

Visual representation of benign and malignant tumor


Benign Tumor
Malignant Tumor
Cells are not cancerous and won't spread.
Cells are cancerous and can spread to other tissues and organs.
verywell

Note: Figure 5 is reprinted from Splane, B. (2022, October 17). What Is a Benign vs. Malignant Tumor? *Verywell Health*. https://www.verywellhealth.com/what-does-malignant-and-benign-mean-514240

BI-RADS was another important entity stored for each patient. Breast imaging studies had classified into seven assessment categories based on the BI-RADS system:

BI-RADS 0: *Incomplete* - Further assessment of the imaging may be required, which could involve obtaining previously unavailable images during the initial reading.

BI-RADS 1: *Negative* - No masses, suspicious calcifications, architectural distortion, or symmetrical features were observed.

BI-RADS 2: *Benign* - There is a 0% possibility of malignancy.

BI-RADS 3: *Probably* benign - The possibility of malignancy is under 2%, and it is recommended to schedule a short-interval follow-up.

BI-RADS 4: *Suspicious for malignancy* - There is a probability of malignancy ranging from 2-94%, which is divided into:

- BI-RADS 4A indicates a low suspicion for malignancy with a 2-9% probability range.
- BI-RADS 4B suggests a moderate suspicion of malignancy with a probability range of 10-49%.
- BI-RADS 4C indicates a high suspicion of malignancy with a probability range of 50-94%.
- A biopsy should be considered if the mammogram results fall under BI-RADS 4B or 4C.

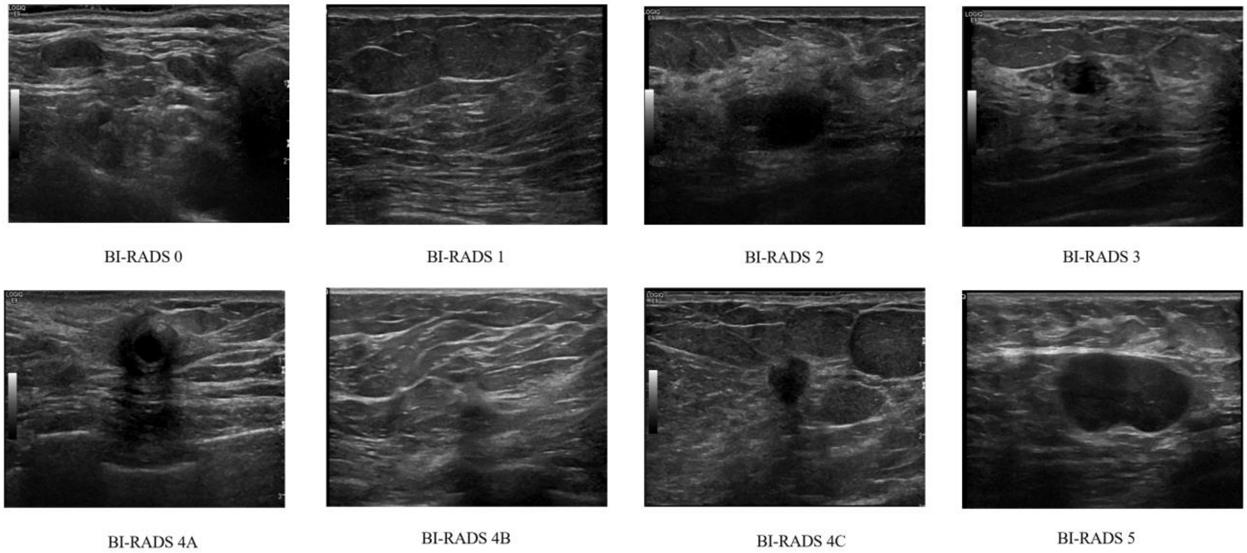
BI-RADS 5: *Highly suggestive of malignancy* - If the probability of malignancy is higher than 95%, it is necessary to take appropriate action.

BI-RADS 6: *Known biopsy-proven malignancy* – Biopsy confirmed a malignancy.

As discussed in section 3.3, patients with the BI-RADS 6 had been eliminated from the studies. Figure 6 illustrates how ultrasound images for all the above-mentioned BI-RADS categories appear on ultrasound imaging.

Figure 6

Ultrasound images for BI-RADS categories



The study by Mercer et al. (2014) states that the visual analog scale is a reliable and practical method for assessing breast density and is acknowledged for its ability to measure it continuously. Nevertheless, visual analog scales have been criticized due to issues such as low consistency among observers. Additionally, discrepancies in breast density estimation terminology have led to errors in classification. In an effort to establish uniformity in mammographic reporting, the BI-RADS lexicon was created, and it has been revised to include categories A, B, C, and D.

The article by Kolata (2023) offers insights into categorizing breast density. According to the article, breast density is classified into four distinct categories under BI-RADS, namely:

Category 1 (A) - Breast tissue is entirely composed of fatty tissue.

Category 2 (B) - Breast tissue comprises scattered fibroglandular densities.

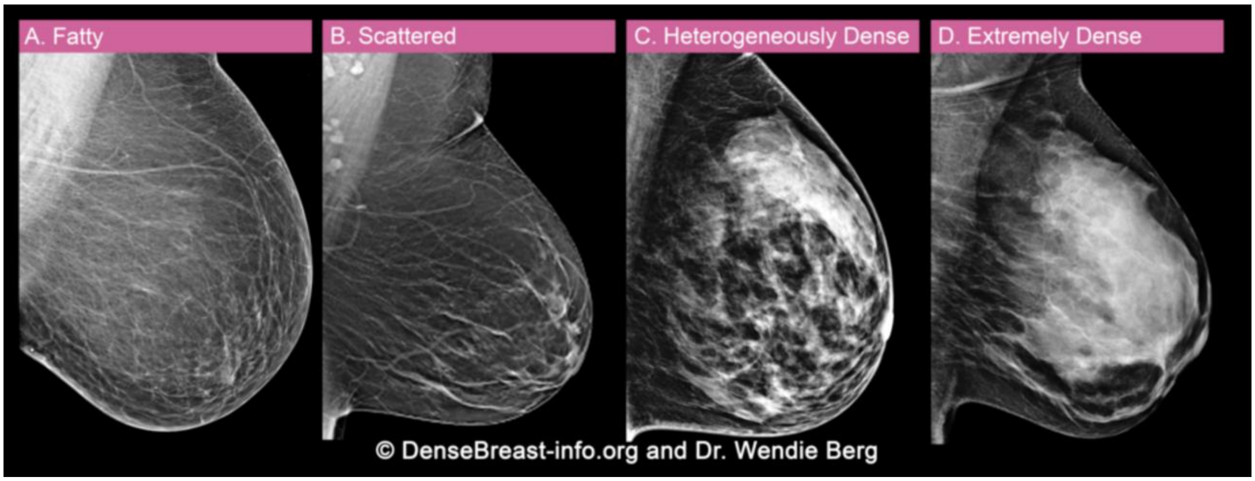
Category 3 (C) - Breast tissue features heterogeneously dense regions.

Category 4 (D) - Breast tissue is extremely dense.

The article explains what each category means, such as that category 1 indicates almost all fatty tissue while category 4 indicates almost all dense glandular and fibrous connective tissue. Figure 7 represents visualization for the mammographic breast density labels. In the current project, we stored all above mentioned categories for mammographic breast density labels for each patient.

Figure 7

Mammographic breast density labels



Note: Figure 7 is reprinted from Dense Breasts: Answers to Commonly Asked Questions. *National Cancer Institute*. Retrieved from https://www.cancer.gov/types/breast/breast-changes/dense-breasts#:~:text=The%20four%20breast%20density%20categories,be%20extremely%20dense%720(D).

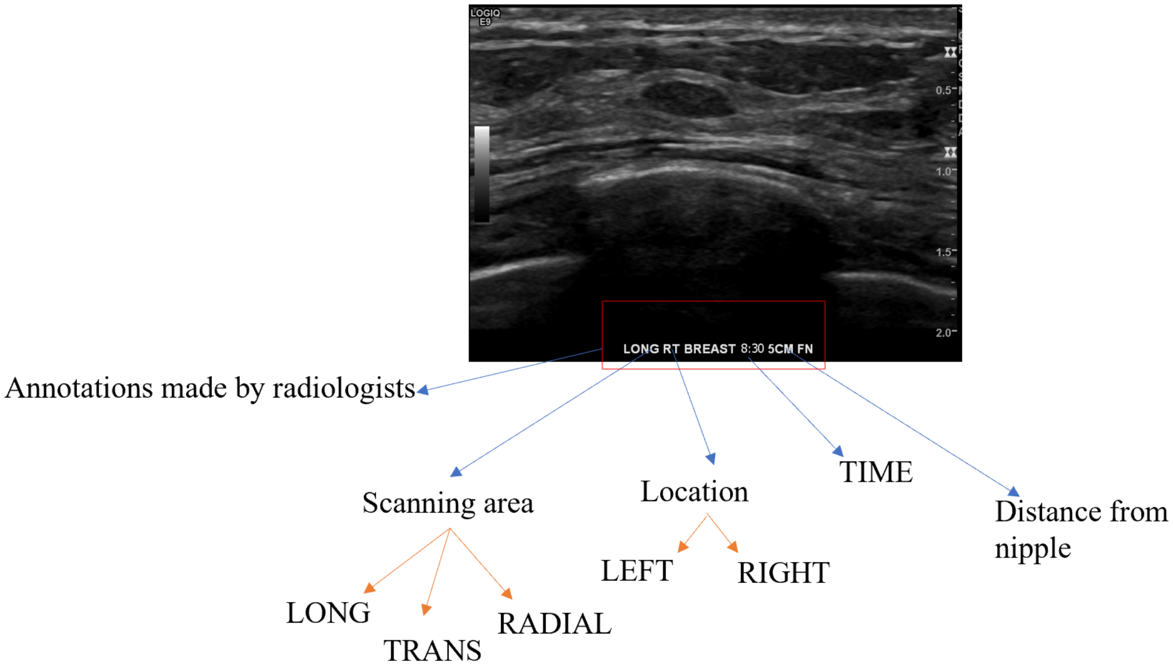
Ultrasound imaging data:

The breast ultrasound images consisted of annotations by the radiologist. Radiologists annotated the ultrasound images, which gave information about the location, size, and characteristics of structures of interest, which helps in the accurate diagnosis and treatment of medical conditions. In addition, annotations are a valuable tool for communication and precision during medical procedures.

Figure 8 illustrates how annotated data from ultrasound images were stored in a database. The figure shows that the annotated data was organized into different fields, such as scanning area, location, time, and distance from the nipple. Each field contained specific information about the ultrasound image annotated by radiologists, which can be used for further analysis and research.

Figure 8

Annotation by radiologist: a detailed explanation



In breast ultrasound imaging, "long", "trans", and "radial" describe the orientation of the ultrasound probe relative to the breast tissue being imaged. "Long" corresponds to the long axis of the breast, "trans" is perpendicular to the long axis of the breast, while "radial" is perpendicular to the line connecting the nipple and the chest wall. These terms help radiologists describe the location and orientation of lesions or abnormalities they detect during the examination.

Next, for each new batch of ultrasound images:

- The annotated data consisted of text descriptors for each image and was stored accordingly.
- Each annotation allowed for the collection to be updated.
- The tools used during the study allowed for the inclusion of additional annotations that could be collected at a later time if needed.

Tools:

The project was focused on building, correcting, and expanding the breast ultrasound patient data and their image collection. Python code was developed to accomplish the following tasks:

- The first task was to write code that allowed for simple filtering of the image data to select relevant subsets of the collection. This code took the master index file as input and produced a smaller index file for only the selected studies.
- The second task was to write downloaders as an interface between the deep learning models and the master images collection (or a filtered subset).
- Finally, I trained a classification model to predict whether each image was malignant or benign. This was done to show that the collection was configured well and that the downloaders could interface with the collection correctly. Different python libraries and frameworks, such as OCR, PyTesseract, CV2, PIL, PyTorch, neural network, etc., were used to complete the tasks mentioned above.

Maclary D. (2019) discussed the relative popularity of R and Python among data scientists. According to him, Python has been gaining popularity among data scientists and is now the language of choice for many. Python is particularly well-suited to machine learning and artificial intelligence applications because powerful libraries like TensorFlow and PyTorch are available. Therefore, although the R language had its strengths, Python was more suitable for this project.

The breast ultrasound patient data may be imbalanced, meaning there can be many more examples of one class (e.g., benign) than the other (e.g., malignant). The study by Batista et al. (2004) determined that an algorithm like KNN can be sensitive to imbalanced data because it relies on the distance between the nearest neighbors, which may be biased towards the majority class. Breast ultrasound images are typically high-dimensional, meaning that they contain a large number of features or pixels. KNN can suffer from the curse of dimensionality, which means that the distance between the nearest neighbors becomes less informative as the number of dimensions increases. This can lead to poor performance and high computational costs.

The relationship between the features in the breast ultrasound images and the target variable (malignant or benign) can be nonlinear, meaning that a linear classifier may not be able to capture this relationship effectively. More complex models, such as neural networks or decision trees, may be more suitable in this case.

Chapter 4: Results

The study aimed to analyze a collection of medical images to determine if a deep-learning model could accurately classify them as either malignant or benign. To achieve this goal, developing an organized breast imaging database was important. The dataset contained every small detail of the ultrasound image along with the patient details. I employed several methods, including filtering the image data to select relevant subsets, creating downloaders to interface with the data collection, and training a classification model using various Python libraries and frameworks.

To ensure the accuracy of the deep-learning classification model, the collected data was thoroughly reviewed and annotated. The annotation process was conducted by expert radiologists trained to identify and mark cancerous lesions in breast tissue. This annotation helped ensure that the data was of high quality and could be used to train the deep-learning model effectively.

Once the data was collected, it was preprocessed to extract relevant features from the ultrasound images. This involved using image processing techniques to filter and enhance the images and selecting relevant regions of interest (ROIs) within the images.

A deep-learning model was then trained using the preprocessed data. This involved using various Python libraries and frameworks to create a convolutional neural network (CNN) that could accurately classify the images as either malignant or benign. The model's performance was evaluated using various metrics, including specificity, accuracy, and recall.

The study results showed that the deep-learning model could accurately classify the images as either malignant or benign, with an accuracy of 65%. This suggests that deep-learning models can be trained using ultrasound imaging and annotation data to predict the presence of cancerous lesions in breast tissue.

This study demonstrated the importance of collecting and managing high-quality ultrasound imaging data to train deep-learning models for breast cancer diagnosis. Developing a well-organized database and employing advanced image processing techniques for text extraction from images makes it possible to achieve a significant level of accuracy in the classification of breast cancer images.

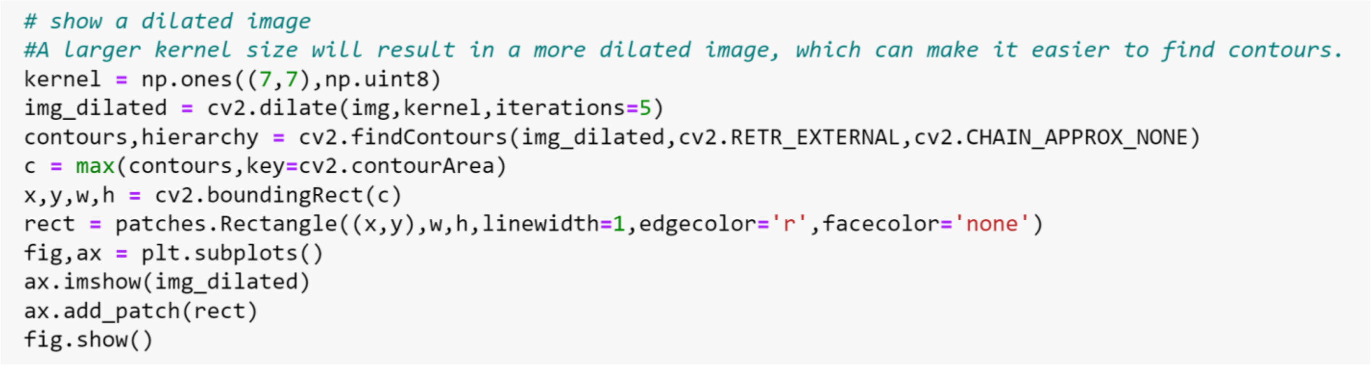
The Mayo breast ultrasound database consisted of 4099 images from 362 patients' ultrasound exams, and patients were examined between 2018 and 2019. Mayo Clinic radiologists worked with skilled technologies to efficiently provide imaging services. They used advanced machines for the screenings. The images were originally stored per the standards of DICOM (digital imaging and communications in medicine).

4.1 Text extraction

The ultrasound images contained the annotations by the radiologist. It was important to extract and store them into a dataset.

Figure 9

Image dilatation and cropping for text extraction

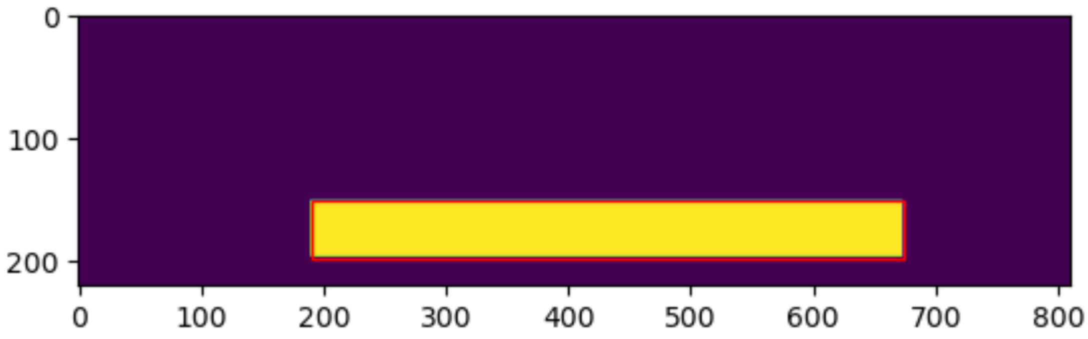


```
# show a dilated image
#A larger kernel size will result in a more dilated image, which can make it easier to find contours.
kernel = np.ones((7,7),np.uint8)
img_dilated = cv2.dilate(img,kernel,iterations=5)
contours,hierarchy = cv2.findContours(img_dilated,cv2.RETR_EXTERNAL,cv2.CHAIN_APPROX_NONE)
c = max(contours,key=cv2.contourArea)
x,y,w,h = cv2.boundingRect(c)
rect = patches.Rectangle((x,y),w,h,linewidth=1,edgecolor='r',facecolor='none')
fig,ax = plt.subplots()
ax.imshow(img_dilated)
ax.add_patch(rect)
fig.show()
```

As shown in Figure 9, a kernel was created and applied to an image. The resulting image was then used to find the external contours. The contour with the largest area was identified, and its bounding rectangle was calculated. A rectangle was drawn around this contour. Figure 10 represents the resulting image with the bounding rectangle displayed.

Figure 10

Bounding rectangle over annotation

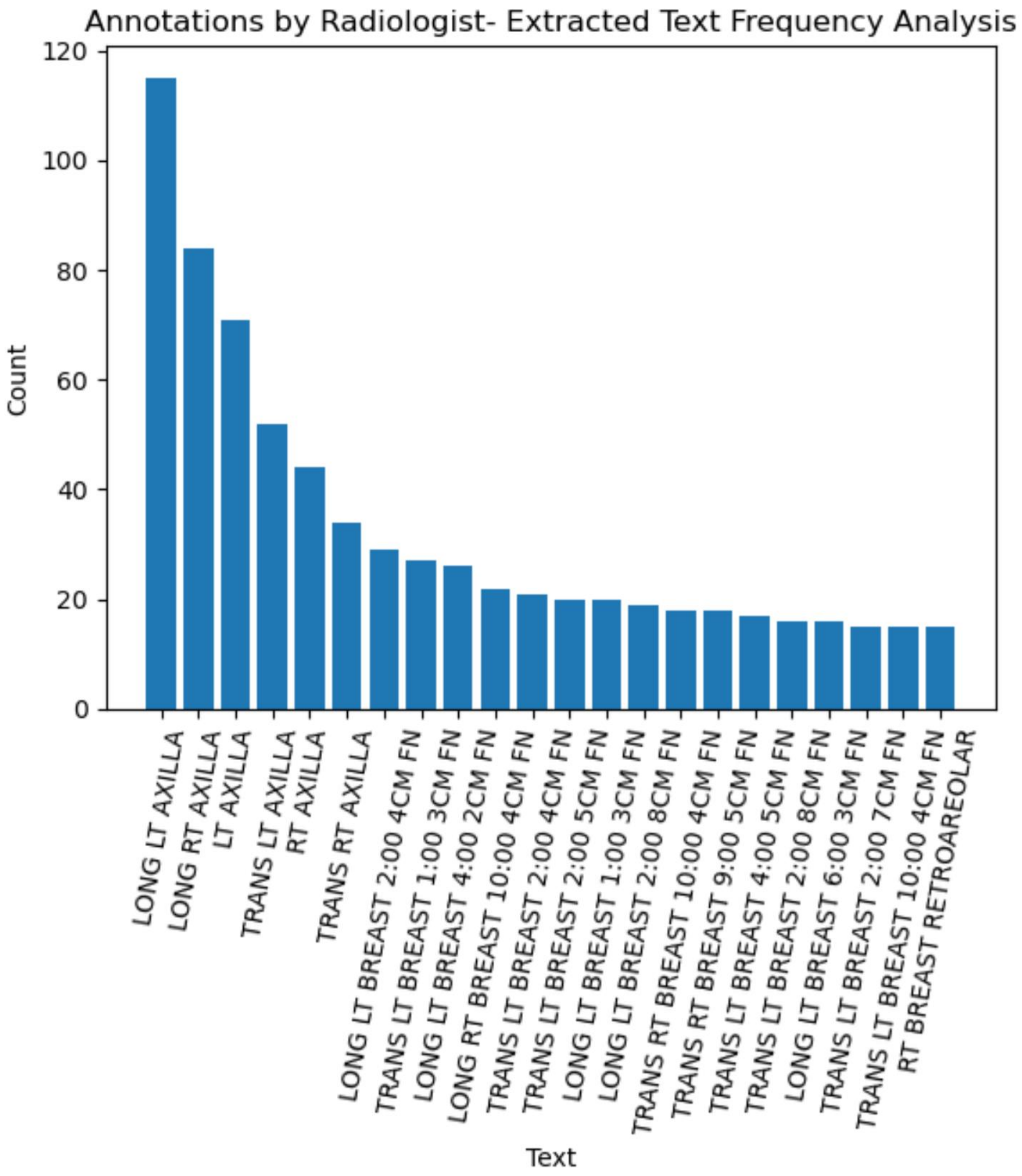


Pytesseract and EasyOCR were used for text extraction from ultrasound images in English. However, there were some images where the annotation text was in a light color which Pytesseract could not retrieve accurately. In such cases, EasyOCR performed better and could extract the text more accurately. After extracting text from the ultrasound images using Pytesseract and EasyOCR, the extracted text was processed using the regular expression library to remove noisy data and special characters. The logic was implemented to keep already cleaned text as it is and clean only the uncleaned text. To evaluate the accuracy of the extracted data, exact matches were counted between the raw text data and cleaned text data, resulting in 83% accuracy.

Frequency analysis was conducted, as shown in Figure 11, to determine how often each annotation was used in the extracted text. The resulting bar graph showed the most frequently used annotation at the top and the least at the bottom. The bars decrease in height gradually as the frequency of the annotations decreases. This analysis provides a clear picture of the distribution of annotations in the extracted text.

Figure 11

Frequency analysis

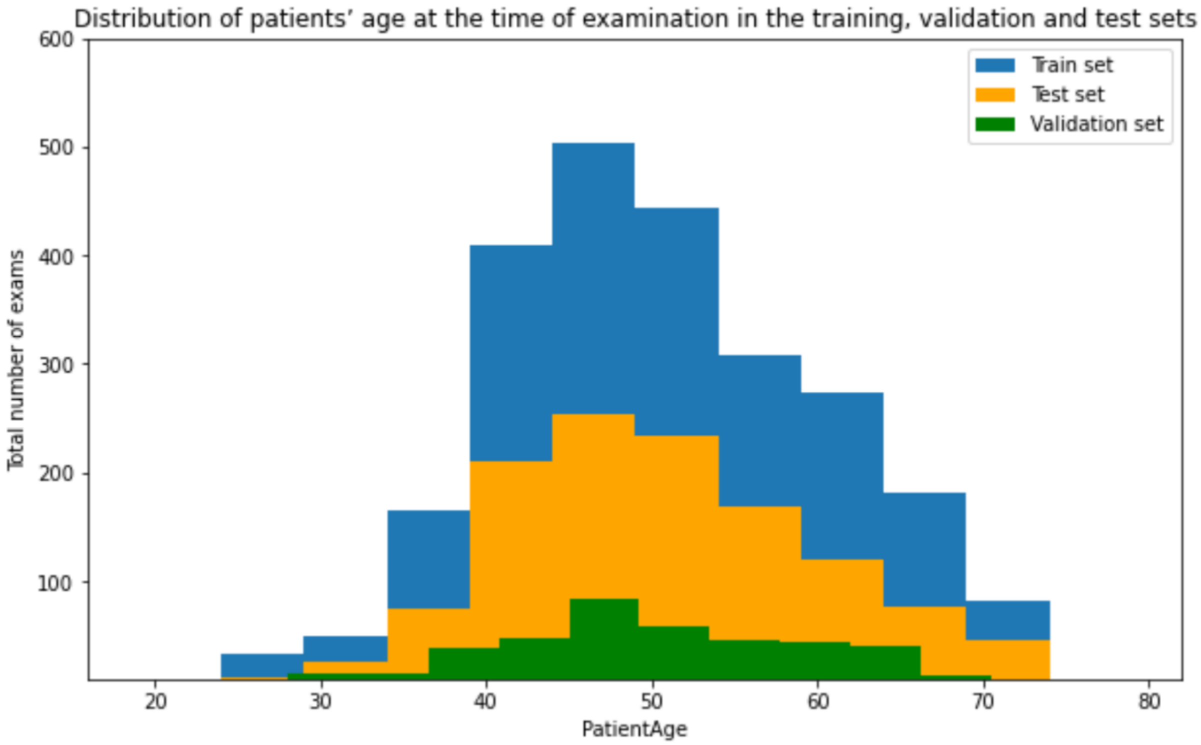


4.2 Statistics:

The organized dataset was ready with all the patient study and annotation data. Next, the dataset was split into three subsets for training the classification model: the training, test, and validation sets. The split was performed in a 60:30:10 ratio, respectively. The data was split into subsets based on each patient's unique ID to prevent biased results. This ensured that the same patient's data did not appear in the training and test sets, which would have skewed the results if the data had been split randomly. Figure 12 demonstrates that at the time of the ultrasound exam, the patient's age ranged between 35 to 65, with a mean age of 55.

Figure 12

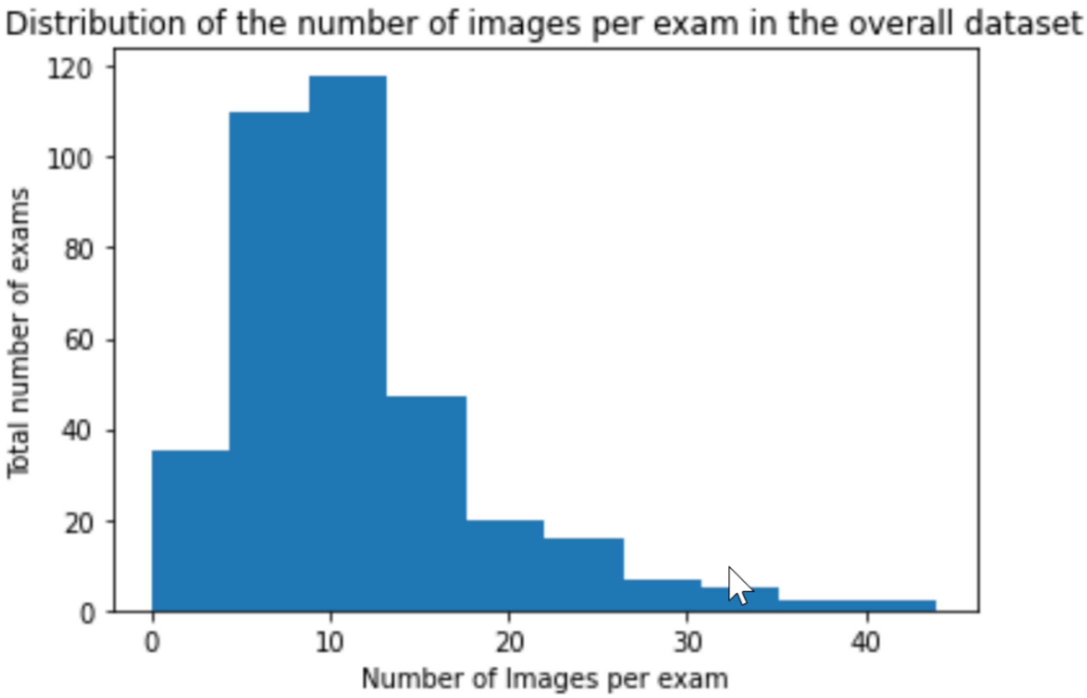
Patient age distribution in the training, validation, and test sets



As discussed earlier, there were a set of 4099 ultrasound images used in this study. The number of images per exam ranges between 5 to 20, with an average of 11 or 12 images. Figure 13 shows the distribution of the dataset's number of images per exam.

Figure 13

Number of images per exam distribution in the entire dataset

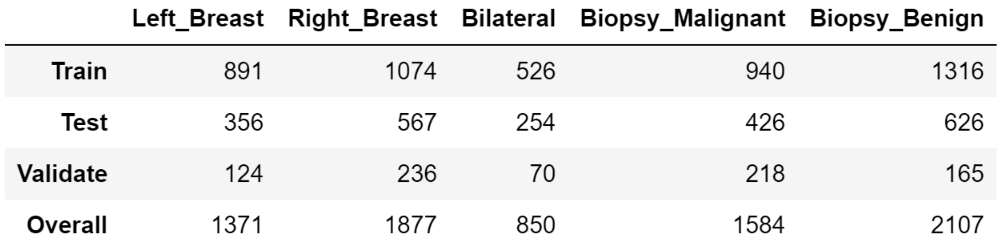


The study description column from the organized dataset describes whether the breast exam was recorded on the left or right breast or both (bilateral). The biopsy findings denote whether the tissue sample is malignant or benign. If the biopsy result indicates malignancy, the tissue sample contains cancer cells. In contrast, if the result is benign, the tissue sample does not contain cancer cells, and the abnormality is likely, not cancerous.

Table 1 represents the distribution of malignant and benign findings and the right, left and bilateral ultrasounds for the train, test, and validation sets.

Table 1

Distribution of biopsy findings amongst the train, test, and validation sets

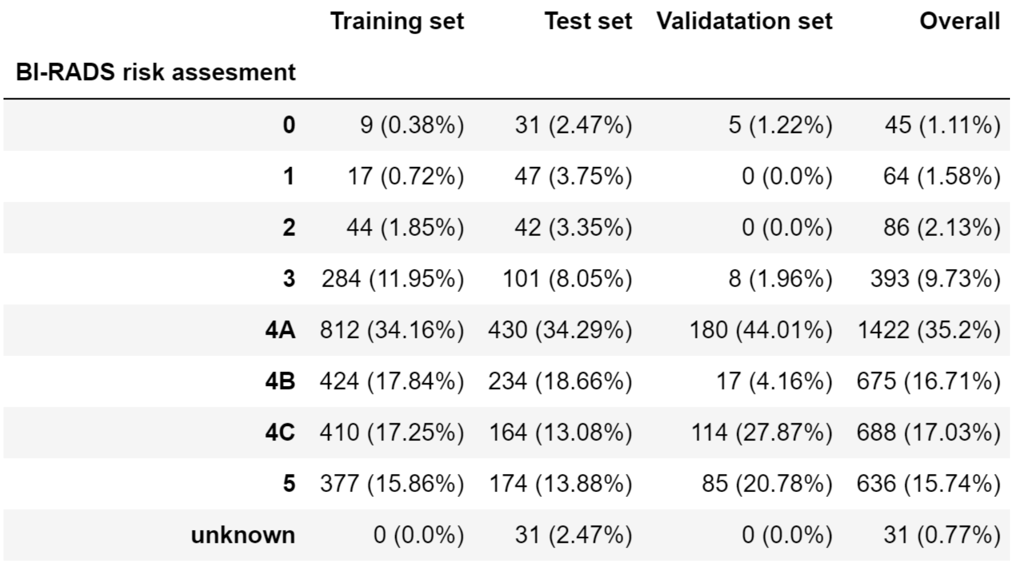


	Left_Breast	Right_Breast	Bilateral	Biopsy_Malignant	Biopsy_Benign
Train	891	1074	526	940	1316
Test	356	567	254	426	626
Validate	124	236	70	218	165
Overall	1371	1877	850	1584	2107

The radiologist allocated the BI-RADS risk assessment labels. It indicated suspicions of malignancy. Tables 2 and 3 summarize the distribution of mammographic breast densities and BI-RADS risk assessment labels in the dataset.

Table 2

Distribution of BI-RADS risk assessment labels for ultrasound images

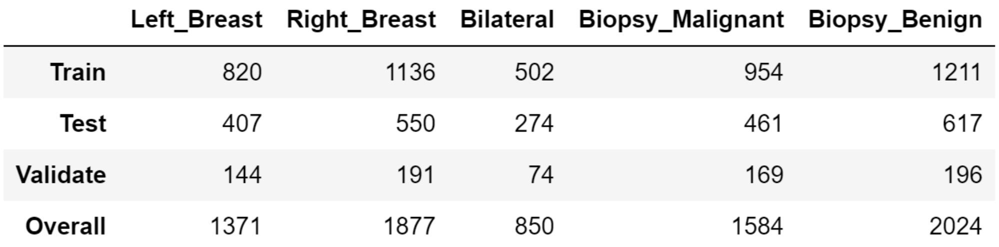


	Training set	Test set	Validatation set	Overall
BI-RADS risk assesment				
0	9 (0.38%)	31 (2.47%)	5 (1.22%)	45 (1.11%)
1	17 (0.72%)	47 (3.75%)	0 (0.0%)	64 (1.58%)
2	44 (1.85%)	42 (3.35%)	0 (0.0%)	86 (2.13%)
3	284 (11.95%)	101 (8.05%)	8 (1.96%)	393 (9.73%)
4A	812 (34.16%)	430 (34.29%)	180 (44.01%)	1422 (35.2%)
4B	424 (17.84%)	234 (18.66%)	17 (4.16%)	675 (16.71%)
4C	410 (17.25%)	164 (13.08%)	114 (27.87%)	688 (17.03%)
5	377 (15.86%)	174 (13.88%)	85 (20.78%)	636 (15.74%)
unknown	0 (0.0%)	31 (2.47%)	0 (0.0%)	31 (0.77%)

The BI-RADS risk assessment labels were obtained from the ultrasound reports of the patients, and the label 'Unknown' was assigned to the exams that contained incomplete or unclear information.

Table 3

Distribution of Mammographic breast density labels



	Left_Breast	Right_Breast	Bilateral	Biopsy_Malignant	Biopsy_Benign
Train	820	1136	502	954	1211
Test	407	550	274	461	617
Validate	144	191	74	169	196
Overall	1371	1877	850	1584	2024

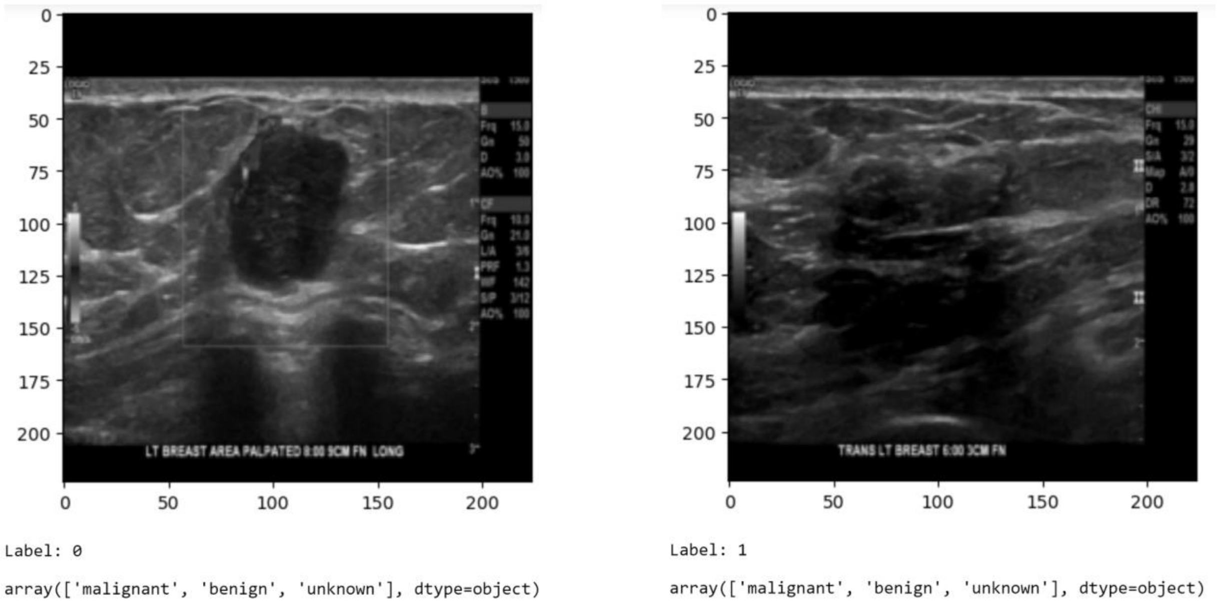
Mammographic density labels were extracted from the patient's mammogram report.

The data distribution statistics tables help a client understand the size of the dataset and the proportion of data used for training, testing, and validation.

Next, the PyTorch data loaders were created to provide an efficient way to load and iterate over a dataset during training or inference. The PyTorch DataLoader from this dataset returned batches of images and their corresponding class indices. The batches were used to train a classification model, where the model learned to predict the class of each image based on its features. The model aimed to classify each image into one of the possible classes, as shown in Figure 14.

Figure 14

Ultrasound images with predicted labels



```
Label: 0
array(['malignant', 'benign', 'unknown'], dtype=object)
```

```
Label: 1
array(['malignant', 'benign', 'unknown'], dtype=object)
```

Label 0- malignant | Label 1- benign

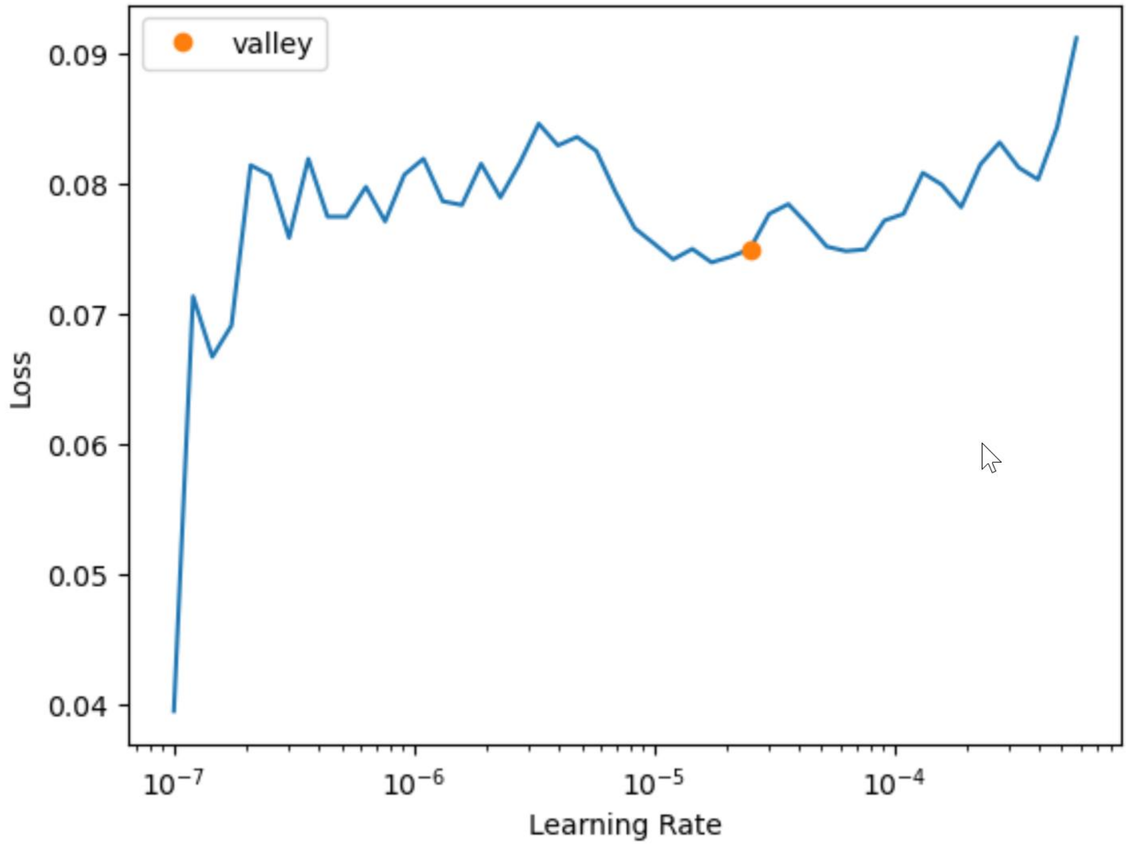
4.3 Accuracy assessment:

The training data was passed to the ResNet50 neural network architecture using PyTorch and FastAI libraries. The model was trained using the training data for 25 epochs, and the accuracy was recorded for each epoch. The study by He et al. (2016) discussed that residual learning was widely used for image classification tasks because of its efficiency and high accuracy. The ResNet architecture, which employed residual learning, succeeded in various image classification tasks. One of the main reasons it was good for image classification was that it used residual blocks, allowing the network to learn more complex features in a deeper architecture without suffering from the problem of vanishing gradients.

Setting the correct hyperparameters was important for ResNet as they determined the architecture and behavior of the model during the training. Farag et al. (2021) discussed that the hyperparameters in ResNet, such as the number of layers, the width of the layers, the learning rate, batch size, weight decay, dropout rate, and activation function, affect the model's capacity to learn and generalize from the training data. Therefore, incorrectly setting these hyperparameters can result in the model overfitting or underfitting the training data, which can ultimately cause poor performance when presented with new or unseen data. For example, the learning rate was a critical hyperparameter determining the step size taken during optimization to update the network's weights. Figure 15 shows the learning rate changes during the training process of a neural network to optimize its performance. The "valley" in the learning rate graph refers to the range of learning rates that can result in the best performance of the model.

Figure 15

Learning rate fluctuations



Haque et al. (2022) discussed the performance evaluation of a machine-learning model for breast cancer diagnosis using a confusion matrix. The authors reported the values of the true positive (TP), true negative (TN), false positive (FP), and false negative (FN) rates of the model. These were used to calculate various metrics such as sensitivity, also known as recall, specificity, and accuracy.

Next, to calculate sensitivity and specificity from a confusion matrix, I used the following formulas:

Sensitivity (recall or true positive rate):

Sensitivity = TP / (TP + FN)

Specificity (true negative rate):

Specificity = TN / (TN + FP)

Accuracy = (TP+TN)/(TP+TN+FP+FN)

Where:

TP: True positives (the number of cases where the model predicted the positive class correctly)

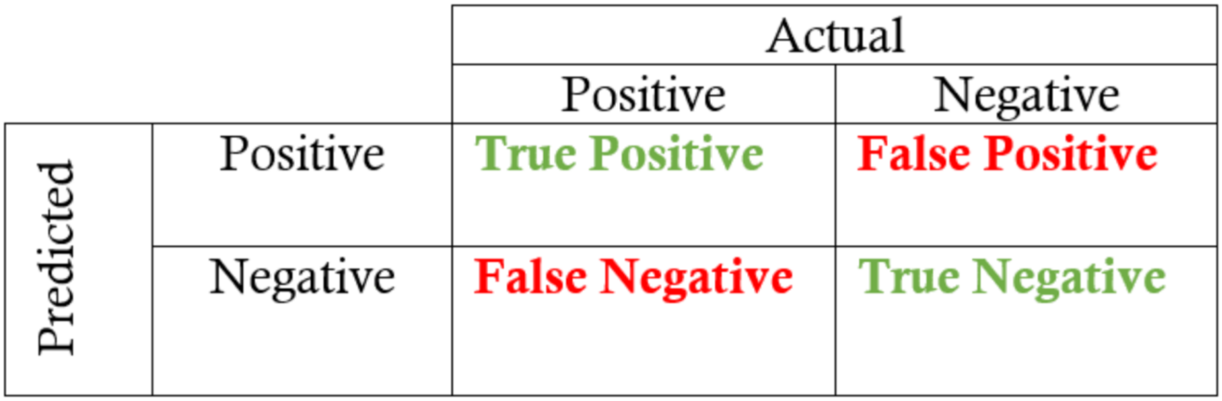
TN: True negatives (the number of cases where the model predicted the negative class correctly)

FP: False positives (the number of cases where the model predicted the positive class incorrectly)

FN: False negatives (the number of cases where the model predicted the negative class incorrectly)

Table 4

Confusion matrix details



		Actual	
		Positive	Negative
Predicted	Positive	**True Positive**	**False Positive**
	Negative	**False Negative**	**True Negative**

Table 4 represents the confusion matrix for this project, which summarizes the model's performance regarding these metrics.

Figure 16

Confusion matrix for a current project

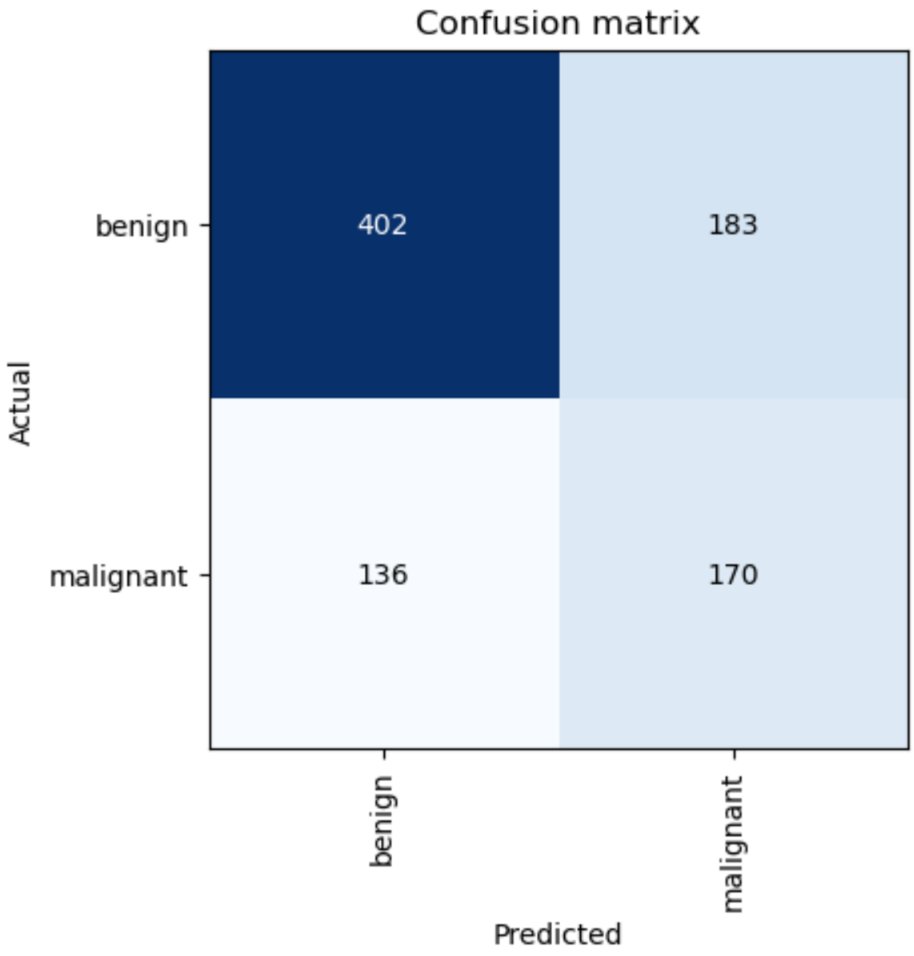


Figure 16 shows the confusion matrix generated from the model performance. After the calculations, the model achieved a sensitivity value of 0.74, correctly identifying 74% of the positive cases in the dataset as malignant tumors. The model also achieved a specificity value of 0.48, correctly identifying 48% of the negative cases in the dataset as breast cancer (benign) cases. Overall, the model's accuracy was 0.6459, accurately classifying 65% of the patients in the dataset.

<u>4.4 Business value:</u>

The project had significant business value as it aimed to potentially improve patient outcomes and reduce the cost of treatment through earlier detection and diagnosis of breast cancer. In the best-case scenario, the software's ability to accurately predict the presence of potentially cancerous lesions could have resulted in fewer unnecessary biopsies and medical procedures, leading to reduced healthcare costs. Furthermore, the software could have improved the efficiency and productivity of radiologists and healthcare professionals, leading to further cost savings.

In addition to the potential cost savings, the software could have significant benefits for medical professionals in helping with their expert decision-making. By providing accurate and efficient predictions of potentially cancerous lesions, the software could improve patients' overall quality of care and help healthcare professionals make more informed and timely decisions.

Chapter 5: Conclusion

5.1 Summary of findings:

The project used two text extraction tools for extracting annotated data from ultrasound images. Based on the findings, Pytesseract and EasyOCR were effective English text extraction tools. However, some images had different colored/light-colored text on them. EasyOCR performed better in such cases where the annotation text was light-colored. After extracting the text, the regular expression library was used to remove noisy data and special characters, and the implementation of logic helped improve the extracted data's accuracy. Finally, the accuracy of the extracted data was evaluated using exact matches between the raw text data and cleaned text data, resulting in an accuracy rate of 83%.

Based on the frequency analysis of text extraction, the annotation "LONG LT AXILLIA" and " LONG RT AXILLIA " were the most commonly used by radiologists. In addition, frequency analysis gives an idea of what keywords are used repetitively, which helps further annotation parsing.

In addition, the model used for classifying breast tumors achieved a sensitivity value of 0.74 and a specificity value of 0.48, indicating a reasonably fair accuracy in identifying benign and malignant tumors. The model's overall accuracy was 0.6459, meaning it accurately classified 65% of the patients in the dataset.

5.2 Interpretation of findings:

The article by Zeitchik (2021) discussed the potential benefits of using AI for breast cancer diagnosis, including reducing healthcare costs and improving patient outcomes. According to him, AI can improve breast cancer diagnosis and screening significantly, but further research and development are needed to realize this potential fully. The findings from this project highlight the importance of using a combination of tools and techniques to achieve accurate text extraction and medical diagnosis results.

EasyOCR performed better than Pytesseract in cases where the annotation text was light-colored. This is because EasyOCR uses a deep learning-based approach, making it more robust to variations in image quality, such as lighting, contrast, and background. On the other hand, Pytesseract is based on traditional optical character recognition (OCR) techniques, which may not perform well under such conditions. However, it's important to note that the overall accuracy rate when both tools were used combinedly was 83%, indicating that they performed reasonably well in extracting text from ultrasound images in English.

The finding that radiologists' most commonly used annotations were "LONG LT AXILLIA" and "LONG RT AXILLIA" means that radiologists frequently use these annotations to describe findings in breast ultrasound images. "LONG LT AXILLIA" refers to a long-axis view of the left axilla, and "LONG RT AXILLIA" refers to a long-axis view of the right axilla. In addition, these annotations may be useful for other radiologists and healthcare professionals who need to interpret the same images or follow up on a patient's condition.

Mendes et al. (2022) used ResNet18 architecture for their study. They noted that fine-tuning on a large dataset could help improve the network's performance on a smaller, more specific dataset, such as a breast cancer imaging dataset. This project used ResNet50 architecture. ResNet50 is a deeper neural network architecture than ResNet18, with more layers and parameters, which allows it to capture more complex features in the data. The classifier using ResNet50 achieved higher accuracy (65%) than the one using ResNet18 (61%). This indicates that ResNet50 performed better than ResNet18 in the given task. The PyTorch data loaders used in this project efficiently passed the data to the ResNet50 model for classification, demonstrating the effectiveness of the data processing pipeline.

5.3 Context of findings:

Studies by Satariano et al. (2023) & Grady D. (2020) have explored the use of AI in medical imaging for diagnosis and screening, focusing specifically on breast cancer diagnosis. This project's findings agree with previous research that suggests AI can significantly improve breast cancer diagnosis accuracy, reduce healthcare costs, and improve patient outcomes.

Regarding population characteristics, this project focused on a specific dataset of breast ultrasound images, which is somewhat similar to the data used by Al-Dhabyani et al. (2020) in their studies. The assessment instruments used in this project were text extraction tools and ResNet50 architecture for breast cancer classification, which were also used by Raza et al. (2023) in their studies. This project's research design and procedures are also consistent with previous research. The project's findings agree with existing literature that suggests the choice of neural network architecture can significantly affect the accuracy of breast cancer classification.

The project's findings extend previous research in AI-assisted breast cancer diagnosis. While previous studies have investigated the use of various machine learning techniques for breast cancer classification using mammography images, this project focuses on using ultrasound images and text extraction techniques to improve the accuracy of diagnosis. Additionally, the project uses a combination of Pytesseract and EasyOCR text extraction tools to improve the accuracy of extracted text data, which has not been widely explored in previous studies. The use of ResNet50 architecture for breast tumor classification extends previous research, as some studies have used other deep learning architectures. Overall, the project's findings provide insights into the potential benefits of using a combination of tools and techniques to achieve accurate diagnosis results, which can contribute to further research in AI-assisted breast cancer diagnosis.

5.4 Implications of findings:

The implications of the findings can be discussed in terms of theory, research, and practice.

Regarding theory, the findings suggest that a combination of tools and techniques can improve the accuracy of breast cancer diagnosis using ultrasound images. Pytesseract and EasyOCR text extraction tools and ResNet50 architecture for breast tumor classification are consistent with previous research in AI-assisted medical diagnosis. In addition, the findings support deep learning-based approaches in image analysis and text extraction.

Regarding research methodology, the findings demonstrate the importance of careful data processing and noise reduction techniques when working with medical imaging data. For example, using regular expressions to remove noisy data and special characters helped to improve the accuracy of the extracted text data by up to 95%. Additionally, using ResNet50 architecture with efficient PyTorch data loaders demonstrated the effectiveness of the data processing pipeline. These findings may help to inform future studies on AI-assisted medical diagnosis using similar techniques.

Regarding practice, the findings suggest that AI-assisted breast cancer diagnosis using ultrasound images has the potential to significantly improve accuracy, reduce healthcare costs, and improve patient outcomes. Radiologists and other healthcare professionals may be interested in using the findings to inform their practice and improve their diagnostic accuracy. Combining text extraction tools and ResNet50 architecture may also be useful for other medical imaging applications beyond a breast cancer diagnosis.

Overall, the findings suggest that AI-assisted medical diagnosis using a combination of image analysis and text extraction techniques can lead to significant improvements in accuracy and efficiency. Furthermore, these findings inform future research and practice in medical imaging and AI-assisted diagnosis.

5.5 Limitations:

While this study has provided valuable insights into the effectiveness of text extraction tools for medical image analysis and the use of deep learning models for breast cancer classification, several potential limitations exist. These limitations could impact the findings' generalizability and reliability and should be considered when interpreting the results. The following are some of the key limitations of the study.

1. Dataset Size: The dataset used in the project was not large enough to represent the overall population of breast cancer patients. The limited diversity of the dataset may have affected the model’s accuracy.
2. Annotations by radiology technician: The annotations used in the dataset are assigned by only one radiology technician. Different radiology technicians may interpret imaging data differently, leading to variations in the assigned annotations. As a single person assigns annotations, the dataset may not reflect the diversity of interpretations in a larger population of radiology technicians. This could limit the usefulness and accuracy of the dataset for broader applications.

3. Annotation placement in ultrasound images: The annotation placement in the ultrasound images needs to be improved. Many of the images had annotations placed in the center bottom of the ultrasound image. In contrast, some of the images had annotations placed in the middle of the image, which affected the accuracy of the text extraction tools.
4. Model generalization: The model's performance may not generalize well to other datasets or real-world scenarios. The model's accuracy may vary depending on the type of breast cancer and the specific patient population. To detect the different types of breast cancer by the model, there’s a need for a more diverse dataset.
5. Tool dependencies: The accuracy of the text extraction tools used in the project may have depended on external libraries, such as the regular expression library, EasyOCR, and PyTesseract. Any changes to these dependencies could affect the accuracy of the extracted text.

The other limitations include internal and external validity, measurements, and statistical analysis.

- Internal Validity: As a retrospective study, this project is subject to certain limitations in terms of internal validity. Retrospective studies rely on data that have already been collected. The classifier takes longer to learn due to the limited dataset size and diversity, which could affect the generalizability of the study's findings to the larger population of breast cancer patients.
- External Validity: The lack of a diverse population of machines in the dataset may limit the study’s external validity. If the dataset used in the study only includes images from a specific type or brand of machines, the model's accuracy may not generalize well to other types of machines used in different settings. This could limit the model’s applicability in

real-world scenarios where diverse populations of machines are used. Next, the study used a specific dataset of breast ultrasound images, and the performance of the text extraction and classification tools may vary on different datasets. Therefore, the findings may not directly apply to other datasets or populations.

- Measurement: The text extraction tool's accuracy was measured based on the noise level between the raw and cleaned text. The ground truth labels data is not available for this dataset. In real-time scenarios, relying solely on this type of evaluation may lead to reliability issues. Ideally, the assessment should be based on a comparison with ground truth labels, but creating such labels can be time-consuming and expensive.
- Statistical analysis: A limitation of this project is the small sample size, which may limit the statistical power of the analysis. Additionally, the choice of statistical tests used to evaluate the performance of the text extraction and classification tools may influence the results.

5.6 Future directions:

One potential future direction is to extend the study to different populations to see if the results are consistent across different groups. For example, the study could be replicated with different age groups, genders, races, or cultural backgrounds to see if the results are consistent. Adding imaging modalities, such as suspicious mammograms and different populations of machines, as well as including genetic variations, such as BRCA1 and BRCA2 categories, could be useful too.

Incorporating imaging modalities can provide additional information about potential breast cancer risk and can be useful in assessing breast cancer risk in different populations. For example, women with suspicious mammograms may be at increased risk for breast cancer. This information can help assess the effectiveness of breast cancer prevention and early detection strategies in different populations. Similarly, certain genetic variations are associated with an increased risk of breast cancer. This information can be useful in identifying high-risk individuals and developing targeted prevention and early detection strategies.

To enhance the accuracy of their predictive models, researchers may explore the use of data augmentation methods. These methods involve creating additional data points by applying various transformations such as flipping, scaling, and rotating to the existing dataset. This can help increase the dataset's size, reduce overfitting, and improve the accuracy of the model's predictions. For example, in image classification tasks, researchers can use techniques such as random cropping, flipping, and rotating to create variations of the original images. This can help the model to better recognize objects from different angles and orientations, leading to more accurate predictions.

5.7 Final thoughts:

In conclusion, this project demonstrates the potential benefits of using the collaboration of tools and techniques to achieve accurate diagnosis results in AI-assisted breast cancer diagnosis. Pytesseract and EasyOCR were effective English text extraction tools, with EasyOCR performing better in cases where the annotation text was light-colored. In addition, the ResNet50 architecture performed better than ResNet18 in the breast tumor classification task, achieving an accuracy rate of 65%. The project's findings extend previous research in AI-assisted breast cancer diagnosis, providing insights into the potential benefits of using a combination of tools and techniques to improve diagnosis accuracy. The implications of the findings can contribute to further research and practical applications, such as lowering healthcare costs by reducing biopsies and improving patient outcomes.

References

Adjetey, C., & Adu-Manu, K. S. (2021). Content-based Image Retrieval using Tesseract OCR Engine and Levenshtein Algorithm. *International Journal of Advanced Computer Science and Applications (IJACSA)*, *12*(7). http://dx.doi.org/10.14569/IJACSA.2021.0120776

Alanazi, S. A., Kamruzzaman, M. M., Sarker, M. N. I., Alruwaili, M., Alhwaiti, Y., Alshammari, N., & Siddiqi, M. H. (2021). Boosting Breast Cancer Detection Using Convolutional Neural Network. *Journal of Healthcare Engineering*, 5528622, 11. https://doi.org/10.1155/2021/5528622

Al-Dhabyani, W., Gomaa, M., Khaled, H., & Fahmy, A. (2020). Dataset of breast ultrasound images. *Data in Brief*, 28, 104863. https://doi.org/10.1016/j.dib.2019.104863

Anastasiadi, Z., Lianos, G. D., Ignatiadou, E., & Haralampos, V. (2017). Breast cancer in young women: An overview. *Updates in Surgery*, *69*(3), 313-317. https://doi.org/10.1007/s13304-017-0424-1

Bard, R. L., DABR, M.D., FASL, & Cutter, Dense Breast Imaging Detection and Image-guided Oncologic Treatment. (2021, November 9). *Imaging technology news*. https://www.itnonline.com/article/dense-breast-imaging-detection-and-image-guided-oncologic-treatment

Batista, G. E., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM Sigkdd Explorations Newsletter*, 6(1), 20-29. https://doi.org/10.1145/1007730.1007735

Chandrakesan, G. (2021, May 27) Day 33: Predict Image Using ResNet50 Pretrained Model. *LinkedIn*. https://www.linkedin.com/pulse/day-33-predict-image-using-resnet50-pretrained-model-chandrakesan/

Colangelo, M. (2023, April 18). Using Deep Learning To Diagnose Breast Cancer With High Accuracy. *LinkedIn*. https://www.linkedin.com/pulse/using-deep-learning-diagnose-breast-cancer-high-margaretta-colangelo/?trk=article-ssr-frontend-pulse_more-articles_related-content-card

Farag, H. H., Kandil, A. H., Mohamed, H. A., Mahmoud, A. H., & Al-Antari, M. A. (2021). Hyperparameters optimization for ResNet and Xception in the purpose of diagnosing COVID-19. *Computer Methods and Programs in Biomedicine*, 200, 3555-3571. DOI: 10.3233/JIFS-210925

Genzel, D., Popat, A., & Narayanan, D. (2015, May 6). Paper to digital in 200+ languages. *Googleblog*. https://ai.googleblog.com/2015/05/paper-to-digital-in-200-languages.html

Geras, K. (2019, October 17). Combination of Artificial Intelligence & Radiologists More Accurately Identified Breast Cancer. *NYU Langone Health*. https://nyulangone.org/news/combination-artificial-intelligence-radiologists-more-accurately-identified-breast-cancer

Grady, D. (2020, Januvary 1). A.I. Is learning to read Mammograms. *The New York Times*. https://www.nytimes.com/2020/01/01/health/breast-cancer-mammogram-artificial-intelligence.html

Haque, M. N., Tazin, T., Khan, M. M., Faisal, S., Ibraheem, S. M., Algethami, H., & Almalki, F. A. (2022). Predicting Characteristics Associated with Breast Cancer Survival Using Multiple

Machine Learning Approaches. *Computational and mathematical methods in medicine.* 1249692. https://doi.org/10.1155/2022/1249692

He, K., Zhang, X., Shaoqing, R., & Sun, J. (2016). Deep Residual Learning for Image Recognition. *IEEE Conference on Computer Vision and Pattern Recognition*, 770–778. https://openaccess.thecvf.com/content_cvpr_2016/papers/He_Deep_Residual_Learning_CVPR_2016_paper.pdf

Healy, M. (2022, January 4). Breast Imaging, Reporting & Data System (BIRADS). *OncoLink.* https://www.oncolink.org/cancers/breast/screening-diagnosis/breast-imaging-reporting-data-system-birads

Jiang, L., & Zhang, Z. (2021). Research on Image Classification Algorithm Based on Pytorch. *Journal of Physics: Conference Series. 2010*(1), 12009. https://doi.org/10.1088/1742-6596/2010/1/012009

Kolata, G. (2023). F.D.A. Will Require Dense Breast Disclosure at Mammogram Clinics. *The new york times.* Retrieved from https://www.nytimes.com/2023/03/09/health/dense-breast-fda-mammogram.html

Kryzhanivska, O. (2017, August 25). The Importance of Baseline Data. *LinkedIn*. https://www.linkedin.com/pulse/importance-baseline-data-olena-kryzhanivska/

Lang, K., Dustler, M., Dahlblom, V., Akesson, A., Andersson, I., & Zackrisson, S. (2021). Identifying normal mammograms in a large screening population using artificial intelligence. *European Radiology, 31*(3), 1687-1692. https://doi.org/10.1007/s00330-020-07165-1

Maclary, D. (2019, October 16). R vs Python: What do Data Scientists prefer? *LinkedIn*. https://www.linkedin.com/pulse/r-vs-python-what-do-data-scientists-prefer-donnie-maclary/

Mercer, C. E., Hogg, P., Kelly, J., Borgen, R., Millington, S. R., Hilton, B., Enion, D., & Whelehan, P. (2014). A mammography image set for research purposes using BI-RADS density classification. *Radiologic Technology*, *85*(6), 609–613. https://pubmed.ncbi.nlm.nih.gov/25002640/

Mori, S., Suen, C. Y., & Yamamoto, K. (1992). Historical review of OCR research and development. *Proceedings of the IEEE, 80*(7), 1029-1058. https://doi.org/10.1109/5.156468

Novac, O. C., Chirodea, M. C., Novac, C. M., Bizon, N., Oproescu, M., Stan, O. P., & Gordan, C. E. (2022). Analysis of the application efficiency of TensorFlow and PyTorch in convolutional neural network. *Sensors*, *22*(22), 8872. https://doi.org/10.3390/s22228872

Pan, H. (2016). The role of breast ultrasound in early cancer detection. *Journal of Medical Ultrasound, 24*, 138-141. https://doi.org/10.1016/j.jmu.2016.10.001

Raza, A., Ullah, N., Khan, J. A., Assam, M., Guzzo, A., & Aljuaid, H. (2023). Deep breast cancer net: A novel deep learning model for breast cancer detection using ultrasound images. *Applied Sciences. 13*(4), 2082. http://dx.doi.org/10.3390/app13042082

Reig, B., Heacock, L., Geras, K. J., & Moy, L. (2020). Machine learning in breast MRI. *Journal of magnetic resonance imaging*. *52*(4), 998–1018. https://doi.org/10.1002/jmri.26852

Satariano, A., & Metz, C. (March, 5, 2023). Using A.I. to detect breast cancer that doctors miss. *The New York Times*.

https://www.nytimes.com/2023/03/05/technology/artificial-intelligence-breast-cancer-detection.html

Siegel, R. L., Miller, K. D., Fuchs, H. E., & Jemal, A. (2021). Cancer statistics, 2021. *CA: A Cancer Journal for Clinicians*, *71*(1), 7-33. https://doi.org/10.3322/caac.21551

Saul, S. (July, 19, 2010). Prone to error: Earliest steps to find cancer. *The New York Times*. https://www.nytimes.com/2010/07/20/health/20cancer.html

Splane, B. (2022, October 17). What is a benign vs. malignant Tumor? *Verywell health*. https://www.verywellhealth.com/what-does-malignant-and-benign-mean-514240

Walsh, R., & Tardy, M. (2023). A comparison of techniques for class imbalance in deep learning classification of breast cancer. *Diagnostics, 13*(1), 67. https://doi.org/10.3390/diagnostics13010067

Wu, G. G., Zhou, L. Q., Xu, J. W., Wang, J. Y., Wei, Q., Deng, Y. B., Cui, X. W., & Dietrich, C. F. (2019). Artificial intelligence in breast ultrasound. *World journal of radiology, 11*(2), 19-26. https://doi.org/10.4329/wjr.v11.i2.19

Zeitchik, S. (2021). Is artificial intelligence about to transform the mammogram? *The Washington Post.* https://www.washingtonpost.com/technology/2021/12/21/mammogram-artificial-intelligence-cancer-prediction/

Zheng, Y., Wang, X., Fan, L., & Shao, Z. (2021). Breast cancer-specific mortality in small-sized tumor with stage IV breast cancer: A population-based study. *The Oncologist*, *26*(2), e241-e250. https://doi.org/10.1002/onco.13567

Zhu, Q., & Poplack, S. (2020). A review of optical breast imaging: Multi-modality systems for breast cancer diagnosis. *European journal of radiology, 129*, 109067. https://doi.org/10.1016/j.ejrad.2020.109067

Appendix

Description of Python code files:

1. Capstone2023.ipynb – the file contains all code that written in python programming language used to generate all results as discussed previously throughout the paper.
2. textExtractionOnly.ipynb- the file contains the code that extracts the text for the images where the text area was blank. There were couple of images for which the text extraction via PyTesseract was not possible hence advanced tool EasyOCR was used through google colab.

Excel files:

1. organized_db.csv- Final organized dataset
2. FilteredText.csv- File generated through google colab notebook

Link for the project code:

https://github.com/pranalishendekar1/Pranali-shendekar