Machine Learning-based Prediction of Molecular Subtypes of Breast Cancer using DCE MRI

by

© Javad Aghadavood Marnani

A Thesis submitted to the School of Graduate Studies in partial fulfillment of the requirements for the degree of Master of Science.

Supervisor: Dr. Hamid Usefi

Co-supervisor: Dr. J. Conceptión Loredo-Osti

Department of Mathematics and Statistics
Memorial University of Newfoundland

May 2023

St. John's, Newfoundland and Labrador, Canada

Abstract

Breast cancer is a prevalent disease that can be classified into four molecular subtypes based on genetic and molecular markers. This study aimed to develop a machine learning-based approach to classify molecular subtypes of breast cancer using radiomics features extracted from dynamic contrast-enhanced magnetic resonance imaging (DCE MRI). The comprehensive dataset used in this study included 4428 radiomics features per patient, as well as clinical features, making it a valuable resource for future research. Our methodology involved several stages, including image preprocessing, feature extraction, initial and final feature selection, and data cleaning techniques, such as data imputation and Local Outlier Factor (LOF), to ensure the quality of the dataset. We conducted hyperparameter tuning and robustness analysis to optimize the performance of the machine learning algorithms. The results were evaluated in three scenarios: 4-label classification, binary, and 3-label classifications. Our approach achieved up to 85% F1 score in binary classifications and improved the overall accuracy of classifying the four molecular subtypes of breast cancer by 12%, which represents a significant improvement over the original study. These findings suggest that machine learning algorithms can be a powerful tool for improving the diagnosis and treatment of breast cancer, paving the way for personalized medicine approaches. Furthermore, the proposed approach can be applied to other datasets and may be useful in other areas of medical research that rely on radiomics features extracted from medical images.

Acknowledgments

I would like to express my deepest gratitude to my esteemed supervisors, Dr. Hamid Usefi and Dr. J Concepcion Loredo-Osti, for their invaluable guidance, support, and mentorship during my master's degree. Their expertise and dedication have been instrumental in shaping my research and academic experience.

I extend my profound appreciation to Dr. Alexander Bihlo, Dr. Lourdes Pena-Castillo, Dr. Candemir Cigsar, and Dr. Armin Hatefi for their invaluable support and teaching.

Furthermore, I would like to thank my family and friends for their unwavering encouragement and support throughout my academic journey. Their continuous support and motivation have been a source of strength and inspiration for me throughout my academic journey.

Contents

List of Figures

List of Tables

List of Abbreviations

AI Artificial Intelligence

AM-FM Amplitude Modulation – Frequency Modulation

ANOVA Analysis of Variance

AUC Area Under the Curve

CDF Cumulative Distribution Function

CNN Convolutional Neural Network

CSV Comma-Separated Values

DBT Digital Breast Tomosynthesis

DCE Dynamic Contrast-Enhanced

DICOM Digital Imaging and Communications in Medicine

DL Deep Learning

DWT Discrete Wavelet Transform

ER Estrogen Receptor

FBM Fractional Brownian Motion

FPS Fourier Power Spectrum

FOS First Order Statistics

GBM Gradient Boosting Machine

GLCM Gray-Level Co-Occurrence Matrix

GLDS Gray Level Difference Statistic

GLRM Generalized Low Rank Models

GLSZM Gray Level Size Zone Matrix

GPU Graphics Processing Unit

GT Gabor Transform

HER2 Human Epidermal Growth Factor Receptor 2

HOS Higher Order Spectra

IDC Invasive Ductal Carcinoma

IV(line) Intravenous line

KNN K-Nearest Neighbors

LBP Local Binary Pattern

LDA Linear Discriminant Analysis

LOF Local Outlier Factor

MF Multiresolution Fractal

ML Machine Learning

MRI Magnetic Resonance Imaging

NA Not Available

NBIA National Biomedical Imaging Archive

NNs Neural Networks

OvO One versus One

OvR One versus the Rest

PACS Picture Archiving and Communication System

PCA Principal Component Analysis

PDF Probability Density Function

PNG Portable Network Graphics

PR Progesterone Receptor

RAM Random Access Memory

RF Random Forest

RNN Recurrent Neural Network

ROI Region of Interest

SFM Statistical Feature Matrix

SVM Support Vector Machines

STD Standard Deviation

TCIA The Cancer Imaging Archive

TVT One Versus One

XGBoosting Extreme Gradient Boosting

NGTDM Neighborhood Gray Tone Difference Matrix

Chapter 1

Introduction

Breast cancer is a complex disease that affects public health worldwide. It is the most common cancer among women, with an estimated 2.3 million new cases and 685,000 deaths in 2020 [1]. Breast cancer is a heterogeneous disease with distinct molecular subtypes that have different clinical presentations, responses to therapy, and prognoses. These subtypes are typically defined by the expression of biomarkers such as estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2). The four main molecular subtypes of breast cancer are Luminal A, Luminal B, HER2+, and Triple Negative (TN) [2, 3]. Traditionally, breast cancer subtype classification has relied on invasive and time-consuming histological and immunohistochemical analyses of biopsy samples. However, these methods may not always provide accurate information on tumor heterogeneity and are prone to sampling errors [4]. In recent years, non-invasive imaging techniques such as magnetic resonance imaging (MRI) have been used to predict the molecular subtypes of breast cancer [5–10]. MRI is a powerful imaging modality that provides high-resolution images of breast tissue which can be used for the detection and characterization of breast tumors. Furthermore, machine learning (ML) is a powerful set of techniques that can be used to analyze complex medical imaging data for the diagnosis and prognosis of various diseases, including breast cancer. Deep learning-based ML algorithms have shown significant improvements in the accuracy of breast cancer diagnosis and classification [11, 12]. In this thesis, an effective machine learning approach is developed to predict the molecular subtypes of breast cancer using dynamic contrast-enhanced (DCE) MRI data. DCE MRI is a non-invasive imaging technique used to detect and characterize breast cancer [13]. DCE MRI uses a gadolinium-based contrast agent injected into the patient's bloodstream through an IV line. Before the contrast agent is injected, a fat-saturated gradient echo T1-weighted pre-contrast sequence is acquired to provide baseline information about the breast tissue. After the pre-contrast sequence is acquired, a series of post-contrast sequences are acquired to capture the uptake and washout of the contrast agent in the breast tissue. This study focuses on pre-contrast and the first three post-contrast sequences of DCE MRI. The study uses locations of tumors annotated by experienced radiologists to filter out slices not containing tumors and crop the slices to focus on tumors further. Three different methods of cropping are used: the original crop based on the locations in the annotation file, 32×32 crop, and 64×64 crop from the middle of rectangles used for the first method of cropping. Segmentation is also used to narrow down slices to some tumors containing slices. Image segmentation is the process of dividing an image into multiple regions for further analysis. In this study, tumor regions were segmented from the cropped slices via thresholding. The Pyfeats Python library [14] is used to generate various features from the cropped images and the corresponding tumor segmentation as essential inputs. Pyfeats can extract a wide range of features from molecular dynamics simulations, including first order statistics, radial and bond angle distribution functions, coordination number distributions, grid-based features, geometric features, collective variables such as principal component analysis (PCA) and diffusion map analysis, free energy landscapes, reaction rates estimation, clustering analyses, gray level co-occurence matrix (GLCM) and gray level run length matrix (GLRM) features, and structural autocorrelation functions. These features provide insights into the structure, dynamics, and thermodynamics of molecular systems. Using four different sequences of DCE MRI and three different methods of cropping, a dataset consisting of 4428 radiomics features is obtained. Clinical features such as MRI technical information, demographics, tumor characteristics, and recurrence information are also included in this dataset. In order to train different machine learning (ML) models, we used the comprehensive dataset that we created. However, the dataset had missing values that needed to be handled using proper data imputation methods such as mean and KNN imputation [15–17].

Developing a machine learning-based approach for predicting the molecular subtypes of breast cancer involves a significant challenge of selecting appropriate features. To address this challenge, initial feature selection methods were performed to eliminate redundant features, resulting in a reduction in computation time. Subsequently, a variety of feature selection methods [18], including ANOVA and a hybrid method combining ANOVA and the Forward Selection Algorithm, were used to identify the most informative features and avoid overfitting for each classification. It was observed that the performance of the models slightly varied depending on the feature selection methods used and data imputation method. However, the primary focus was on ANOVA feature selection and KNN imputation.

We explored three different scenarios of classification based on class numbers, including 4-label, binary, and 3-label classifications. In the first scenario, we classified all four subtypes of breast cancer simultaneously. In the second scenario, we used different methods to make binary classifications, including one vs. the rest, one vs. one, and two vs. two. In the last scenario, we considered three-label classification using different methods of either eliminating or combining one class.

To develop our breast cancer classification model, we investigated various classifiers, such as support vector machines (SVM), logistic regression (LR), random forests (RF), XGBoost (XGB), linear discriminant analysis (LDA), and neural networks (NN). We utilized grid search or random search [19] with cross-validation on the training dataset to fine-tune the hyperparameters of these classifiers.

To ensure reliable results and avoid overfitting, we split the data into training and testing sets using a 90% – 10% ratio. We repeated this process five times by varying the random state each time. Finally, we constructed a 90% confidence interval for the average number of features and the F1 score using the t-student distribution with four degrees of freedom.

Our study demonstrates the effectiveness of our approach in predicting the molecular subtypes of breast cancer using DCE MRI data and extracted features. We achieved up to 85% F1 score in binary classifications. Moreover, Our approach achieved great performance compared to the original research [3] in multi-classification, where all four molecular subtypes available. The overall accuracy of classifying the four molecular subtypes of breast cancer was improved by 12%, indicating a considerable improvement over the original study that utilized the same dataset. We observed that the precision and recall values were consistent across all subtypes, indicating that our approach is effective in predicting all subtypes with similar accuracy. In conclusion, our study highlights the potential of machine learning techniques in improving the diagnosis and classification of breast cancer subtypes using non-invasive imaging data. By accurately predicting the molecular subtypes of breast cancer, our approach can lead to the development of personalized treatment plans and improve patient outcomes. However, further validation studies are necessary to assess the generalizability of our approach across different datasets and populations. To advance the field of breast cancer diagnosis and classification, future research should focus on developing machine learning models that are both interpretable and robust. Additionally, using larger datasets or exploring different types of MRI scans could help improve the accuracy and generalizability of these models.

This thesis makes several contributions to the field of breast cancer diagnosis and classification using DCE MRI and machine learning algorithms. These contributions include:

1. Demonstrating that machine learning algorithms can accurately distinguish between molecular subtypes of breast cancer using DCE MRI data, providing a non-invasive and efficient method for predicting cancer subtypes.

2. Reducing the workload of radiologists by providing methods for predicting molecular subtypes of breast cancer with greater precision and recall, which can help reduce errors and potentially prevent irreparable damages.

3. Identifying the most important classes of features that can be used to generate features from DCE MRI data, as well as determining certain clinical features that are effective in predicting cancer types.

4. Evaluating previous research and ensuring reproducibility using a well-known dataset in breast cancer DCE MRI, which can help to establish the validity and reliability of the findings and improve the overall quality of research in the field.

This thesis comprises the following chapters:

- **Chapter 1 -** Introduction: An overview of the research problem, objectives, and context.

- **Chapter 2 -** Background and Related Work: A review of recent research on applying machine learning algorithms to predict molecular subtypes of breast cancer using DCE MRI.

- **Chapter 3 -** Methodology: A detailed description of the steps involved in processing raw DCE MRI data to prepare it for use with machine learning algorithms.

- **Chapter 4 -** Results and Discussion: A review of the achieved results for three classification scenarios, along with a discussion of the findings and their significance.

- **Chapter 5 -** Conclusion: A summary of the major findings in relation to the research objectives and questions, along with a brief overview of the study's limitations and future directions for research.

By reading these chapters, readers will gain a comprehensive understanding of the research problem, the methods used, the results achieved, and the implications of the findings.

Chapter 2

Background and Related Work

Breast cancer is a complex and heterogeneous disease with various molecular subtypes that exhibit distinct clinical presentations and responses to treatment [20]. Accurate and early prediction of molecular subtypes can help in developing personalized treatment plans and improving patient outcomes. Dynamic contrast-enhanced magnetic resonance imaging (DCE MRI) [13, 21] is a non-invasive imaging technique that can provide valuable information for predicting molecular subtypes of breast cancer. Machine learning algorithms have shown promising results in predicting molecular subtypes of breast cancer using DCE MRI data on extracted features. Several studies have investigated the use of machine learning algorithms on various extracted features from medical images to predict molecular subtypes of breast cancer.

Saha et al. [3] analyzed 922 patients with invasive breast cancer and pre-operative MRI using a computer algorithm to extract 529 features of the tumor and surrounding tissue. Machine-learning-based models were trained and evaluated to predict molecular, genomic, and proliferation characteristics. The models showed promising results in predicting Luminal A subtype, Triple Negative breast cancer, ER status, and PR status, but were limited to binary classifications with AUC of 0.69 for Luminal A vs. other subtypes and 0.566 for Luminal B vs. other subtypes. This inspired further research into the MRI dataset to improve the results, not only in binary states but also in multi-class classifications. This thesis utilized the same DCE MRI dataset; However, only high-quality images from a subset of 200 patients were considered in this study.

Sutton et al. [22] conducted a retrospective study to differentiate breast cancer molecular subtypes using machine-learning-based models with features extracted from MR images of 178 patients. The study achieved an accuracy of 69.9%, 62.9%, and 81.0% for the three subtypes considered. However, limitations of the study included the use of in-house software for feature extraction and insufficient details regarding the machine learning methods and reproducibility.

Son et al. [23] conducted a study aimed at predicting molecular subtypes of breast cancer using a radiomics signature developed from synthetic mammography reconstructed from Digital Breast Tomosynthesis (DBT). The study included 365 patients with three subtypes, and the radiomics signature achieved an area under the curve (AUC) of 0.838, 0.556, and 0.645 for the TN, HER2, and luminal subtypes, respectively. To obtain radiomics features, the researchers segmented regions of interest

(ROIs) using an open-source software called “PyRadiomics”. A total of 72 radiomics features were obtained for each view. The researchers used the elastic-net approach to select appropriate features and to build the machine learning model. Parameter tuning of the elastic-net was performed through ten-fold cross-validation. The radiomics signature was the only independent predictor of the molecular subtype, and its combination with clinical features improved the accuracy of distinguishing the TN subtype. This indicates that the radiomics signature has the potential to serve as a biomarker for TN breast cancer treatment direction.

Sun et al. [12] utilized an ensemble learning based prediction model to distinguish between luminal and non-luminal breast cancer subtypes, and achieved an 85.2% accuracy using 5-fold cross-validation.

Fan et al. [5] extracted 90 features from DCE MRI, including 88 imaging features related to morphology and texture, dynamic features from tumor and BPE, and 2 clinical information-based parameters (age and menopausal status). The study used the Weka software platform for data mining and machine learning. A logistic regression model was used for discrimination of the four molecular subtypes. The Kruskal-Wallis test was used to test the statistical significance of the selected features across the subtypes. An evolutionary algorithm-based optimization method was used to search for optimal feature subsets for classification. The performance of the classifier was evaluated using a receiver-operating characteristic analysis, and the area under the curve was computed. The EA chromosome with the highest AUC was selected to establish the optimal feature pool and build the optimal classifier. The classifier achieved an AUC value of 0.869 with high overall classification performance.

While machine learning algorithms on extracted features have shown promising results in predicting molecular subtypes of breast cancer using DCE MRI, there are still some challenges that need to be addressed. One of the major challenges is the limited availability of high-quality DCE MRI data for breast cancer patients. Another challenge is the lack of standardization in the methods used to define molecular subtypes. In addition, there is a need for further validation studies to assess the generalizability of the machine learning models across different datasets and populations. Further research is needed to address the challenges and improve the accuracy and reliability of these algorithms. With the increasing availability of DCE MRI data and advancements in machine learning techniques, these algorithms have the potential to revolutionize the diagnosis and treatment of breast cancer. However, it is important to note that these studies are limited by their small sample sizes and need to be validated in larger cohorts. Furthermore, the performance of these algorithms needs to be improved to achieve clinically acceptable sensitivity and specificity. Various feature classes and feature selection methods should be explored to improve the performance and interpretability of the machine learning models. The upcoming methodology chapter will provide detailed explanations of the data preprocessing and methods utilized in this study.

Chapter 3

Methodology

Breast MRI is a widely used imaging modality to evaluate the extent of disease in breast cancer patients. Recent studies have demonstrated its potential in predicting both short-term and long-term patient outcomes, as well as identifying pathological and genomic characteristics of tumors.

In this study, we will utilize a well-known dynamic contrast-enhanced (DCE) MRI dataset[1] [24], which has been made available through the Cancer Imaging Archive (TCIA) [25], to differentiate between four molecular subtypes of breast cancer.

3.1 Data Description

In terms of design, the dataset is a single-institutional, retrospective collection of 922 biopsy-confirmed invasive breast cancer patients, over a decade, having the following

[1] `https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=70226903`

data components:

1. **Pre-operative dynamic contrast-enhanced (DCE) MRI:** The provided dataset includes axial breast MRI images acquired in the prone position using 1.5T or 3T scanners. The following MRI sequences are available in DICOM format (368.4 GB):

   - A non-fat saturated T1-weighted sequence
   - A fat-saturated gradient echo T1-weighted pre-contrast sequence
   - Mostly three to four post-contrast sequences

   To ensure data uniformity, a cohort of 200 patients with pre-contrast sequences and the first three post-contrast sequences were selected for this study. This decision was based on the fact that many patients in the larger pool of 922 patients did not have all the necessary sequences, or the image quality was inadequate for our project. By limiting our analysis to this subset of patients, we aimed to ensure the consistency and quality of the data used in our research.

   Pre-Contrast and Post-Contrast Sequences in Dynamic Contrast-Enhanced MRI for Breast Cancer:

   DCE MRI is a non-invasive imaging technique that is particularly useful for detecting and characterizing breast cancer because it can provide information about the blood vessels and blood flow within the breast tissue. The contrast agent used in DCE MRI is typically a gadolinium-based agent, which is injected into the patient's bloodstream through an intravenous (IV) line.

Before the contrast agent is injected, a fat-saturated gradient echo T1-weighted pre-contrast sequence is acquired. This sequence provides baseline information about the breast tissue, including its internal structure, fat content, and any existing lesions or abnormalities. The fat-saturation technique is used to suppress the signal from fat tissue, which can interfere with the detection of small lesions in the breast.

After the pre-contrast sequence is acquired, the contrast agent is injected, and a series of post-contrast sequences are acquired at regular intervals, typically every 60-120 seconds. These sequences capture the uptake and washout of the contrast agent in the breast tissue, providing information about the blood flow and vascular permeability in the tissue.

The post-contrast sequences are typically acquired using a fast, three-dimensional gradient echo sequence, with a high temporal resolution to capture the rapid changes in contrast enhancement that occur in the breast tissue. The images are acquired in the axial plane, which allows for the most accurate assessment of lesion size and location.

After the images are acquired, they are processed using specialized software to create a series of dynamic contrast-enhanced images. These images show the changes in contrast enhancement in the breast tissue over time, allowing radiologists to identify areas of abnormal blood vessel growth and leakage, which are often associated with breast cancer.

In addition to DCE MRI, other imaging techniques such as mammography, ultrasound, and magnetic resonance imaging (MRI) can also be used to detect and diagnose breast cancer. However, DCE MRI is particularly useful for detecting small lesions and assessing the extent of disease in the breast tissue, making it an important tool in the diagnosis and treatment of breast cancer.

2. **File Path mapping tables:** This CSV file[2] is used to filter out different sequences and slices for each patient.

3. **Location of tumors in DCE MRI:**

   The DCE MRI sequences in this dataset consist of between 50 and 220 slices providing a comprehensive view of the breast tissue. To isolate the slices that contain tumors, radiologists annotated the dataset by using a three-dimensional box to identify the precise location of the primary tumor. The annotations were recorded in a CSV file, with each row corresponding to a unique patient and indicating the start and end row, column, and slice numbers of the tumor box (six coordinates per patient).

   The annotations were performed in two parts by a panel of fellowship-trained radiologists using a graphical user interface developed in MATLAB. In the first

[2] https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=70226903

part, a panel of six radiologists annotated a subset of 271 patients by drawing boxes around any areas of mass and non-mass enhancement for up to five lesions. For the remaining 651 patients, a panel of four radiologists was formed, and the annotation procedure was modified to focus on the largest biopsied lesion. The radiologists had access to relevant radiology and pathology reports and the PACS system, if needed.

Overall, the annotations provide precise localization of the tumors in the pre-contrast, first post-contrast, and subtracted sequences, which can be used for further analysis and research purposes.

4. **Clinical and Other Features:** Apart from the Image Annotations CSV file, there is another file that contains various patient features. These features were collected from multiple sources, such as clinical notes, radiology reports, and pathology reports, and have been used in several previously published studies on radiogenomics, outcomes prediction, and other areas. For this study, only a subset of these features will be utilized, which will be further described in the following sections. The CSV file also includes the molecular subtypes of breast cancer, which will serve as the label for this study.

   Molecular subtypes of breast cancer: Breast cancer is a complex disease that can be classified into different molecular subtypes based on the expression of specific markers. The most commonly used markers to classify breast cancer are ER, PR, and HER2.

Luminal A breast cancer is characterized by the presence of ER and/or PR and the absence of HER2 expression. Luminal A tumors have a low proliferative index, meaning that they grow slowly, and are typically associated with a good prognosis. They are often treated with hormonal therapy, such as tamoxifen or aromatase inhibitors [26].

Luminal B breast cancer is also characterized by the presence of ER and/or PR, but in contrast to Luminal A, HER2 is positive and has a higher proliferative index. Luminal B tumors are associated with a worse prognosis than Luminal A tumors. Treatment for Luminal B breast cancer usually involves a combination of hormonal therapy and chemotherapy [27].

HER2+ breast cancer is characterized by overexpression of the HER2 protein but is typically ER and PR negative. HER2+ tumors are associated with a more aggressive disease course and a worse prognosis than luminal tumors. Treatment for HER2+ breast cancer usually involves a combination of chemotherapy and targeted therapy, such as trastuzumab or pertuzumab [28].

Triple Negative (TN) breast cancer is characterized by the lack of expression of ER, PR, and HER2. Triple Negative tumors are typically more aggressive and have a poorer prognosis than other subtypes of breast cancer. Treatment for Triple Negative breast cancer usually involves chemotherapy, as there are no targeted therapies currently available for this subtype [29].

In conclusion, molecular subtyping of breast cancer is crucial for personalized treatment and prediction of prognosis. Each subtype has its unique characteristics and treatment options, and understanding these subtypes can help guide clinical decision-making. It is also worth mentioning that in some instances, the literature may combine Luminal A and Luminal B subtypes into a single label known as “Luminal Like”.This is because these subtypes share some characteristics, such as positive ER and/or PR expression, but may differ in terms of HER2 expression and proliferation index. Therefore, the Luminal Like subtype may be used to describe a broader category of breast cancer that has a more favorable prognosis than HER2-positive or Triple Negative subtypes.

Table 3.1: Molecular subtypes based on ER, PR, HER2

ER	PR	HER2	Molecular Type
+	−	−	Luminal A
−	+	−	Luminal A
+	+	−	Luminal A
+	−	+	Luminal B
−	+	+	Luminal B
+	+	+	Luminal B
−	−	+	HER2+
−	−	−	Triple Negative

3.2 Data Collection

All components of the dataset were gathered from Cancer Imaging Archive (TCIA). To filter and download the images it is required to use NBIA Data Retriever[3]. Through this data retriever, the images related to 200 patients have been downloaded. It should be noted that using NBIA Data Retriever there are two options including Descriptive Directory Name and Classic Directory Name, which the latter should be selected. It is also possible to use some other applications and software to see the images in a three-dimensional space or even perform some processes such as data segmentation and cropping images. Some articles have utilized software to perform similar tasks [30–32]. However, in this project, the software was only used for visualization purposes, while all the cropping and segmentation processes were carried out using Python. These applications are as follows:

1. **3D Slicer[4]:** It is a free and open-source software package for image analysis and visualization in medical research [33]. It provides a platform for medical image processing and visualization, as well as for the creation of 3D models for surgical planning and simulation.

2. **RadiAnt DICOM Viewer[5]:** RadiAnt DICOM Viewer is a free and user-friendly software for medical image visualization and analysis in DICOM format.

[3] https://wiki.cancerimagingarchive.net/display/NBIA/Downloading+TCIA+Images

[4] https://www.slicer.org/

[5] https://www.radiantviewer.com/

It provides a simple interface for viewing, measuring, and analyzing medical images that can be used for diagnostic purposes, research, and education. In a 3-dimensional space, an image from the first post-contrast sequence of patient number 37 can be viewed.

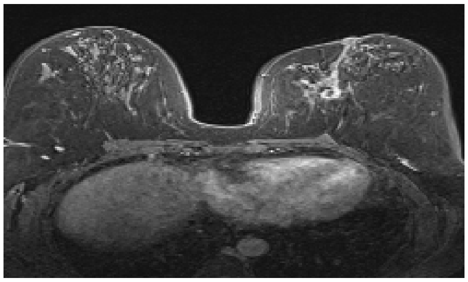

Figure 3.1: One single slice out of 144 slices (2D)

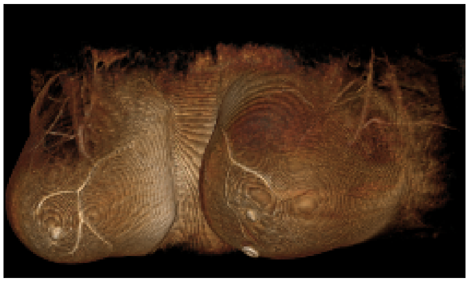

Figure 3.2: Combination of all the slices (3D)

The characteristics of slices

1. **Size:** There are three different sizes depending on the patients and attributes of MRI scanner. Slices are originally 320 × 320 or 448 × 448 or 512 × 512 images. In the Clinical and Other Features CSV file, columns seven and eight indicate the pixel dimension for each patient.

2. **Color:** All the slices are grayscale images. Grayscale images are digital images that only contain shades of gray, without any other colors. They are commonly used in fields such as photography, medical imaging, and printing due to their ability to accurately represent brightness and contrast. Grayscale images are

represented by a two-dimensional pixel matrix, unlike color images which have three or four dimensions. The simplicity of grayscale images makes them easy to work with and analyze, and they require less storage space compared to color images.

3. **Extension:** All the slices are originally in DICOM (Digital Imaging and Communications in Medicine) extension [34]. DICOM is a widely used standard for storing and transmitting medical images. This extension enables the customization of DICOM files to meet specific requirements of various medical imaging applications. By using DICOM extension, medical professionals can enhance their ability to effectively manage and analyze medical images, ultimately leading to improved patient care. However, DICOM images require significant storage space due to their high resolution and complex data structures. To reduce storage requirements and make the images more accessible, the original DICOM images were converted to PNG format which offers high quality and compression. Despite the conversion, the essential diagnostic information is retained and the images can be analyzed with the same level of accuracy as the original DICOM images.

3.3 Data Filtration

The original dataset was found to be highly imbalanced, with only 59 patients having the HER2+ subtype and some cases lacking post-contrast 3 images. To address this, we selected a subset of 200 patients from a larger pool of 922 biopsy-confirmed invasive breast cancer patients, ensuring that each molecular subtype was represented by 50 patients to achieve a completely balanced dataset. This allowed us to avoid the need for resampling methods. We also ensured that all images used in the study were of high quality by selecting only the DICOM images that did not primarily consist of white or black slices, as these can reduce image quality. The table below shows the selected patients and their corresponding labels. All selected patients were stored in four RAR files, each containing all sequences. However, some patients did not have post-contrast 3 or higher quality DICOM images, so we filtered out all but four sequences: pre-contrast, post-contrast 1, post-contrast 2, and post-contrast 3. To obtain the final sample of 200 patients, we used a File Path mapping tables file to exclude unwanted patients, and an Image Annotations CSV file to select only the slices that involved tumors, with patient IDs in the first column and start and end slice information in the sixth and seventh columns, respectively. We narrowed down further by considering only three middle slices for each patient, reducing computation time while extracting features. After converting the DICOM images to PNG format, we saved the selected slices in a separate folder. This process was performed for all four sequences, resulting in a total of 600 tumor-containing slices per sequence.

Label	**Patient Number**
Label0 (Luminal A)	3,4,6,18,19,21,23,25,26,33,34,35,41,49,72,73,120,121,124,125,126,127,138,145,165,199,207,212,224,291,312,316,354,355,356,357,358,412,413,434,440,441,447,512,528,600,601,602,690,691
Label1 (Luminal B)	8,22,27,51,56,86,119,133,134,141,143,146,177,214,222,239,262,302,308,328,345,348,353,369,411,466,473,491,541,544,552,578,608,610,622,663,666,693,707,722,782,784,812,822,826,847,867,883,885,891
Label2 (HER2+)	1,17,44,68,74,96,105,144,159,161,179,202,297,306,321,344,366,373,386,395,400,403,422,429,431,444,468,470,519,525,558,559,567,577,604,674,679,683,687,748,758,763,775,797,835,838,873,894,903,907
Label3 (Triple Negative)	9,10,11,20,37,42,43,48,59,64,77,78,97,99,106,132,148,158,172,178,187,209,211,218,219,220,362,375,398,438,456,460,463,478,482,485,514,520,523,565,568,631,636,645,656,662,723,871,906,908

Table 3.2: Patient Identifiers and Labels: Fixed Numbers Assigned for the Study

3.4 Cropping the Images

Even after excluding the slices that do not contain the tumor, we still need to process the remaining slices since only a small portion of them involves the tumor. To achieve this, we crop the slices to focus on the tumors. The Image Annotations CSV file includes the start and end coordinates for the tumor's location, with the second to fifth columns representing the start row, end row, start column, and end column, respectively. Using these four coordinates, we draw a rectangle around the tumor using one of three methods:

1. **Original Crop:** We drew a rectangle around the tumor using the coordinates provided by the radiologist. As the sizes of the rectangles varied, we resized the cropped images to a uniform size of 64×64 pixels.

2. **32×32 Crop:** In some cases, the rectangles drawn by the radiologist did not contain the tumor effectively. To improve this, we drew other squares of equal sides that fit better around the tumors. We found the center coordinates of the original rectangles and drew four lines with a size of 16 pixels from the center to determine the locations of the new squares. These squares often worked better than the original crops, although they were sometimes too small to contain the tumors completely. To address this, we also used 64×64 pixel crops. It should be noted that we did not resize the images in this method of cropping, as all the squares were already of the same size of 32×32 pixels.

3. **64×64 Crop:** In cases where breast cancer tumors grew and extended to involve adjacent organs, the previous cropping methods were not effective. In these situations, 64 × 64 pixel crops worked well when the tumors were not concentrated in a small area. We found the center coordinates of the original rectangles and drew squares of equal sides with a size of 64 × 64 pixels around the tumors.

We repeated these processes for all 4601 slices in each of the four sequences and saved the cropped images in separate folders within the directory.

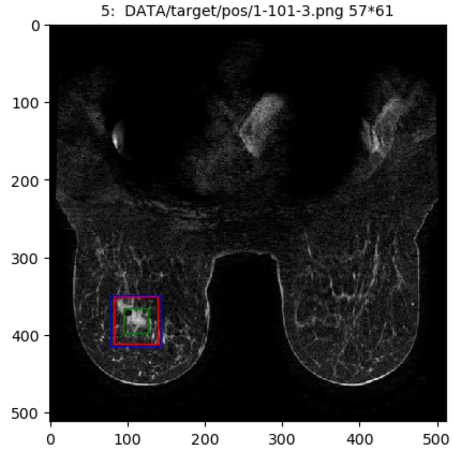

Figure 3.3: Pre-contrast sequence, Patient #3, Slice#101

Red: Original (57×61)

Green: 32×32

Blue: 64×64

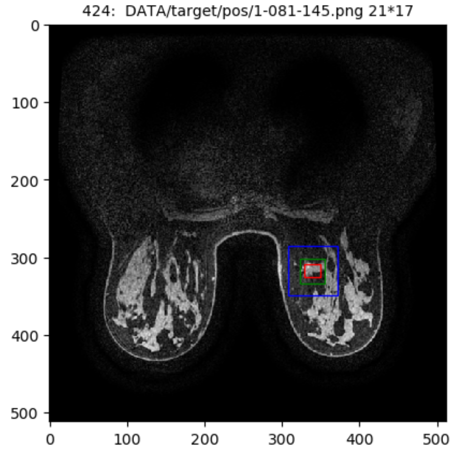

Figure 3.4: Pre-contrast sequence, Patient #145, Slice#81

Red: Original (21×17)

Green: 32×32

Blue: 64×64

3.5 Data Segmentation Using Masks

3.5.1 Image segmentation

Medical image segmentation involves dividing an image into regions or segments that correspond to specific anatomical structures or tissue types, which is a critical task in aiding clinicians to accurately diagnose various diseases and conditions. In dynamic contrast-enhanced magnetic resonance imaging (DCE MRI), segmentation is particularly important for analyzing the uptake and washout of contrast agents in tissues over time. However, the complexity and heterogeneous nature of the data, along with the presence of noise and artifacts, make segmentation challenging. Fortunately, there are various segmentation techniques developed specifically for medical images, including thresholding, region growing, active contours, and machine learning-based methods such as convolutional neural networks (CNNs) and clustering. In our study, we tried several segmentation methods and found that the thresholding method was the most effective.

With accurate segmentation of the cropped DCE MRI images, we can focus on detecting lesions or areas of breast tissue suspected to have cancer. This enables higher accuracy in extracting features and classifying breast cancer into four molecular subtypes, which is crucial for determining the most effective treatment plan.

3.5.2 Generating Masks Using Thresholding

The thresholding method is commonly used to segment medical images, including those obtained through DCE MRI [35–37]. In our study, we applied this method to the cropped DCE MRI images by setting a threshold value for the grayscale intensity of the image. Pixels with intensity values above this threshold were retained, while those below were discarded. This process effectively separated the image into foreground (object) and background regions.

The choice of threshold value is an important hyperparameter, and tuning it can potentially improve segmentation results. For our study, we established a threshold value of T=50 through sample segmentation, which led us to consider only grayscale pixels in the cropped images with intensity values higher than this cutoff. This approach allowed us to generate a binary matrix for each cropped image, where the elements were either zero or one. A value of one indicated that the corresponding pixel had an intensity value higher than 50 and would be used for feature extraction. This method was highly effective in identifying the region of interest (ROI), such as lesions or areas of breast tissue suspected to have cancer.

Experimenting with a broader range of sample segmentation techniques and using different classes of masks could help determine optimal threshold values and extract features more effectively. [38]. However, this would be prohibitively expensive due to the high number of slices and the vast array of features that would need to be extracted.

3.6 Feature Extraction

In image classification problems, there are two main approaches: using pixel matrices directly to feed machine learning classifiers or extracting features to improve results. For this study, we focused on feature extraction as it provides uniformity for all images and can significantly improve results. To extract features, we used Pyfeats[6], a Python library that covers a wide range of feature classes, including textural, morphological, histogram-based, multi-scale, and moment-based features. Pyfeats generates these features using cropped images and their corresponding tumor segmentation as inputs. However, some features generated NA values, requiring data imputation. To reduce dimensionality while retaining informative features, different feature selection processes were explored. Ultimately, the feature extraction and selection process identified the most informative features, which were used to classify medical images with high accuracy. In the following section, we will explore the main feature classes extracted through Pyfeats.

1. First Order Statistics (FOS): FOS are statistical measures that describe the distribution of intensity in an image. These features are computed from the histogram of the image, which represents the probability density function of individual pixels. FOS features include mean, standard deviation, median, mode, skewness, kurtosis, energy [39], entropy [40], minimum and maximum gray levels, coefficient of variation, percentiles, and histogram width.

[6] https://pypi.org/project/pyfeats/

2. The Gray Level Co-occurrence Matrix (GLCM): A set of features proposed by Haralick that estimates the second-order joint conditional probability density functions [41]. GLCM features include measures such as angular second moment, contrast, correlation, sum of squares: variance, inverse difference moment, sum average, sum variance, sum entropy, entropy, difference variance, difference entropy, and information measures of correlation.

3. Gray Level Difference Statistics (GLDS): A texture extraction algorithm that uses first-order statistics of local property values based on absolute differences between pairs of gray levels or of average gray levels. GLDS features include measures such as homogeneity, contrast, energy, entropy, and mean [42]. These features provide information about the texture of an image.

4. The Neighborhood Gray Tone Difference Matrix (NGTDM): It is a method of feature extraction that captures visual properties of texture [43]. NGTDM features include measures such as coarseness, contrast, busyness, complexity, and strength.

5. The Statistical Feature Matrix (SFM): It is a method of measuring the statistical properties of pixel pairs at various distances within an image for statistical analysis [44]. SFM features include measures such as coarseness, contrast, periodicity, and roughness.

6. Law's Texture Energy Measures (LTE/TEM): These are derived from three

simple vectors that are convolved with themselves to obtain masks of different sizes [45]. These masks are convolved with an image to extract texture features, such as texture energy from LL, EE, and SS kernels, as well as average texture energy from LE and EL, ES and SE, and LS and SL kernels.

7. Fractal Dimension Texture Analysis (FDTA): It estimates the roughness of natural surfaces using the Fractional Brownian Motion (FBM) Model [46, 47]. FDTA uses parameters such as fractal dimension or Hurst coefficient to represent a fractal surface and obtain a multiresolution fractal (MF) feature vector by observing the image at different resolutions.

8. The Gray Level Run Length Matrix (GLRLM): GLRLM is a feature extraction method that identifies consecutive picture points with the same gray level value [48]. GLRLM features include measures such as short-run emphasis, long-run emphasis, gray level non-uniformity, and run length non-uniformity.

9. The Fourier Power Spectrum (FPS): It is a method of feature extraction that uses the discrete Fourier transform for digital images [49]. Texture features are based on ring-shaped or wedge-shaped samples of the discrete Fourier power spectrum and include measures such as radial sum and angular sum.

10. Shape Parameters: They are a set of features that describe the geometric shape of an object in an image [50]. These features include measures such as x-coordinate maximum length, y-coordinate maximum length, area, perimeter,

and perimeter2/area.

11. The Gray Level Size Zone Matrix (GLSZM): It is a feature extraction method that quantifies gray level zones in an image based on connected voxels. GLSZM features include measures such as small zone emphasis, large zone emphasis, gray level non-uniformity, and zone-size non-uniformity [51].

12. Higher Order Spectra (HOS): Radon transform turns 2D images into a line parameter domain [52, 53]. HOS captures higher moments of a signal, with bispectrum as its Fourier transform. Bispectrum is symmetric and calculated in non-redundant region, with entropy as the extracted feature.

13. Local Binary Pattern (LBP): LBP is a texture descriptor that converts pixels to circular bit-streams [54]. Uniformity is used to identify patterns with few transitions. Energy and entropy of LBP images at different scales are used as feature descriptors.

14. Gray-scale Morphological Analysis: This method extracts geometric properties of components in multilevel binary morphological analysis [55, 56]. It involves generating three binary images and calculating pattern spectra. Grayscale morphological features include mean CDF and mean PDF of pattern spectra using the cross as a structural element.

15. Histogram: The histogram is the distribution of grey levels in the region of interest (ROI) of an image.

16. Multi-region Histogram: It involves identifying equidistant regions of interest (ROIs) by eroding the image outline based on its size [57]. The histogram is then computed for each of these ROIs separately.

17. Amplitude Modulation – Frequency Modulation (AM-FM): AM-FM involves multi-scale representations of images using least-square approximations [58]. It calculates instantaneous amplitude, phase, and frequency for specific image components. The input image is processed through bandpass filters, producing IA, IP, and IF for each block. AM-FM features include histograms of reconstructed images at different scales: low, medium, high, and dc.

18. Discrete Wavelet Transform (DWT): DWT uses inner product with a function family to transform signals. For 2D signals like images, 2D DWT can be used by applying DWT on rows and columns, followed by down-sampling [59]. This yields four sub-images at each level, which are further decomposed into approximation and detail sub-images, each created by convolution with half-band filters. DWT features include mean and standard deviation of the absolute value of detail sub-images.

19. Gabor Transform (GT): GT convolves an image with a Gabor function, which is a sinusoidal plane wave with a certain frequency and orientation modulated by a Gaussian envelope [60]. Gabor filters have frequency and orientation representations similar to the human visual system, making them useful for texture

segmentation and classification. GT features include mean and standard deviation of the absolute value of detail sub-images.

20. Zernikes' Moments: Zernikes' Moments are orthogonal complex moments used in image processing, computer vision, and related fields [61]. They are based on a set of complete orthogonal polynomials defined over the unit disc in polar coordinate space. Zernike's Moments consist of 25 orthogonal moments invariants with respect to translation.

21. Hu's Moments: Hu's Moments are a set of image moments used in image processing, computer vision, and related fields [62]. They consist of 7 moments that are invariant with respect to translation, scale, and rotation.

After generating 369 features for each cancer-containing slice using the methods described earlier, we found that some of the features contain missing (NA) or infinite (Inf) values. To address this issue, we applied data preprocessing techniques to handle missing and infinite values.

3.7 Data Preprocessing

3.7.1 Data Averaging

To address the issue of having too many cancer-containing slices in our dataset, we implemented a data-averaging approach. Specifically, we only considered three middle slices out of all cancer-containing slices. After performing feature generation, we took the average over the extracted features from these three slices and created a single sample per patient. This approach reduced the sample size from 600 to 200, which can help improve the computational efficiency of our models while still retaining the relevant information from the original dataset.

3.7.2 Data Splitting

To ensure reliable results and avoid overfitting, we split the data into training and testing sets using a $90\% - 10\%$ ratio. We repeated this process five times by varying the random state each time. We performed all data preprocessing steps after splitting the data to prevent test data from influencing the models' generalization abilities. By evaluating the models on unseen test data, we could determine their accuracy and effectiveness.

3.7.3 Data Scaling

To improve the performance of machine learning algorithms, it is important to normalize or standardize features in datasets containing different scales. Scaling features can help reduce computation time and improve model accuracy. For normalization, we performed standardization on the dataset since most features appeared to have a normal distribution based on their histogram. It is possible to confirm normality using statistical non-parametric tests such as the Shapiro-Wilk and Kolmogorov-Smirnov tests [63]. During standardization, we calculated the mean and standard deviation of each feature on the training data and scaled all features to have a mean of 0 and standard deviation of 1. We applied the same transformation to the test data to ensure the model was not influenced by information from the test set during scaling.

3.7.4 Data Cleaning

To address missing and infinite values in the dataset, we employed three techniques: mean imputation, KNN imputation, and dropping features with missing or infinite values [64]. Additionally, we utilized Local Outlier Factor (LOF) to detect and address the outliers. We evaluated the computational efficiency and effectiveness of each method, and found that KNN imputation with 10 nearest neighbors showed the best performance, so we focused on this method for imputation. We emphasized the importance of selecting an appropriate threshold for generating masks, as it can significantly impact how missing values are handled when generating features and affect subsequent analyses.

1. **Mean Imputation:** Mean imputation is a commonly used method for handling missing data in datasets and performed well in our study without requiring hyper-parameter tuning. However, it may introduce bias and reduce variability. Replacing infinite values with the maximum feature value had minimal impact on the results as they were rare in the datasets (with at most two per dataset).

2. **KNN Imputation:** KNN imputation replaces missing data with estimates

based on the values of the K nearest neighbors in the dataset. It can handle both numerical and categorical variables but may be computationally expensive and requires careful consideration of K and distance metric. In our study, we performed a grid search with the values of 5, 10, 15, and 20 to determine the optimal value of K. Additionally, we replaced infinite values with the maximum feature value. While KNN outperformed other imputation methods, the optimal K value and choice of distance metric can vary based on the dataset characteristics.

3. **Dropping Features With Missing or Infinite Values:** Dropping features with missing or infinite values can be an effective way to handle missing data in datasets. In our study, we found that most datasets had at most 30 columns containing missing values, and we chose to drop these features without sacrificing too much information. Dropping features also prevented information leakage between the training and test sets. However, for the previous imputation methods, we performed imputation on the training data and applied the same transformation to the test data to avoid information leakage and ensure reliable results.

4. **Local Outlier Factor (LOF):** LOF is an unsupervised algorithm used to detect anomalies in datasets by measuring the local density of points and identifying those with significantly lower densities as potential outliers [65]. The two key hyperparameters in LOF are contamination and n−neighbors. The contam-

ination determines the expected percentage of outliers in the dataset, while n−neighbors determines the number of neighbors to consider when computing the local density of each point. A higher value of n−neighbors can improve the algorithm's robustness to noise and density fluctuations but increase computational complexity. In our study, we set contamination to 'auto'and n−neighbors to 10, and applied LOF only to the training data without altering the test data. We also found that applying LOF over the selected features after feature selection was more effective than applying it over all the features.

3.8 Data Integration

We used four sequences (pre-contrast, post-contrast1, post-contrast2, and post-contrast3) and three different methods of cropping (original, 32×32, and 64×64) to extract 369 features for each combination, resulting in a total of 4428 (12×369) features per patient. This dataset accurately represents the molecular subtypes of breast cancer, and the study achieved great results using this dataset. Additionally, we added 23 clinical features from the Clinical and Other Features CSV file to each dataset. These features are listed below:

- **MRI technical information:** days to MRI (from the date of diagnosis), manufacturer, manufacturer model name, field strength (tesla), patient position during MRI, contrast agent, TR (repetition time), TE (echo time), acquisition matrix, slice thickness, flip angle, field of view

- **Demographics:** date of birth (days), menopause at diagnosis, race and ethnicity

- **Tumor characteristics:** tumor location (side of cancer: left: -1,right: +1,not given: 0)

- **MRI findings:** multicentric/multifocal, contralateral breast involvement, lymphadenopathy or suspicious nodes

- **SURGERY:** days to surgery (from the date of diagnosis)

- **Recurrence:** recurrence events (no: 0, yes: 1)

- **Follow up:** days to last local recurrence free assessment (from the date of diagnosis), age at last contact

Although technical information regarding MRI machines may not be directly related to the molecular subtype of breast cancer, they have been considered as candidates for feature selection to mitigate the impact or noise generated by the machines. Features related to various types of treatments and those closely related to molecular subtypes, such as tumor grades, were not selected for inclusion in the dataset. More detailed information regarding feature values can be found in the aforementioned CSV file.

3.9 Feature Selection

Feature selection involves identifying and selecting the most relevant and informative subset of features from a larger set for a given machine learning task. This process offers several advantages, such as reducing data dimensionality, improving model accuracy and interpretability, reducing overfitting, and speeding up training. By selecting only the most important features, we can enhance the efficiency and effectiveness of the machine learning algorithm, and reduce the complexity and resource requirements of the model. Additionally, feature selection helps to identify and remove redundant, irrelevant, or noisy features that can negatively impact model performance.

In this study, three initial feature selection methods were used to eliminate redundant features in each dataset. Afterwards, two final feature selection methods were separately applied to filter the remaining features and identify the most suitable ones for accurately classifying the molecular subtypes of breast cancer [66–70].

3.9.1 Initial Feature Selections

In order to reduce computation time for the major feature selections, it is necessary to initially drop some redundant features. These features may have very low variance or high correlation with other features.

3.9.1.1 Features with low variance

Some features have a very small variance or even are constant, meaning they are not informative and can be removed without losing any information. A conservative threshold of std=0.01 was selected to filter out these features. Some features have maximum values of 10^{-24}, and not filtering them out could cause scaling problems.

3.9.1.2 Features with low variety

There are other features where a large number of values are the same and cannot be useful for classifying cancer types. These features may not necessarily have low variance but do not convey any information. A conservative threshold of $P_{0.05} = P_{0.95}$ was used to filter out these features by removing those where the 5th percentile and 95th percentile values of the feature's distribution are the same.

3.9.1.3 Features with strong correlation

Some features are strongly correlated, meaning that only one of them is needed for the major feature selection. Pearson correlation was used to measure the linear relationship between two variables, with values ranging from -1 to 1. A strict threshold of 0.98 was selected to reduce computation time without significant information loss. Only one feature was selected from among those with a correlation greater than or equal to 0.98 to proceed with the final feature selection process. The formula for Pearson correlation is provided below:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where:

r_{xy} is the Pearson correlation coefficient between variables x and y.

n is the number of observations. x_i and y_i are the values of variables x and y for the i_{th} observation, respectively. $\bar{x}$ and $\bar{y}$ are the means of variables x and y, respectively.

3.9.2 Final Feature Selections

After performing an initial feature selection to reduce computation time, we explored two additional feature selection methods: ANOVA and a hybrid approach that combines ANOVA with forward selection. The goal was to identify the most important features for classifying the molecular subtypes of breast cancer. ANOVA is a statistical test that measures the significance of the differences between the means of different groups, and it can be used to identify features that are significantly associated with the outcome variable. Forward selection is a wrapper-based feature selection method that starts with a single feature and iteratively adds the best-performing feature until a stopping criterion is met. The hybrid approach first uses ANOVA to identify a subset of potentially relevant features and then applies forward selection to further refine the subset. By combining these two approaches, we aimed to identify a smaller set of highly informative features for breast cancer subtype classification.

3.9.2.1 ANOVA Feature Selection

ANOVA feature selection is a well-known method for selecting the most important features for a classification problem. The method involves calculating the F score, which measures the difference between the mean of the feature values for each class and the variance within each class. The higher the F score, the more relevant the feature is for classification. The F score can be calculated using the following formula:

$$F = \frac{\text{MSB}}{\text{MSW}} = \frac{\frac{\sum_{i=1}^{k} n_i(\bar{x}_{i.} - \bar{x}_{..})^2}{k-1}}{\frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_{i.})^2}{N-k}}$$

where k is the number of classes, n_i is the number of observations in the i^{th} class, N is the total number of observations, $\bar{x}_{i.}$ is the mean of the i^{th} class, $\bar{x}_{..}$ is the overall mean, and x_{ij} is the j^{th} observation in the i^{th} class.

A high F score or equivalently low p-*value* indicates that the feature is statistically significant and should be included in the model. To find the optimal number of features, a for loop can be used to try different numbers, typically ranging from one to 150. However, for some computationally intensive classifiers, it may be necessary to consider lower values of iteration to ensure efficient performance. For each iteration of the loop, grid search or random search with cross-validation can be used to tune the hyperparameters and evaluate the performance of the model. The goal is to find the set of features that provides the best balance between accuracy and efficiency.

The ANOVA feature selection method can help to reduce the dimensionality of the data and improve the accuracy of the classification model. By selecting only the most relevant features, we can reduce the noise and improve the interpretability of the model. Additionally, the method can help to identify and remove redundant or irrelevant features that may negatively impact the performance of the model.

3.9.2.2 Hybrid Feature Selection

This method consists of a combination of ANOVA feature selection and forward feature selection. First, we start by setting k to 150 in ANOVA feature selection, which selects the most important features based on the F score. Then, we use a for loop to try different numbers of features, from 1 to 100, using forward feature selection. Forward feature selection is an iterative method that starts with an empty set of features and adds one feature at a time, based on the performance improvement on the validation set [71]. At each iteration, we evaluate the performance of the model using grid search or random search with cross-validation to tune the hyperparameters and select the best set of features. This process continues until we reach the desired number of features or the performance improvement becomes negligible. To evaluate the performance of the model, we use the test data to calculate the F1 score as the metric for each set of features. The hybrid feature selection method combines the benefits of ANOVA feature selection and forward feature selection to identify the best features for classification. This combined approach can enhance the accuracy and efficiency of the classification model. Nevertheless, it is a time-consuming process and can be prohibitively expensive for some machine learning classifiers. Therefore, the focus of this thesis is on the first final feature selection method, ANOVA.

In both feature selection methods, we explored a variety of machine learning classifiers to evaluate their performance in classifying breast cancer subtypes. The details of these classifiers will be discussed in the next section. By comparing the performance of multiple classifiers on the selected features, we aimed to identify the best model for breast cancer subtype classification.

3.10 Classification Using ML Algorithms

This section provides an overview of the machine learning (ML) algorithms utilized to classify breast cancer subtypes based on the selected features. We explored several well-established classifiers, including support vector machines (SVM), logistic regression (LR), linear discriminant analysis (LDA), random forests (RF), extreme gradient boosting (XGB), and neural networks (NNs). These classifiers have been widely used for classification tasks in the biomedical field and have demonstrated promising results in various studies. Our goal is to compare the performance of these ML classifiers on the breast cancer dataset and identify the most accurate model for breast cancer subtype classification. To ensure reliable results and avoid overfitting, we split the data into training and testing sets using a $90\% - 10\%$ ratio. We repeated this process five times by varying the random state each time. This allowed us to construct a 90% confidence interval for the average number of features and the F1 score using the t-student distribution with four degrees of freedom. We presented the results in three different scenarios based on the number of classes. The F1 score was the main metric used to evaluate the models, which is a useful measure as it is an average between precision and recall, calculated with the following formula:

$$F1\ score = 2 \times \frac{precision \times recall}{precision + recall}$$

$$precision = \frac{true\ positives}{true\ positives + false\ positives}$$

$$recall = \frac{true\ positives}{true\ positives + false\ negatives}$$

3.10.1 ML Classifiers

1. **Support Vector Machines (SVM):** SVM is a supervised learning algorithm that is commonly used for both classification and regression tasks [72], which finds the hyperplane that maximally separates the classes in the feature space. SVM is capable of handling both linear and non-linear relationships by using kernel functions. To optimize SVM's performance specifically for breast cancer subtype classification, we employed a random search technique with 5-fold cross-validation to fine-tune the hyperparameters. This involved tuning hyperparameters such as kernel type, C, degree, and gamma on a training dataset. The optimal hyperparameters for multiclsification were kernel='rbf', C=0.5, gamma='scale', and degree=3, with an average for numerical parameters and mode for string ones.

2. **Logistic Regression (LR):** LR is a supervised learning algorithm that models the probability of the target variable using a logistic function, making it

well-suited for classification tasks [73]. The assumption of a linear relationship between the features and the log-odds of the target variable underpins LR. To optimize LR for classifying breast cancer subtypes, we employed a random search technique with 5-fold cross-validation to fine-tune its hyperparameters. This involved tuning hyperparameters such as penalty type, C, solver algorithm, and l1-ratio (if using elasticnet regularization). The optimal hyperparameters for multiclsification were penalty='elasticnet', C=0.05, solver='saga', and L1-ratio=0.65. We used an average for numerical parameters and mode for string ones.

3. **Linear Discriminant Analysis (LDA):** LDA is a supervised learning algorithm that finds the linear combination of features that maximizes the separation between classes in the feature space, making it well-suited for classification tasks [74]. It assumes that the covariance matrices of the classes are equal and that the features are normally distributed. To optimize LDA for breast cancer subtype classification, we used a grid search technique with 5-fold cross-validation to fine-tune its hyperparameters. This involved tuning hyperparameters such as solver, shrinkage, and n-components. We chose grid search over random search to explore hyperparameters comprehensively since LDA is a quick classifier. We found that the optimal hyperparameters for multiclsification were solver='lsqr', shrinkage=0.04, and n-components=1. We used an average for numerical parameters and mode for string ones. Additionally, we

increased the number of iterations of ANOVA feature selection to 250 for better performance.

4. **Random Forests (RF):** RF is an ensemble learning method that combines multiple decision trees to improve classification accuracy, making it well-suited for non-linear relationships and high-dimensional data [75]. To optimize RF for breast cancer subtype classification, we employed a random search technique with 5-fold cross-validation to fine-tune its hyperparameters, which included criterion, n-estimators, max-depth, min-samples-leaf, and min-samples-split. To reduce the computation time, we limited the number of iterations in random search to 200. We found that the optimal hyperparameters for multiclsification were n-estimators=200, criterion='gini', max-depth=6, min-samples-split=2, and min-samples-leaf=1. We used an average for numerical parameters and mode for string ones. RF performed well in our experiments, notably in binary classifications, and outperformed other classifiers.

5. **Extreme Gradient Boosting (XGB):** XGB is an ensemble learning method that improves model accuracy by combining multiple decision trees, making it well-suited for handling non-linear relationships and high-dimensional data [76]. It iteratively adds decision trees to correct errors. To optimize XGB for breast cancer subtype classification, we employed a random search technique with 5-fold cross-validation to fine-tune its hyperparameters, which included learning rate, n-estimators, max-depth, subsample, and regularization-lambda. Due to

the time-consuming nature of XGB and the large parameter spaces of methods like RF, we limited the number of iterations in the random search algorithm to 10. Additionally, we had to limit the iteration loops to 100 for feature selection to avoid affecting the performance of the ANOVA feature selection and prevent complications in the outer iteration. We found that the optimal hyperparameters for multiclsification were learning-rate=0.01, n-estimators=200, max-depth=3, subsample=0.8, and reg-lambda=0.07. We used an average for numerical parameters and mode for string ones.

6. **Neural Networks (NNs):** Neural networks are a type of machine learning model commonly used for classification tasks [77]. They consist of an input layer, one or more hidden layers, and an output layer. The hidden layers contain nonlinear activation functions that allow the network to capture complex relationships between the features and the target variable. Neural networks are capable of handling non-linear relationships and high-dimensional data, making them a popular choice in many applications. In this study, we experimented with several neural networks that had different structures and activation functions, as well as drop-out layers. We also used early stopping to prevent overfitting and reduce computation time. To optimize the performance of neural networks for classifying breast cancer subtypes, we performed a random search technique with 5-fold cross-validation to fine-tune the following hyperparameters: the number of hidden layers, the number of neurons in each layer, the acti-

vation function, and the dropout rate. To enable us to perform ANOVA feature selection, we constrained the iteration loops to 50. After conducting the random search, we found that the optimal hyperparameters for multiclsification were n-hidden-layer=4, activation=‘Relu’, hidden-layer-sizes=30, dropout=0.03, and early-stopping=15. We used an average for numerical parameters and mode for string ones.

3.11 Hardware Specifications

All programming and coding for this study were conducted using the Python language. To enable the processing of large MRI image data and the utilization of machine learning algorithms, we employed a high-performance computing server with the following hardware specifications. The server is equipped with 251 GB of RAM and features two powerful NVIDIA GPUs, each with a memory of 50 GB, as well as an integrated graphics controller with a memory of 256 MB. Both NVIDIA GPUs are Quadro RTX 6000/8000 models. These hardware specifications allowed us to perform computationally intensive tasks efficiently, contributing to the successful completion of this research project.

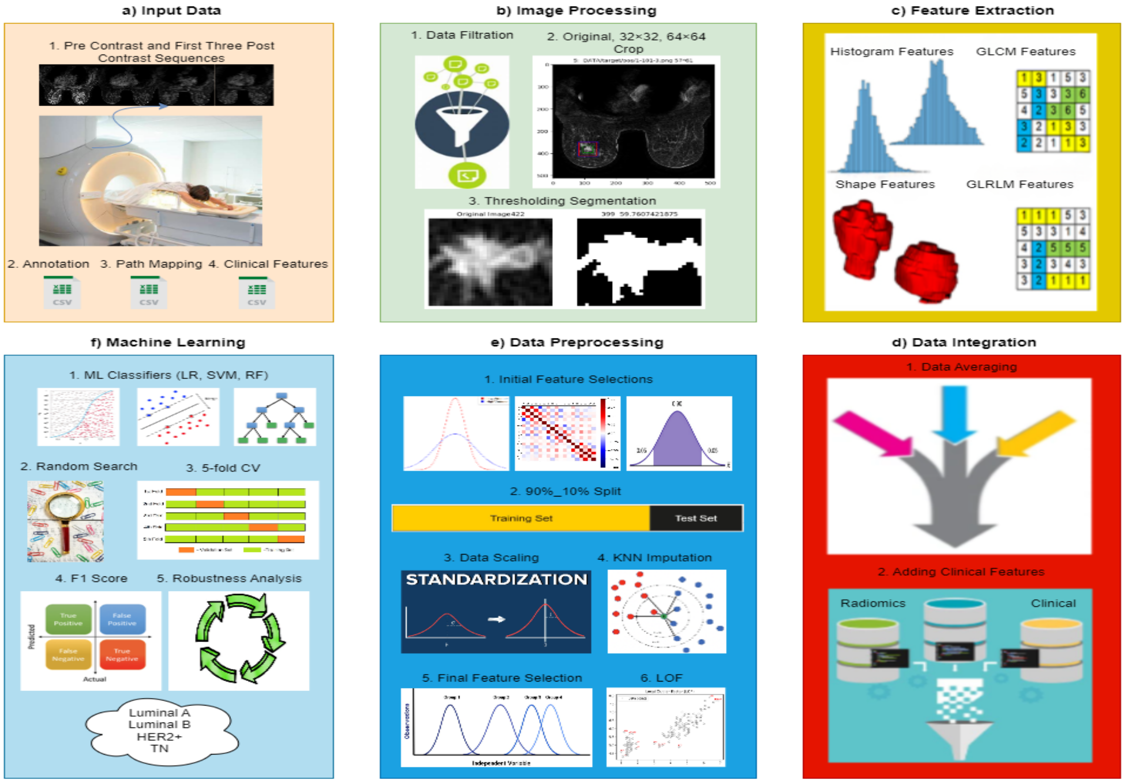

Figure 3.5: The proposed methodology pipeline for developing a machine learning model to classify molecular subtypes of breast cancer is shown. The pipeline consists of the following steps: (a) Input data, which includes images and three csv files; (b) Image processing, where unwanted slices and sequences are filtered out, and images are cropped with three different methods to segment them using thresholding; (c) Extraction of radiomics features using Pyfeats library in Python; (d) Merging clinical features with the average of three middle tumor-containing slices to create one sample per patient (n=200); (e) Data processing, including initial feature selections, data splitting, standardization, KNN imputation, ANOVA feature selection, and outlier detection using Local Outlier Factor (LOF); (f) Three machine learning classifiers are trained and hyperparameters are tuned using random search with 5-fold cross-validation on the training data. The performance of each model is evaluated on 5 different test subsets, with F1 score as the main metric.

Chapter 4

Results and Discussion

In this chapter, we present the findings of our research on breast cancer molecular subtypes using the dataset composed of 4428 extracted radiomics features and 23 clinical features. We presented the results in three different scenarios based on the number of classes. To be able to perform fare comparisons with the original study (Saha et al. [3]), we applied our framework to the dataset from Saha et al. consisting of 529 radiomics features and 200 samples with IDs given in Table 3.2. To evaluate the performance of our extracted features in both 4-label and one versus the rest classifications, we used the F1 score for all evaluations and maintained consistent settings across experiments, such as data cleaning methods, the number of folds in cross-validation, and the number of iterations in random search. Our results indicate that, in almost all cases, the machine learning classifiers trained on our dataset outperformed those reported in the original study.

4.1 4-label Classification

In the first scenario, we attempted to distinguish between all molecular subtypes of breast cancer simultaneously, which was a challenging task. Using random forest, we achieved the highest F1 score of 52.45%. Although the results in this case may not be very significant, there is a noticeable improvement compared to the original research.

	Our Study		**Saha et al.**
Classifier	**Avg No. of Features**	**Avg F1 Score**	**Avg F1 Score**
SVM	avg=51.60 90%CI=(9.36,93.84)	avg=46.37% 90%CI=(0.36,0.57)	avg=36.65% 90%CI=(0.35,0.38)
LR	avg=36.40 90%CI=(15.70,57.09)	avg=45.73% 90%CI=(0.39,0.53)	avg=38.14% 90%CI=(0.32,0.44)
LDA	avg=113.20 90%CI=(6.46,219.94)	avg=43.32% 90%CI=(0.34,0.53)	avg=38.30% 90%CI=(0.33,0.43)
RF	avg=30.69 90%CI=(48.53,107.06)	avg=52.45% 90%CI=(0.48,0.57)	avg=38.02% 90%CI=(0.36,0.40)
XGB	avg=51.00 90%CI=(23.16,78.85)	avg=48.05% 90%CI=(0.39,0.57)	avg=40.04% 90%CI=(0.36,0.43)
NNs	avg=27.98 90%CI=(3.27,52.65)	avg=49.15%, 90%CI=(0.39,0.59)	avg=34.07% 90%CI=(0.32,0.37)

Table 4.1: Comparison of six machine learning models in a 4-Label classification task: Evaluation of average F1 score and feature count across five random test subsets. A parallel analysis was performed on the features extracted by Saha et al., with identical settings for consistency in comparison.

Table 4.1 demonstrates that in multiclassification where all four molecular subtypes are available, the random forest algorithm achieves the highest performance, with an average F1 score of 52.45%. The average number of features across all classifiers ranges almost from 30 to 50, although the variance is relatively high, and the average number of features can vary in different test subsets. In LDA, ANOVA feature selection explores a larger number of features (250) to obtain the optimal k, resulting in a higher average number of features. LDA is computationally efficient compared to other classifiers, but RF, XGB, and NNs also perform well, albeit with slightly lower results. However, these classifiers are slower than RF and require significant computational resources to explore a broader range of hyperparameters in random search or increase the number of features in ANOVA feature selection.

4.2 Binary Classifications

In the second scenario, we concentrate on binary classifications derived from four cancer subtypes. Our results show a significant improvement in binary classifications, enabling us to accurately predict the molecular subtype of breast cancer with a high macro-average F1 score. We explore three distinct approaches:

- **One Versus the Rest (OvR)**
- **One Versus One (OvO)**
- **Two Versus Two (TvT)**

4.2.1 One Versus the Rest (OvR)

OvR is a binary classification approach that is derived from multi-class classification [78]. In OvR, a multi-class classification problem is transformed into multiple binary classification problems, where each class is treated as a separate binary classification problem. For each binary classification problem, one class is treated as the positive class (1), and all other classes are treated as the negative class (0). In this particular case, the dataset was discovered to be imbalanced, with a ratio of 50 to 150 between the two labels. To rectify this issue, an effective technique known as SMOTE (Synthetic Minority Over-sampling Technique) was utilized to balance the two labels. However, using this method did not yield improvements in the results, prompting us to revert to using the original data.

	Our Study		**Saha et al.**
Classifier	**Avg No. of Features**	**Avg F1 Score**	**Avg F1 Score**
SVM	avg=61.6 90%CI=(1,131.13)	avg=75.04% 90%CI=(0.70,0.81)	avg=64.87% 90%CI=(0.62,0.67)
LR	avg=12.66 90%CI=(2.93,22.40)	avg=70.84% 90%CI=(0.65,0.77)	avg=64.94% 90%CI=(0.61,0.68)
RF	avg=49.40 90%CI=(2.81,95.99)	avg=77.89% 90%CI=(0.74,0.82)	avg=67.38% 90%CI=(0.64,0.70)

Table 4.2: OvR, Luminal A vs. the Rest

	Our Study		Saha et al.
Classifier	**Avg No. of Features**	**Avg F1 Score**	**Avg F1 Score**
SVM	avg=17.8 90%CI=(3.07,32.53)	avg=70.00% 90%CI=(0.64,0.77)	avg=60.97% 90%CI=(0.57,0.65)
LR	avg=68.4 90%CI=(28.09,108.71)	avg=54.25% 90%CI=(0.47,0.62)	avg=58.25% 90%CI=(0.54,0.62)
RF	avg=76.80 90%CI=(29.81,123.79)	avg=65.57% 90%CI=(0.60,0.71)	avg=58.21% 90%CI=(0.56,0.60)

Table 4.3: OvR, Luminal B vs. the Rest

	Our Study		Saha et al.
Classifier	**Avg No. of Features**	**Avg F1 Score**	**Avg F1 Score**
SVM	avg=65.6 90%CI=(13.98,117.22)	avg=72.99% 90%CI=(0.57,0.90)	avg=67.40% 90%CI=(0.60,0.74)
LR	avg=20.60 90%CI=(1,42.59)	avg=69.61% 90%CI=(0.54,0.87)	avg=62.38% 90%CI=(0.59,0.65)
RF	avg=96.00 90%CI=(39.96,152.04)	avg=77.11% 90%CI=(0.67,0.87)	avg=61.72% 90%CI=(0.60,0.63)

Table 4.4: OvR, HER2+ vs. the Rest

	Our Study		Saha et al.
Classifier	**Avg No. of Features**	**Avg F1 Score**	**Avg F1 Score**
SVM	avg=74.40 90%CI=(19.85,128.96)	avg=64.78% 90%CI=(0.63,0.67)	avg=67.06% 90%CI=(0.63,0.71)
LR	avg=80.00 90%CI=(20.37,139.64)	avg=65.17% 90%CI=(0.62,0.69)	avg=64.18% 90%CI=(0.60,0.68)
RF	avg=76.20 90%CI=(25.26,127.14)	avg=71.19% 90%CI=(0.63,0.80)	avg=65.78% 90%CI=(0.63,0.68)

Table 4.5: OvR, TN vs. the Rest

In one versus the rest classifications, the random forest classifiers performed the best in all cases, except for the Luminal B vs. the rest case, where SVM had a higher F1 score than random forest. In all four cases, the best model using the generated features in our study outperformed the features generated by Saha et al. In the first three cases of Luminal A vs. the rest, Luminal B vs. the rest, and HER2+ vs. the rest, our methodology using extracted features achieved around 10% higher F1 scores compared to the extracted features by Saha et al. In the last case of TN vs. the rest, the difference was approximately 5%. The improvement was attributed to the utilization of a wider range of extracted features and diverse cropping techniques. As recall, precision, and accuracy were very similar, if not identical, to the F1 score, we focused on the F1 score, which is a harmonic average between precision and recall and places more emphasis on the smaller value. The number of selected features varied depending on the case and classifier, as well as the test subset. The confidence interval obtained for the average number of features provided an understanding of the minimum and maximum number of features used for classification. Below are the box plots illustrating the performance of the best model in each OvR binary classification.

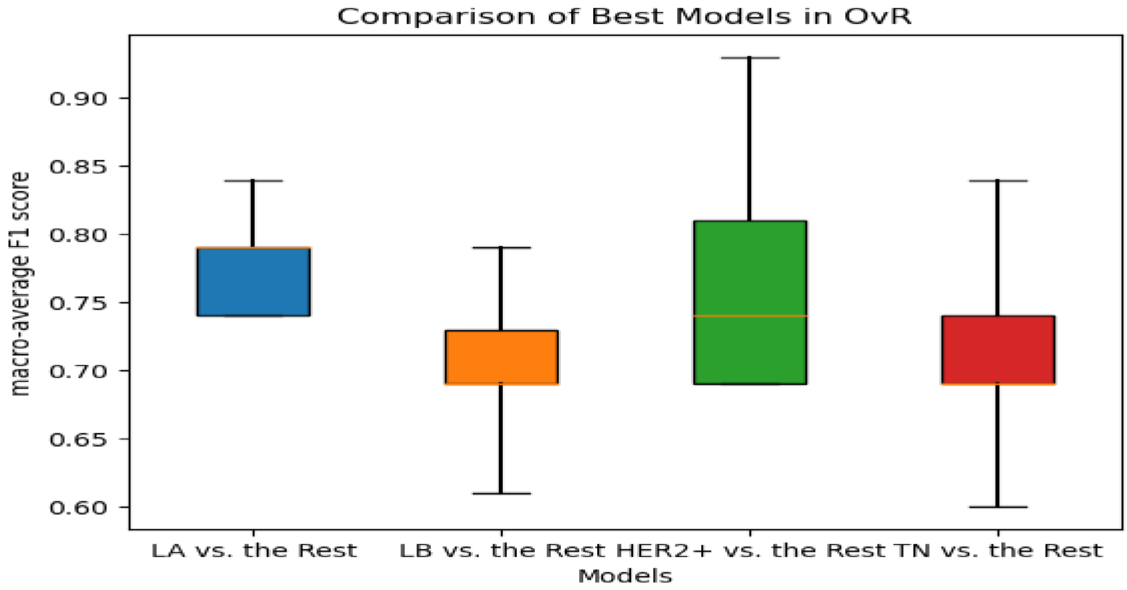

Figure 4.1: The Random Forest (RF) model outperforms all other classifiers in every category, except for the Luminal B vs. the non-Luminal B classification where the Support Vector Machine (SVM) model showed better performance in terms of F1 score.

According to Figure 4.1, the model shows a low variance among test subsets for Luminal A versus the rest, with a minimum F1 score of approximately 75%, indicating robustness. However, for other scenarios with higher variances, incorporating clinical information may enhance the accuracy of predicting the molecular subtypes of breast cancer.

4.2.2 One Versus One (OvO)

OvO is another binary classification approach that is derived from multi-class classification. In OvO, all possible pairs of classes are considered and a binary classifier is trained for each pair. In our study, with four different labels (0, 1, 2, and 3), six binary classifiers were trained: 0 vs. 1, 0 vs. 2, 0 vs. 3, 1 vs. 2, 1 vs. 3, and 2 vs. 3.

During inference, each of the binary classifiers makes a prediction, and the class with the most votes is chosen as the final prediction. OvO is typically used when the number of classes is relatively small, as the number of binary classifiers trained is proportional to the square of the number of classes, which can become computationally expensive for a large number of classes.

Compared to OvR, OvO requires more models to be trained, but each model is trained on a smaller and balanced subset of the data. OvO can also be more robust to imbalanced class distributions, as each binary classifier only needs to distinguish between two classes. However, in this particular case, only half of the available samples were used (i.e., 100 samples out of a total of 200), and the results obtained were quite high, up to 85% F1 score.

Luminal A vs. Luminal B		
Classifier	**Avg No. of Features**	**Avg F1 Score**
SVM	avg=56.6 90%CI=(7.16,106.03)	avg=85.81% 90%CI=(0.77,0.95)
LR	avg=72.00 90%CI=(27.81,116.17)	avg=77.44% 90%CI=(0.63,0.92)
RF	avg=17.40 90%CI=(2.02,32.78)	avg=83.72% 90%CI=(0.75,0.93)

Table 4.6: OvO, Luminal A vs. Luminal B

Luminal A vs. HER2+		
Classifier	**Avg No. of Features**	**Avg F1 Score**
SVM	avg=11.00 90%CI=(1,23.04)	avg=79.68% 90%CI=(0.70,0.89)
LR	avg=46.60 90%CI=(1,95.85)	avg=81.74% 90%CI=(0.74,0.90)
RF	avg=31.20 90%CI=(1,70.20)	avg=85.90% 90%CI=(0.75,0.97)

Table 4.7: OvO, Luminal A vs. HER2

Luminal A vs. TN		
Classifier	**Avg No. of Features**	**Avg F1 Score**
SVM	avg=44.20 90%CI=(1,92.18)	avg=67.38% 90%CI=(0.59,0.76)
LR	avg=53.60 90%CI=(9.86,97.33)	avg=63.71% 90%CI=(0.50,0.78)
RF	avg=69.00 90%CI=(29.95,108.05)	avg=77.25% 90%CI=(0.66,0.88)

Table 4.8: OvO, Luminal A vs. TN

Luminal B vs. HER2+		
Classifier	**Avg No. of Features**	**Avg F1 Score**
SVM	avg=19.60 90%CI=(1,48.86)	avg=77.44% 90%CI=(0.63,0.92)
LR	avg82.20 90%CI=(12.35,152.04)	avg=71.71% 90%CI=(0.55,0.89)
RF	avg=23.40 90%CI=(1,61.11)	avg=79.88% 90%CI=(0.68,0.92)

Table 4.9: OvO, Luminal B vs. HER2+

Luminal B vs. TN		
Classifier	**Avg No. of Features**	**Avg F1 Score**
SVM	avg=42.80 90%CI=(6.23,79.37)	avg=75.80% 90%CI=(0.67,0.84)
LR	avg=53.20 90%CI=(19.54,86.86)	avg=73.16% 90%CI=(0.62,0.84)
RF	avg=26.80 90%CI=(1,55.78)	avg=77.86% 90%CI=(0.70,0.86)

Table 4.10: OvO, Luminal B vs. TN

HER2+ vs. TN		
Classifier	**Avg No. of Features**	**Avg F1 Score**
SVM	avg=28.40 90%CI=(1,64.49)	avg=79.39% 90%CI=(0.72,0.87)
LR	avg=53.40 90%CI=(3.70,103.10)	avg=75.67% 90%CI=(0.66,0.85)
RF	avg=48.60 90%CI=(17.28,79.92)	avg=79.96% 90%CI=(0.68,0.92)

Table 4.11: OvO, HER2+ vs. TN

In one versus one classifications, the random forest model outperformed other models in all cases except for Luminal A versus Luminal B, where SVM achieved a higher F1 score. Across all cases, random forest required fewer features on average to predict molecular subtypes of breast cancer compared to SVM and LR. Below are the box plots illustrating the performance of the best model in each OvO binary classification.

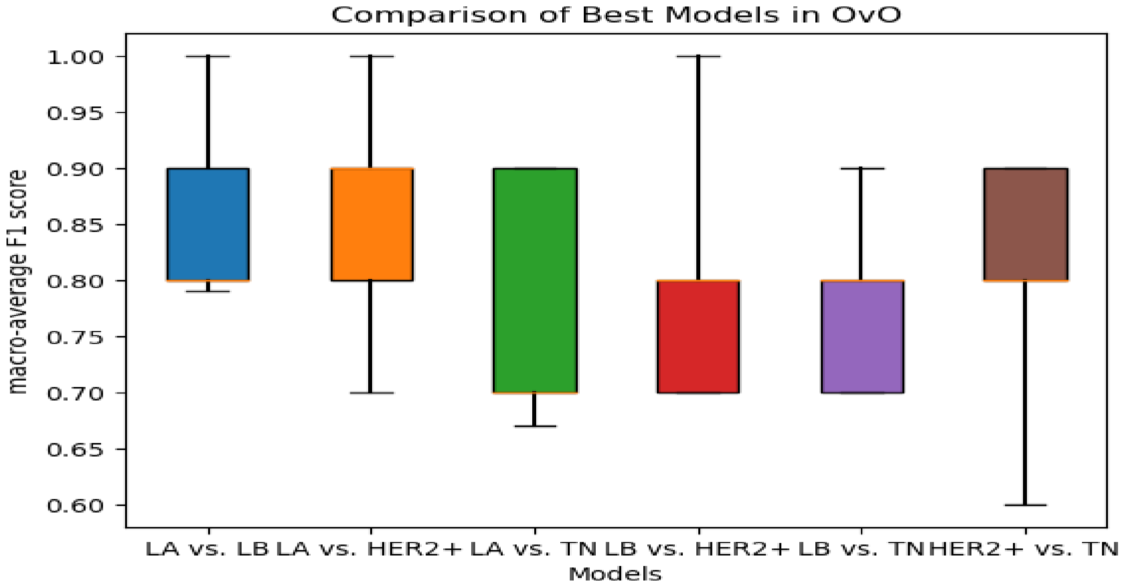

Figure 4.2: In all categories except for Luminal A vs. Luminal B, the Random Forest (RF) model outperforms all other classifiers in terms of F1 score. However, for the Luminal A vs. Luminal B classification, the Support Vector Machine (SVM) model showed better performance.

Based on the results presented in Figure 4.2, the best model shows low variance among test subsets for most cases, particularly for Luminal A versus Luminal B, with a minimum F1 score of approximately 80%, indicating robustness. The best models also perform well in Luminal A versus HER2+ and Luminal A versus TN cases. However, for HER2+ versus TN, the variance among different test subsets is relatively higher, which may make this scenario less robust compared to others.

4.2.3 Two Versus Two (TvT)

TvT is another binary classification approach that is similar to OvO, but instead of considering all possible pairs of classes, it considers combinations of two labels against combinations of two other labels. For example, in our study with four different labels (0, 1, 2, and 3), three binary classifiers would be trained: 01 vs. 23, 02 vs. 13, and 03 vs. 12. This approach can reduce the number of binary classifiers that need to be trained compared to OvO, while still providing a more robust approach than OvR. TvT can also be more efficient than OvO, as a smaller number of binary classifiers need to be trained and evaluated during inference.

Similar to OvO, the dataset in this case is perfectly balanced, thereby eliminating the need to explore any resampling methods.

Luminal A,Luminal B vs. HER2+, TN		
Classifier	**Avg No. of Features**	**Avg F1 Score**
SVM	avg=65.00 90%CI=(19.09,110.91)	avg=73.39% 90%CI=(0.64,0.83)
LR	avg=53.40 90%CI=(14.76,92.04)	avg=73.58% 90%CI=(0.64,0.83)
RF	avg=65.60 90%CI=(30.54,100.66)	avg=75.54% 90%CI=(0.69,0.82)

Table 4.12: TvT, Luminal A, Luminal B vs. HER2+, TN

Luminal A, HER2+ vs. Luminal B, TN		
Classifier	**Avg No. of Features**	**Avg F1 Score**
SVM	avg=19.00 90%CI=(5.77,32.23)	avg=74.80% 90%CI=(0.71,0.79)
LR	avg=28.60 90%CI=(1,60.91)	avg=69.87% 90%CI=(0.64,0.76)
RF	avg=25.00 90%CI=(11.30,38.70)	avg=74.97% 90%CI=(0.69,0.81)

Table 4.13: TvT, Luminal A, HER2+ vs. Luminal B, TN

Luminal A, TN vs. Luminal B, HER2+		
Classifier	**Avg No. of Features**	**Avg F1 Score**
SVM	avg=35.60 90%CI=(7.95,63.25)	avg=78.95% 90%CI=(0.70,0.88)
LR	avg=39.20 90%CI=(14.54,63.86)	avg=76.91% 90%CI=(0.68,0.86)
RF	avg=32.40 90%CI=(10.48,54.32)	avg=75.94% 90%CI=(0.68,0.84)

Table 4.14: TvT, Luminal A, TN vs. Luminal B, HER2+

4.3 3-Label Classifications

In the last scenario, we focus on classifications with three labels derived from four types of cancers. We consider two different approaches:

4.3.1 3-Label with Elimination of One Label

For the first approach, we drop one label in four stages and classify the remaining three labels. In this case, we are using 150 samples out of 200.

Luminal A, Luminal B, HER2+		
Classifier	**Avg No. of Features**	**Avg F1 Score**
SVM	avg=66.60 90%CI=(13.26,119.94)	avg=63.70% 90%CI=(0.56,0.72)
LR	avg=105.00 90%CI=(59.16,150.84)	avg=65.09% 90%CI=(0.54,0.76)
RF	avg=81.40 90%CI=(34.10,128.70)	avg=63.65% 90%CI=(0.56,0.72)

Table 4.15: 3-Label, Luminal A, Luminal B, HER2+

Luminal A, Luminal B, TN		
Classifier	**Avg No. of Features**	**Avg F1 Score**
SVM	avg=73.80 90%CI=(35.91,111.69)	avg=58.10% 90%CI=(0.49,0.68)
LR	avg=63.20 90%CI=(24.62,101.78)	avg=48.45% 90%CI=(0.37,0.61)
RF	avg=90.80 90%CI=(60.89,120.71)	avg=55.24% 90%CI=(0.46,0.65)

Table 4.16: 3-Label, Luminal A, Luminal B, TN

Luminal A, HER2+, TN		
Classifier	**Avg No. of Features**	**Avg F1 Score**
SVM	avg=45.00 90%CI=(1,97.75)	avg=61.80% 90%CI=(0.49,0.75)
LR	avg=50.20 90%CI=(9.72,90.68)	avg=60.73% 90%CI=(0.49,0.73)
RF	avg=44.20 90%CI=(1,101.78)	avg=63.92% 90%CI=(0.57,0.71)

Table 4.17: 3-Label, Luminal A, HER2+, TN

Luminal B, HER2+, TN		
Classifier	**Avg No. of Features**	**Avg F1 Score**
SVM	avg=40.60 90%CI=(15.80,65.40)	avg=56.18% 90%CI=(0.50,0.63)
LR	avg=23.00 90%CI=(1,47.33)	avg=50.39% 90%CI=(0.47,0.54)
RF	avg=43.00 90%CI=(1,89.32)	avg=60.57% 90%CI=(0.55,0.66)

Table 4.18: 3-Label, Luminal B, HER2+, TN

In this case, the models can predict three molecular subtypes of breast cancer simultaneously with a reasonably good F1 score. Although the scores for binary classifications are slightly higher, the models perform well compared to the first scenario where all four subtypes were available. Binary or 3-label classifications can also be useful in situations where clinical features and other medical resources indicate a very low possibility of being infected with one or two cancer types, allowing us to use these models for more accurate predictions.

4.3.2 3-Label with Combining One Label

For the second approach, two labels are combined to create six 3-label classifications. However, in this study, we only consider the case where Luminal A and Luminal B are merged to form a single label called Luminal-like, and this label is compared against two other labels, HER2+ and TN. This is a common comparison made in the literature.

Luminal Like, HER2+, TN		
Classifier	**Avg No. of Features**	**Avg F1 Score**
SVM	avg=48.40 90%CI=(7.72,89.08)	avg=57.08% 90%CI=(0.51,0.64)
LR	avg=48.80 90%CI=(22.14,75.46)	avg=53.50% 90%CI=(0.46,0.61)
RF	avg=88.80 90%CI=(38.70,138.89)	avg=62.00% 90%CI=(0.55,0.69)

Table 4.19: 3-Label, Luminal Like, HER2+, TN

Similarly to previous cases, the random forest algorithm performs the best in distinguishing between Luminal Like, HER2+, and TN subtypes. However, there is a significant variance in the number of features among different test subsets when compared to SVM and LR algorithms.

Chapter 5

Conclusion

This chapter summarizes the principal findings that pertain to the research objectives and assesses their importance and contribution. Furthermore, the limitations of the study are discussed, and future research directions are suggested.

5.1 Principal Findings

1. The feature classes that are most useful for predicting molecular subtypes of breast cancer using DCE MRI are FOS, GLCM, GLDS, GLRLM, GLSZM, and DWT.

2. Among the 23 clinical features, we found that age, menopause at diagnosis, race and ethnicity, and lymphadenopathy or suspicious nodes were the most significant based on p-*value* obtained from ANOVA feature selection. These p-*values* ranged from 10^{-6} to 10^{-2}, depending on the classification category.

3. The RF classifier and ANOVA feature selection performed the best in most cases. LOF and KNN imputation are helpful for processing the dataset, provided that their hyperparameters are fine-tuned. However, resampling methods did not significantly improve performance, even in cases of imbalanced datasets.

4. In OvR classifications, the methodology consistently predicts Luminal A versus the other subtypes, achieving a minimum F1 score of 75% (Figure 4.1). In OvO classifications, the methodology performs well in distinguishing between Luminal A and Luminal B, with a significant F1 score of 80% up to 100% (Figure 4.2). This distinction is crucial as Luminal A and Luminal B subtypes share some characteristics, such as positive ER and/or PR expression, but can differ in terms of HER2 expression and proliferation index.

5. Our methodology has broad applicability across a range of medical images, including MRI, CT, ultrasound, and mammography, and can be used to distinguish between two or more disease classes.

5.2 Study Limitations

There are several limitations to this study. Firstly, the study was resource-intensive, with some sections taking several days to complete. As a result, exploring some ideas in more depth was prohibitively expensive. For example, XGB and NNs showed promise, but running them with a high number of iterations in ANOVA feature selection or considering a larger number of iterations in random search was practically impossible, resulting in lower-than-expected results. Additionally, due to time constraints, we were unable to explore a wide range of thresholds in the thresholding technique used to generate binary masks. We used a threshold of T=50, but using other thresholds would have required repeating the entire processes of feature extraction, feature selection, and hyperparameter tuning for classification, making it impossible to explore a wider range. Another limitation of this study is the inherent complexity and challenge of medical images, even for specialists. Therefore, the contribution of radiologists to advise on medical or biological areas would be useful to enhance the accuracy and reliability of the methodology.

5.3 Future Work

Future research in this field could focus on exploring additional medical imaging modalities, such as CT or ultrasound, and even combining multiple medical images to extract features and improve prediction accuracy. Experimenting with various thresholds for generating masks and exploring additional feature classes while tuning corresponding hyperparameters to extract the most informative features would also be beneficial. Using larger datasets with high-quality images that contain all sequences and cancer types would help to increase the generalizability of the methodology. Additionally, creating additional datasets by cropping images in different ways could provide further insights. However, these processes are computationally expensive, and utilizing a powerful server with high GPU and CPU capabilities would be necessary to enable deeper research.

Bibliography

[1] WHO. Breast cancer: prevention and control. *World Health Organization*, 2020.

[2] Aleix Prat, Estela Pineda, Barbara Adamo, Patricia Galván, Aranzazu Fernández, Lydia Gaba, Marc Díez, Margarita Viladot, Ana Arance, and Montserrat Muñoz. Molecular subtypes of breast cancer: a review. *Biological Research*, 47(1):1–14, 2014.

[3] Ashirbani Saha, Michael R Harowicz, Lars J Grimm, Connie E Kim, Sujata V Ghate, Ruth Walsh, and Maciej A Mazurowski. A machine learning approach to radiogenomics of breast cancer: a study of 922 subjects and 529 DCE-MRI features. *British Journal of Cancer*, 119(4):508–516, 2018.

[4] Britta Weigelt, Frederick L Baehner, and Jorge S Reis-Filho. Challenges in breast cancer pathology and molecular diagnostics. *Nature Reviews Cancer*, 10(11):727–736, 2010.

[5] Ming Fan, Hui Li, Shijian Wang, Bin Zheng, Juan Zhang, and Lihua Li. Radiomic analysis reveals DCE-MRI features for prediction of molecular subtypes of breast

cancer. *Plos One*, 12(2):e0171683, 2017.

[6] Lars J Grimm, Jing Zhang, and Maciej A Mazurowski. Computational approach to radiogenomics of breast cancer: luminal A and luminal B molecular subtypes are associated with imaging features on routine breast mri extracted using computer vision algorithms. *Journal of Magnetic Resonance Imaging*, 42(4):902–907, 2015.

[7] Hui Li, Yitan Zhu, Elizabeth S Burnside, Erich Huang, Karen Drukker, Katherine A Hoadley, Cheng Fan, Suzanne D Conzen, Margarita Zuley, Jose M Net, et al. Quantitative MRI radiomics in the prediction of molecular classifications of breast cancer subtypes in the tcga/tcia data set. *NPJ Breast Cancer*, 2:16012, 2016.

[8] Veli S Öztürk, Yasemin D Polat, Aykut Soyder, Ahmet Tanyeri, Can Z Karaman, and Füsun Taşkın. The relationship between mri findings and molecular subtypes in women with breast cancer. *Current Problems in Diagnostic Radiology*, 49(6):417–421, 2020.

[9] Yuhong Huang, Lihong Wei, Yalan Hu, Nan Shao, Yingyu Lin, Shaofu He, Huijuan Shi, Xiaoling Zhang, and Ying Lin. Multi-parametric mri-based radiomics models for predicting molecular subtype and androgen receptor expression in breast cancer. *Frontiers in Oncology*, 11:706733, 2021.

[10] Francesca Galati, Veronica Rizzo, Giuliana Moffa, Claudia Caramanico, Endi Kripa, Bruna Cerbelli, Giulia D'Amati, and Federica Pediconi. Radiologic-pathologic correlation in breast cancer: Do mri biomarkers correlate with pathologic features and molecular subtypes? *European Radiology Experimental*, 6(1):39, 2022.

[11] Yang Zhang, Jeon-Hor Chen, Yezhi Lin, Siwa Chan, Jiejie Zhou, Daniel Chow, Peter Chang, Tiffany Kwong, Dah-Cherng Yeh, Xinxin Wang, et al. Prediction of breast cancer molecular subtypes on dce-mri using convolutional neural network with transfer learning between two centers. *European Radiology*, 31:2559–2567, 2021.

[12] Rong Sun, Zijun Meng, Xuewen Hou, Yang Chen, Yifeng Yang, Gang Huang, and Shengdong Nie. Prediction of breast cancer molecular subtypes using dce-mri based on cnns combined with ensemble learning. *Physics in Medicine & Biology*, 66(17):175009, 2021.

[13] Lindsay W Turnbull. Dynamic contrast-enhanced mri in the diagnosis and management of breast cancer. *NMR in Biomedicine: An International Journal Devoted to the Development and Application of Magnetic Resonance In Vivo*, 22(1):28–39, 2009.

[14] M. Poghosyan. Pyfeats: A python library for feature extraction. https://pypi.org/project/pyfeats/, 2019.

[15] A Rogier T Donders, Geert JMG Van Der Heijden, Theo Stijnen, and Karel GM Moons. A gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology*, 59(10):1087–1091, 2006.

[16] Shichao Zhang. Nearest neighbor selection for iteratively knn imputation. *Journal of Systems and Software*, 85(11):2541–2552, 2012.

[17] Anil Jadhav, Dhanya Pramod, and Krishnan Ramanathan. Comparison of performance of data imputation methods for numeric dataset. *Applied Artificial Intelligence*, 33(10):913–933, 2019.

[18] Manoranjan Dash and Huan Liu. Feature selection for classification. *Intelligent Data Analysis*, 1(1-4):131–156, 1997.

[19] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(2), 2012.

[20] Charles M Perou, Therese Sørlie, Michael B Eisen, Matt van de Rijn, Stefanie S Jeffrey, Christian A Rees, Jonathan R Pollack, Douglas T Ross, Hilde Johnsen, Lars A Akslen, and et al. Breast cancer subtypes: molecular characterization, clinical management, and future perspectives. *Clinical Advances in Hematology & Oncology: H&O*, 1(6):415–423, 2003.

[21] Fahmi Khalifa, Ahmed Soliman, Ayman El-Baz, Mohamed Abou El-Ghar, Tarek El-Diasty, Georgy Gimel'farb, Rosemary Ouseph, and Amy C Dwyer. Models

and methods for analyzing dce-mri: A review. *Medical Physics*, 41(12):124301, 2014.

[22] Elizabeth J Sutton, Brittany Z Dashevsky, Jung Hun Oh, Harini Veeraraghavan, Aditya P Apte, Sunitha B Thakur, Elizabeth A Morris, and Joseph O Deasy. Breast cancer molecular subtype classifier that incorporates mri features. *Journal of Magnetic Resonance Imaging*, 44(1):122–129, 2016.

[23] Jinwoo Son, Si Eun Lee, Eun-Kyung Kim, and Sungwon Kim. Prediction of breast cancer molecular subtypes using radiomics signatures of synthetic mammography from digital breast tomosynthesis. *Scientific Reports*, 10(1):21566, 2020.

[24] A Saha, MR Harowicz, LJ Grimm, et al. Dynamic contrast-enhanced magnetic resonance images of breast cancer patients with tumor locations. *The Cancer Imaging Archive*, 2021.

[25] Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, et al. The cancer imaging archive (tcia): maintaining and operating a public information repository. *Journal of Digital Imaging*, 26:1045–1057, 2013.

[26] Therese Sørlie, Charles M Perou, Robert Tibshirani, Turid Aas, Stephanie Geisler, Hilde Johnsen, Trevor Hastie, Michael B Eisen, Matt Van De Rijn, Stefanie S Jeffrey, et al. Gene expression patterns of breast carcinomas distin-

guish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences*, 98(19):10869–10874, 2001.

[27] Maggie CU Cheang, Stephen K Chia, David Voduc, Dongxia Gao, Samuel Leung, Jacqueline Snider, Mark Watson, Sherri Davies, Philip S Bernard, Joel S Parker, et al. Ki67 index, her2 status, and prognosis of patients with luminal b breast cancer. *JNCI: Journal of the National Cancer Institute*, 101(10):736–750, 2009.

[28] Ryan H Engel and Virginia G Kaklamani. Her2-positive breast cancer: current and future treatment strategies. *Drugs*, 67:1329–1341, 2007.

[29] Rebecca Dent, Maureen Trudeau, Kathleen I Pritchard, Wedad M Hanna, Harriet K Kahn, Carol A Sawka, Lavina A Lickley, Ellen Rawlinson, Ping Sun, and Steven A Narod. Triple-negative breast cancer: clinical features and patterns of recurrence. *Clinical Cancer Research*, 13(15):4429–4434, 2007.

[30] Jagadeesan Jayender, Eva Gombos, Sona Chikarmane, Donnette Dabydeen, Ferenc A Jolesz, and Kirby G Vosburgh. Statistical learning algorithm for in situ and invasive breast carcinoma segmentation. *Computerized Medical Imaging and Graphics*, 37(4):281–292, 2013.

[31] Jennifer Xiao, Habib Rahbar, Daniel S Hippe, Mara H Rendi, Elizabeth U Parker, Neal Shekar, Michael Hirano, Kevin J Cheung, and Savannah C Partridge. Dynamic contrast-enhanced breast mri features correlate with invasive breast cancer angiogenesis. *NPJ Breast Cancer*, 7(1):42, 2021.

[32] Darryl McClymont, Andrew Mehnert, Adnan Trakic, Dominic Kennedy, and Stuart Crozier. Fully automatic lesion segmentation in breast mri using mean-shift and graph-cuts on a region adjacency graph. *Journal of Magnetic Resonance Imaging*, 39(4):795–804, 2014.

[33] Andriy Fedorov, Reinhard Beichel, Jayashree Kalpathy-Cramer, Julien Finet, Jean-Christophe Fillion-Robin, Sonia Pujol, Christian Bauer, Dominique Jennings, Fiona Fennessy, Milan Sonka, et al. 3d slicer as an image computing platform for the quantitative imaging network. *Magnetic Resonance Imaging*, 30(9):1323–1341, 2012.

[34] W Dean Bidgood Jr, Steven C Horii, Fred W Prior, and Donald E Van Syckle. Understanding and using dicom, the data interchange standard for biomedical imaging. *Journal of the American Medical Informatics Association*, 4(3):199–212, 1997.

[35] Dipak Kumar Patra, Tapas Si, Sukumar Mondal, and Prakash Mukherjee. Breast dce-mri segmentation for lesion detection by multi-level thresholding using student psychological based optimization. *Biomedical Signal Processing and Control*, 69:102925, 2021.

[36] Tapas Si, Dipak Kumar Patra, Sukumar Mondal, and Prakash Mukherjee. Breast dce-mri segmentation for lesion detection using chimp optimization algorithm. *Expert Systems with Applications*, 204:117481, 2022.

[37] Aida Fooladivanda, Shahriar B Shokouhi, Nasrin Ahmadinejad, and Mohammad R Mosavi. Automatic segmentation of breast and fibroglandular tissue in breast mri using local adaptive thresholding. In *2014 21th Iranian Conference on Biomedical Engineering (ICBME)*, pages 195–200. IEEE, 2014.

[38] Wei Li, Kun Yu, Chaolu Feng, Dazhe Zhao, et al. Molecular subtypes recognition of breast cancer in dynamic contrast-enhanced breast magnetic resonance imaging phenotypes from radiomics data. *Computational and Mathematical Methods in Medicine*, 2019, 2019.

[39] Gábor J Székely and Maria L Rizzo. Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272, 2013.

[40] Maria Dolores Esteban and Domingo Morales. A summary on entropy statistics. *Kybernetika*, 31(4):337–346, 1995.

[41] P Mohanaiah, P Sathyanarayana, and L GuruKumar. Image texture feature extraction using glcm approach. *International Journal of Scientific and Research Publications*, 3(5):1–5, 2013.

[42] Ayodeji Olalekan Salau and Shruti Jain. Feature extraction: a survey of the types, techniques, applications. In *2019 International Conference on Signal Processing and Communication (ICSC)*, pages 158–164. IEEE, 2019.

[43] Isabella Fornacon-Wood, Hitesh Mistry, Christoph J Ackermann, Fiona Blackhall, Andrew McPartlin, Corinne Faivre-Finn, Gareth J Price, and James PB O'Connor. Reliability and prognostic value of radiomic features are highly dependent on choice of feature extraction platform. *European Radiology*, 30:6241–6250, 2020.

[44] Chung-Ming Wu and Yung-Chang Chen. Statistical feature matrix for texture analysis. *CVGIP: Graphical Models and Image Processing*, 54(5):407–419, 1992.

[45] Arden Sagiterry Setiawan, Julian Wesley, Yudy Purnama, et al. Mammogram classification using law's texture energy measure and neural networks. *Procedia Computer Science*, 59:92–97, 2015.

[46] Francesca Biagini, Yaozhong Hu, Bernt Øksendal, and Tusheng Zhang. *Stochastic calculus for fractional Brownian motion and applications*. Springer Science & Business Media, 2008.

[47] Manoj Kumar Biswas, Tirthankar Ghose, Sudipta Guha, and Prabir Kumar Biswas. Fractal dimension estimation for texture images: a parallel approach. *Pattern Recognition Letters*, 19(3-4):309–313, 1998.

[48] Fritz Albregtsen and Birgitte Nielsen. Texture classification based on cooccurrence of gray level run length matrices. *Australian Journal of Intelligent Information Processing Systems*, 6(1):38–45, 2000.

[49] Takashi Matsuyama, Shu-Ichi Miura, and Makoto Nagao. Structural analysis of natural textures by fourier transformation. *Computer Vision, Graphics, and Image Processing*, 24(3):347–362, 1983.

[50] Yang Mingqiang, Kpalma Kidiyo, Ronsin Joseph, et al. A survey of shape feature extraction techniques. *Pattern Recognition*, 15(7):43–90, 2008.

[51] Niels W Schurink, Simon R van Kranen, Sander Roberti, Joost JM van Griethuysen, Nino Bogveradze, Francesca Castagnoli, Najim el Khababi, Frans CH Bakers, Shira H de Bie, Gerlof PT Bosma, et al. Sources of variation in multicenter rectal mri data and their effect on radiomics feature reproducibility. *European Radiology*, pages 1–11, 2022.

[52] Yuan Shao and Mehmet Celenk. Higher-order spectra (hos) invariants for shape recognition. *Pattern Recognition*, 34(11):2097–2113, 2001.

[53] Peter Toft. The radon transform. *Theory and Implementation (Ph. D. Dissertation)(Copenhagen: Technical University of Denmark)*, 1996.

[54] Esa Prakasa. Texture feature extraction by using local binary pattern. *INKOM Journal*, 9(2):45–48, 2016.

[55] Yidong Chen and Edward R Dougherty. Gray-scale morphological granulometric texture classification. *Optical Engineering*, 33(8):2713–2722, 1994.

[56] Frank Yeong-Chyang Shih and Owen Robert Mitchell. Decomposition of gray-scale morphological structuring elements. *Pattern Recognition*, 24(3):195–203, 1991.

[57] Conrad Sanderson and Brian C Lovell. Multi-region probabilistic histograms for robust and scalable identity inference. In *Advances in Biometrics: Third International Conference, ICB 2009, Alghero, Italy, June 2-5, 2009. Proceedings 3*, pages 199–208. Springer, 2009.

[58] Victor Murray, Marios S Pattichis, Eduardo Simon Barriga, and Peter Soliz. Recent multiscale am-fm methods in emerging applications in medical imaging. *EURASIP Journal on Advances in Signal Processing*, 2012(1):1–14, 2012.

[59] Kamarul Hawari Ghazali, Mohd Fais Mansor, Mohd Marzuki Mustafa, and Aini Hussain. Feature extraction technique using discrete wavelet transform for image classification. In *2007 5th Student Conference on Research and Development*, pages 1–4. IEEE, 2007.

[60] Nhat-Duc Hoang. Image processing-based spall object detection using gabor filter, texture analysis, and adaptive moment estimation (adam) optimized logistic regression models. *Advances in Civil Engineering*, 2020:1–16, 2020.

[61] Alireza Khotanzad and Yaw Hua Hong. Invariant image recognition by zernike moments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(5):489–497, 1990.

[62] P Bhaskara Rao, D Vara Prasad, and Ch Pavan Kumar. Feature extraction using zernike moments. *International Journal of Latest Trends in Engineering and Technology*, 2(2):228–234, 2013.

[63] Bee Wah Yap and Chiaw Hock Sim. Comparisons of various types of normality tests. *Journal of Statistical Computation and Simulation*, 81(12):2141–2155, 2011.

[64] José M Jerez, Ignacio Molina, Pedro J García-Laencina, Emilio Alba, Nuria Ribelles, Miguel Martín, and Leonardo Franco. Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artificial Intelligence in Medicine*, 50(2):105–115, 2010.

[65] Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. Lof: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, pages 93–104, 2000.

[66] Jiliang Tang, Salem Alelyani, and Huan Liu. Feature selection for classification: A review. *Data classification: Algorithms and Applications*, page 37, 2014.

[67] Mark A Hall. *Correlation-based feature selection for machine learning.* PhD thesis, The University of Waikato, 1999.

[68] Mark A Hall. Correlation-based feature selection of discrete and numeric class machine learning. 2000.

[69] Majid Afshar and Hamid Usefi. Dimensionality reduction using singular vectors. *Scientific Reports*, 11(1):1–13, 2021.

[70] Majid Afshar and Hamid Usefi. Optimizing feature selection methods by removing irrelevant features using sparse least squares. *Expert Systems with Applications*, 200:116928, 2022.

[71] Dimitrios Ververidis and Constantine Kotropoulos. Sequential forward feature selection with low computational cost. In *2005 13th European Signal Processing Conference*, pages 1–4. IEEE, 2005.

[72] Steve R Gunn et al. Support vector machines for classification and regression. *ISIS Technical Report*, 14(1):5–16, 1998.

[73] Lei Liu. Research on logistic regression algorithm of breast cancer diagnose data by machine learning. In *2018 International Conference on Robots & Intelligent System (ICRIS)*, pages 157–160. IEEE, 2018.

[74] Suresh Balakrishnama and Aravind Ganapathiraju. Linear discriminant analysis-a brief tutorial. *Institute for Signal and Information Processing*, 18(1998):1–8, 1998.

[75] Angshuman Paul, Dipti Prasad Mukherjee, Prasun Das, Abhinandan Gangopadhyay, Appa Rao Chintha, and Saurabh Kundu. Improved random forest for classification. *IEEE Transactions on Image Processing*, 27(8):4012–4024, 2018.

[76] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm Sigkdd International Conference on knowledge Discovery and Data Mining*, pages 785–794, 2016.

[77] Guoqiang Peter Zhang. Neural networks for classification: a survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 30(4):451–462, 2000.

[78] Ryan Rifkin and Aldebaro Klautau. In defense of one-vs-all classification. *The Journal of Machine Learning Research*, 5:101–141, 2004.