UNIVERSITY OF SCIENCE AND TECHNOLOGY OF HANOI

A blue text on a black background

Description automatically generated

Breast Cancer EDA

|  |  |
| --- | --- |
| Nguyễn Đức Thành | BI12-420 |
| Hà Hải Long | BI12-256 |
| Tăng Minh Trí | BI12-445 |
| Trần Ngọc Hải | BI12-150 |
| Phạm Trung Hiếu | BI12-159 |

Contents

[1. Introduction 3](#_Toc164116196)

[2. Overview 3](#_Toc164116197)

[3. Dataset Overview 4](#_Toc164116198)

[*3.1.* *Data Preparation* 4](#_Toc164116199)

[*3.2.* *Data Analysis* 5](#_Toc164116200)

[4. Models 7](#_Toc164116201)

[*4.1.* *Pytorch Model* 7](#_Toc164116202)

[*4.2.* *Evaluation and Prediction* 7](#_Toc164116203)

[5. Conclusion 8](#_Toc164116204)

[6. Future work 8](#_Toc164116205)

# Introduction

Breast cancer remains a significant health concern worldwide, impacting millions of women each year. In the realm of medical research, Exploratory Data Analysis (EDA) serves as a foundational approach to uncovering insights and patterns within complex datasets. This report embarks on a journey of EDA specifically tailored to breast cancer, aiming to shed light on various aspects of this disease through a data-driven lens.

EDA is a crucial preliminary step in understanding the nuances of breast cancer data. By visually exploring and summarizing key characteristics of the dataset, we can gain valuable insights into the distribution, relationships, and potential trends within the data. Through this process, we aim to identify important features, patterns, and potential predictive factors associated with breast cancer incidence, diagnosis, and prognosis. The primary objectives of this EDA report are twofold: first, to provide a comprehensive overview of the breast cancer dataset under examination, detailing its structure, content, and quality; and second, to conduct a systematic exploration of the data, uncovering meaningful insights that can inform future research, clinical practice, and public health initiatives. By leveraging the power of data visualization, statistical analysis, and machine learning techniques, we seek to unravel the complexities of breast cancer and contribute to the collective knowledge base surrounding this disease. Through rigorous analysis and interpretation of the data, we endeavor to not only deepen our understanding of breast cancer but also pave the way for more effective prevention, early detection, and treatment strategies.

This report serves as a testament to the transformative potential of EDA in the fight against breast cancer. By harnessing the wealth of information contained within the data, we strive to empower researchers, healthcare professionals, and policymakers with actionable insights that can ultimately improve outcomes and quality of life for those affected by this disease.

# Overview

The purpose of this presentation is to perform exploratory data analysis (EDA) on breast cancer datasets obtained from the Radiological Society of North America (RSNA). This analysis will provide participants with insights into the demographic characteristics of breast cancer patients, the distribution of various breast cancer subtypes, imaging modalities used in diagnosis, and patterns in disease progression. By examining key trends and associations in the data, participants will gain a better understanding of the breast cancer diagnosis and treatment landscape as represented by the RSNA datasets. The insights gained from this EDA will help healthcare professionals, researchers, and policymakers make more informed decisions about how to improve outcomes and reduce the global burden of breast cancer.

# Dataset Overview

* 1. *Data Preparation*

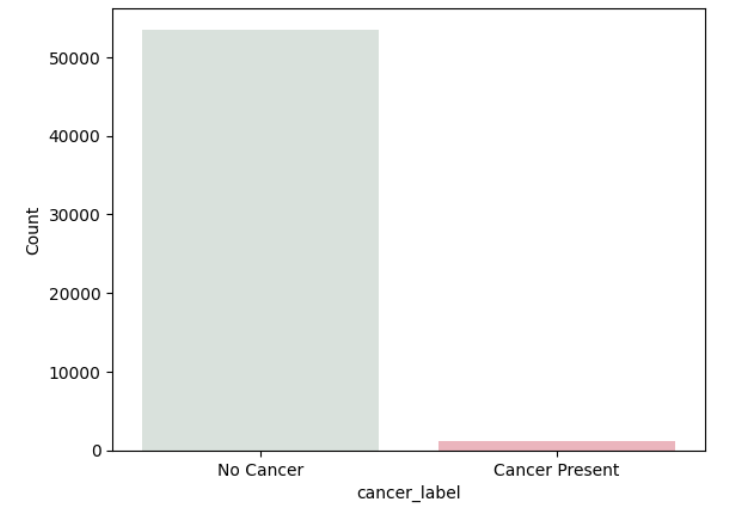
Our group dataset is taken from many different sources. Each dataset shows its factors and characteristics, but they all represent factors that affect breast cancer. Some important attributes can be mentioned:

|  |  |
| --- | --- |
| **Attributes** | **Meaning** |
| **site\_id** | The identification of the site where the medical imaging was performed |
| **patient\_id** | The unique identifier assigned to each patient. |
| **image\_id** | The identification number assigned to each individual image. |
| **Age** | The age of the patient at the time of imaging. |
| **biopsy information** | Indicates whether a biopsy was performed or not. |
| **BIRADS** | Breast Imaging Reporting and Data System score, indicating the suspicion level for breast cancer presence. |
| **implant density** | Indicates if there is an implant present in the breast during imaging. |
| **machine\_id** | Identification number of the machine used for imaging. |
| **difficult\_negative\_case** | Indicates if it’s a challenging case to confirm as negative for cancer or other abnormalities. |
| **Laterality\_L** | Indicates which side (left or right) is being imaged. |
| **Laterality\_R** | Indicates which side (left or right) is being imaged. |
| **View\_AT, CC, LM, LMO, ML, MLO** | Different angles and positions from which images are captured during mammography. |
| **Density** | Attributes (**density\_A**, **density\_B**, **density\_C**, **density\_D**): Indicate breast density categories from A (least dense) to D (most dense). |

We focus on breast cancer detection using the RSNA dataset and PyTorch. The analysis covers various aspects of the project, including data preprocessing, model development, evaluation metrics, and visualization techniques. It includes data loading from CSV files, exploratory data analysis to understand the dataset's characteristics, and model development using deep learning with PyTorch. The notebook also incorporates visualizations using libraries like Matplotlib, Plotly Express, and Lets-Plot to enhance data exploration and model performance interpretation. Overall, we want to provide a detailed and structured approach to breast cancer detection, making it a valuable resource for understanding the application of deep learning to medical image analysis tasks.

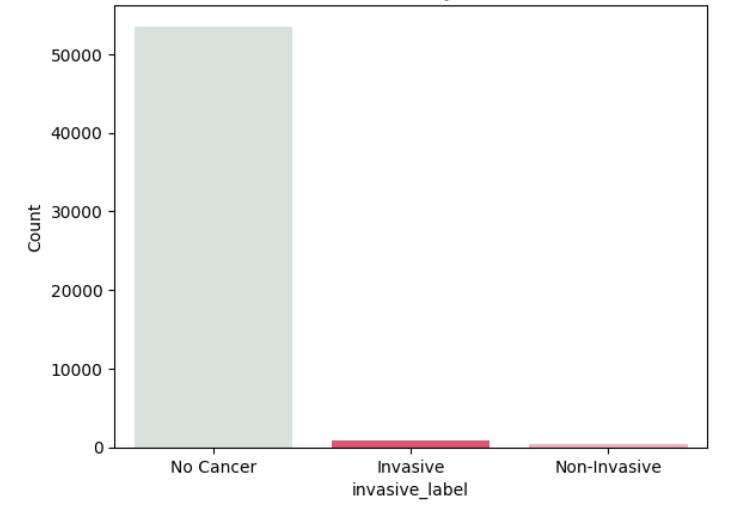
* 1. *Data Analysis*

Data analysis is a critical process in which raw data is analyzed and interpreted to uncover meaningful insights, patterns, and trends. Through various statistical and computational techniques, data analysis sifts through vast amounts of information to extract valuable knowledge that can inform decision-making. As for our team, analyses of factors related to breast cancer were performed, along with visuals to help highlight those characteristics.



**Figure 1. Cancer Distribution.**

The chart "Cancer Distribution" reveals a notable difference in the counts of individuals with cancer compared to those without cancer. By visually illustrating the distribution of cancer cases, this visualization aids in identifying high-risk populations and guiding targeted interventions for cancer prevention and treatment. This chart underscores the importance of visual data representation in elucidating complex healthcare phenomena such as cancer distribution and offering actionable insights for public health initiatives and clinical practice.

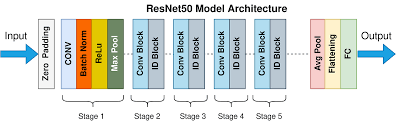


**Figure 2. Cancer Distribution by Invasiveness.**

The chart "Cancer Distribution by Invasiveness" reveals distinct patterns in the distribution of cancer cases across the three categories. The significantly taller bar representing "No Cancer" indicates a higher prevalence of non-cancerous cases, while the bars for "Invasive" and "Non-Invasive" cancers display varying counts, with invasive cases being more prevalent than non-invasive ones. This visualization serves as a valuable tool for assessing the prevalence of different cancer types within a population and guiding decision-making processes in healthcare management and research. Further investigations into the underlying factors contributing to the observed distribution patterns are warranted to enhance our understanding of cancer epidemiology and treatment outcomes.

# Models

* 1. *Pytorch Model*

The PyTorch model leverages both the imaging data and the meta data to make predictions. More specifically, it uses the age, laterality, view, and implant features from the meta data. Previously, age was min-max scaled, and the remaining meta features were dummy encoded. A pre-trained ResNet50 model was used to output predicted probabilities, which were then combined with the meta-data features and used as inputs to a fully connected network that returned the final predictions.

We use the ResNet50 model as the backbone to extract features from images. We change the first convolution layer of ResNet50 to receive single-channel images (x-ray images are usually black and white) instead of three-channel RGB images. We changed the last fully connected layer to have 500 output features instead of ImageNet's number of layers. The linear (fully connected) layer combines features from ResNet and metadata to make the final classification decision. This layer receives 508 features (500 from ResNet and 8 from metadata) and outputs a single value indicating the probability of cancer.

* 1. *Evaluation and Prediction*

By leveraging mammography images and metadata information, we have developed a sophisticated model to predict the likelihood of breast cancer. Our model considers various factors, such as patient age, affected breast side, view angle, and the presence of implants, to enhance the accuracy of the diagnosis.

The methodology employed in this study involves a multi-step process, starting with the processing of mammography images and metadata information. We extract relevant features from the images and metadata, including patient demographics and imaging characteristics. Subsequently, we define a custom dataset class, `MammographyDataset`, to organize and prepare the data for model training. The dataset is then split into training and validation sets, and data loaders are initialized for efficient batch processing.

Numerical Results:

* Loss on validation set: 0.69
* Accuracy on validation set: 58.10%
* Sensitivity on validation set: 0.33
* Specificity on validation set: 0.84

Our results demonstrate promising accuracy and sensitivity in predicting the likelihood of breast cancer based on the combined information from metadata and image features. The model shows potential for enhancing the efficiency and accuracy of breast cancer diagnosis, contributing to improved patient care and outcomes.

# Conclusion

To conclude, this comprehensive EDA on breast cancer datasets from RSNA has shed light on the demographics of patients, imaging methods and modalities, as well as patterns connected with the disease. We have identified significant results through rigorous data cleaning, visualization, and analysis to guide further research, clinical practice, and public health initiatives in the fight against breast cancer.

For example, patient demographics, biopsy information, BIRADS scores, implant density, machine id’s and imaging angles were mentioned in the dataset overview. These are essential components that help us understand what causes breast cancer to be detected or diagnosed. In this project, we focused on how deep learning can be used to develop models for detecting breast cancer using the RSNA dataset and PyTorch.

This study underscores the transformative potential of data-driven approaches in advancing our understanding of breast cancer and guiding future interventions. By leveraging the power of data analytics and machine learning, we aim to contribute to the collective efforts aimed at improving prevention, early detection, and treatment strategies for breast cancer. The insights gained from this analysis serve as a foundation for ongoing research endeavors and collaborative initiatives aimed at reducing the global burden of breast cancer and improving the quality of life for affected individuals.

# Future work

In order to be able to locate the causative agent out of cancer, we will attempt to modify the way we train models in the future. Specifically, we will switch from training them on picture sets with the png extension to dicom image sets. In order to get good results, we will also work to enhance model evaluations.