IMPROVING THE FEATURE SELECTION, MULTI-CLASS CLASSIFICATION, AND IMBALANCED DATASET OF BREAST CANCER

EMAD ABD AL RAHMAN

UNIVERSITI SAINS MALAYSIA

2025

IMPROVING THE FEATURE SELECTION, MULTI-CLASS CLASSIFICATION, AND IMBALANCED DATASET OF BREAST CANCER

by

EMAD ABD AL RAHMAN

Thesis submitted in fulfilment of the requirements
for the degree of
Doctor of Philosophy

February 2025

ACKNOWLEDGEMENT

First and foremost, I extend my deepest gratitude to Allah SWT for His countless blessings and for endowing me with the strength and perseverance needed to complete this thesis. It is through His will that all endeavors find their beginning and end.

I would like to express my sincere appreciation to my advisors, Dr. Nur Intan Raihana Ruhaiyem, Dr. Majed Bouchahma, and Dr. Kamarul Imran Musa, for their invaluable guidance, unwavering support, and continuous encouragement throughout this process. Their insights and expertise have been instrumental in shaping this research, and their patience and understanding have made this journey both enlightening and enriching.

To my wife and children, your support and encouragement have been my stronghold. The countless nights spent working on this thesis were made bearable by your understanding and love. Your belief in me has been a source of motivation that kept me focused and driven, and I am endlessly grateful for your sacrifices. I am also thankful to my brother and sisters for their continuous encouragement and belief in my abilities. Your unwavering faith in me has been a constant source of strength and inspiration, pushing me to strive for excellence.

Last, but certainly not least, I dedicate this achievement to my beloved parents, who, although not present to witness this moment, have always been a beacon of hope and ambition in my life. I am filled with a mix of emotions—sadness for their absence, pride for achieving what they had always envisioned for me, and prayers for their peace and happiness in a better place. Their memory and legacy have been a guiding light throughout this journey, inspiring me to fulfill the dreams they harbored for me.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF SYMBOLS

$\emptyset_i$	Shapley Value
v	Payoff or impact of a subset of features
m	ML Model
cp	Cutoff percentile
cm	Consistency margin
t	Cutoff threshold
ct	Consistency threshold
$\mathcal{M}$	A 2-D Matrix

LIST OF ABBREVIATIONS

USM	Universiti Sains Malaysia
CDC	Center for Disease Control and Prevention
MRI	Magnetic Resonance Imaging
CT	Computed Tomography
PET	Positron-Emission Tomography
ER	Estrogen Receptor
PR	Progesterone Receptor
ERBB2	Human Epidermal Growth Factor 2
HER2	Human Epidermal Growth Factor Receptor 2
TNM	Tumor, Node, Metastasis
FDA	Food and Drug Administration
BI-RADS	Breast Imaging Reporting and Data System
DCE	Dynamic Contrast-Enhanced
DBT	Digital Breast Tomosynthesis
AWBUS	Automated Whole Breast Ultrasound
AUC	Area Under Curve
PMRT	Post Mastectomy Radiation
MBC	Metastatic Breast Cancer
DDSM	Digital Database for Screening Mammography
USF	University of South Florida
ACR	American College of Radiology
JPEG	Joint Photographic Experts Group
CBIS	Curated Breast Imaging Subset

DICOM	Digital Imaging and Communications in Medicine
ROI	Region of Interest
MIAS	Mammographic Image Analysis Society
PEIPA	Pilot European Image Processing Archive
DAT-DDS	Digital Audio File – Digital Data Storage
WBCD	Wisconsin Breast Cancer Dataset
FNA	Fine Needle Aspirate
CC	Craniocaudal
MLO	Mediolateral Oblique
GT	Ground Truth
SEER	Surveillance, Epidemiology and End Results
NCI	National Cancer Institute
CAD	Computer-Aided Diagnosis
CADe	Computer-Aided Detection
CADx	Computer-Aided Diagnostic
MDT	Multidisciplinary Team
CDS	Clinical Decision Support
EPR	Electronic Patient Record
AI	Artificial Intelligence
ML	Machine Learning
MSK	Memorial Sloan Kettering
WFO	Watson for Oncology
GCNN	Graph Convolutional Neural Networks
MATE	Multidisciplinary Team Assistant and Treatments Elector
ANN	Artificial Neural Networks

SVM	Support Vector Machine
RF	Random Forest
LDA	Linear Discriminant Analysis
LR	Logistic Regression
DT	Decision Trees
DL	Deep Learning
NN	Neural Networks
CNN	Convolutional Neural Network
DREAM	Dialogue for Reverse Engineering Assessments and Methods
ANN	Artificial Neural Network
GLCM	Grey-Level Co-Occurrence Matrix
DCE	Dynamic Contrast Enhanced
BBN	Bayesian Belief Network
ROC	Receiver Operating Characteristic
DWI	Diffusion-Weighted Imaging
SER	Signal Enhancement Ratio
ADC	Apparent Diffusion Coefficient
NB	Naïve Bayes
KNN	K-Nearest Neighbors
EUS	Ensemble Under-Sampling
SCBDL	Soft Clustered Based Direct Learning
MLP	Multi-Layer Perceptron
EX-DBC	Expert System for Diagnosis of Breast Cancer
UCI	University of California, Irvine
PSOWNN	Particle Swarm Optimized Wavelet Neural Network

RBF	Radial Basis Function
BANN	Boosting ANN
RLMPSO	Reinforcement Learning-based Memetic Particle Swarm Optimizer
MC	Micro-Calcification
SAE	Stacked Autoencoder
ResNets	Residual Networks
FFDM	Full Field Digital Mammography
ILSVRC	ImageNet Large Scale Visual Recognition Challenge
FC	Fully Connected
DCNN	Deep Convolutional Neural Network
LWT	Lifting Wavelet Transform
CATP	Computer-Aided Treatment Prediction
PCA	Principal Component Analysis
ELM	Extreme Learning Machine
MFO	Moth Flame Optimization
MC-SVF	Multi-Class Shapley Value Filter
SMOTE	Synthetic Minority Over-Sampling Technique
OVO	One-vs-One
OVA	One-vs-All
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
MC-RUS	Multi-Class Random Under Sampler
MC-BMG	Multi-Class Binary Model Generator

MENINGKATKAN PEMILIHAN CIRI, PENGELASAN BERBILANG KELAS, DAN SET DATA KANSER PAYUDARA YANG KETIDAKSEIMBANGAN

ABSTRAK

Fakta bahawa kanser payudara adalah penyakit yang kerap berlaku dalam kalangan wanita menjadikannya satu isu serius dalam kesihatan global. Pengesanan awal adalah sangat penting untuk meningkatkan pilihan rawatan dan kadar kelangsungan hidup, namun, kerumitan dalam mendiagnosis dan menentukan pelan rawatan yang paling berkesan menghadirkan cabaran besar. Dalam beberapa tahun kebelakangan ini, teknik Kecerdasan Buatan (AI) dan Pembelajaran Mesin (ML) telah muncul sebagai alat yang berkuasa dalam usaha melawan kanser payudara, membuka kemungkinan baru untuk memperbaiki kaedah pengesanan dan rawatan. Kajian ini bertujuan untuk menangani cabaran utama dalam aplikasi AI/ML bagi perancangan rawatan kanser payudara, termasuk dataset yang tidak seimbang, kaedah pemilihan ciri yang kurang optimum, dan kerumitan tugas pengelasan pelbagai kelas. Fokus kajian ini dibuktikan dengan memenuhi keperluan yang belum dipenuhi untuk alat pengiraan yang lebih baik, yang dapat memperibadikan dan mengoptimumkan strategi rawatan untuk pesakit kanser payudara. Kami bertujuan untuk meningkatkan prestasi model pada dataset tidak seimbang dengan memperbaiki prosedur pemilihan ciri, memperhalusi model pengelasan pelbagai kelas, serta membangunkan model ramalan untuk perancangan rawatan peribadi. Pencapaian objektif tesis ini berpotensi menyumbang secara signifikan kepada penciptaan pelan rawatan yang optimum untuk setiap pesakit. Keberkesanan kaedah pemilihan ciri sedia ada dan prestasi pelbagai algoritma ML diuji secara menyeluruh menggunakan 5-lipat cross-validation pada

dataset SEER dan tiroid. Keputusan menunjukkan bahawa XGBoost menunjukkan prestasi yang sangat baik, manakala CatBoost dan Pokok Keputusan terjejas oleh strategi pemilihan ciri mereka. Pengenalan MC-SVF, satu kaedah pemilihan ciri yang dipertingkatkan, meningkatkan prestasi algoritma ini dengan ketara, sekali gus menekankan batasan kaedah pemilihan ciri tradisional dalam menangani dataset tidak seimbang dan senario pelbagai kelas secara efektif. Kajian ini juga menunjukkan penambahbaikan besar model pengelasan pelbagai kelas dalam menguruskan ketidakseimbangan kelas dan mencapai pembezaan kelas, terutamanya untuk kelas minoriti, walaupun terdapat penurunan sedikit dalam ketepatan untuk meningkatkan kadar penarikan semula sehingga 60%. Dari perspektif sains komputer, kajian ini memberikan sumbangan penting kepada sistem sokongan keputusan yang didorong oleh AI. MC-SVF bukan sahaja menangani batasan kaedah pemilihan ciri sedia ada, tetapi juga menetapkan penanda aras baru dalam menangani dataset tidak seimbang dan pengelasan pelbagai kelas. Kajian ini berakhir dengan mencadangkan dua model rawatan ramalan yang mengintegrasikan MC-SVF, membuka jalan untuk algoritma adaptif dan metodologi yang lebih berkesan dalam merangkumi hasil rawatan yang kompleks. Inovasi-inovasi ini berpotensi memajukan AI dalam penjagaan kesihatan sambil menyumbang secara meluas kepada penyelidikan pembelajaran mesin. Arah kajian masa depan termasuk mempelbagaikan dataset, membangunkan algoritma adaptif, dan merumuskan metodologi untuk menangkap hasil rawatan yang kompleks dengan lebih tepat, sekali gus membuka jalan untuk kemajuan dalam sistem sokongan keputusan yang didorong oleh AI untuk penjagaan kanser payudara.

IMPROVING THE FEATURE SELECTION, MULTI-CLASS CLASSIFICATION, AND IMBALANCED DATASET OF BREAST CANCER

ABSTRACT

The fact that breast cancer is prevalent among women makes it a serious problem in global health. Its early detection is crucial for improving treatment options and increasing survival rates, yet the complexity of diagnosing and determining the most effective treatment plan presents significant challenges. Recent years have seen the rise of AI and ML techniques as powerful tools in the fight against breast cancer, opening new possibilities for improving detection and treatment methods. This research aims to address key challenges in the application of AI/ML to breast cancer treatment planning, including imbalanced datasets, suboptimal feature selection methods, and the complexity of multi-class classification tasks. The study justifies its focus by addressing the unmet need for improved computational tools that can personalize and optimize treatment strategies for breast cancer patients. We aim to enhance model performance on imbalanced datasets by improving feature selection procedures, refine multi-class classification models, and to develop predictive models for personalized treatment planning. Successful completion of the thesis's objectives will allow it to make a substantial contribution to the creation of optimal treatment plans for individual patients. The effectiveness of existing feature selection methods and the performance of different ML algorithms were thoroughly tested using 5-fold cross-validation on the SEER and thyroid datasets. Results showed that XGBoost performed very well, while CatBoost and Decision Trees were hindered by their feature selection strategy. The introduction of the MC-SVF, an enhanced feature selection method, substantially improved the performance of these algorithms,

underscoring traditional feature selection methods' limitations in addressing imbalanced datasets and multi-class scenarios effectively. The research further demonstrated the enhanced multi-class classification model's substantial improvements in managing class imbalances and achieving class distinction, especially for minority classes, at the expense of some precision for increased recall rates by up to 60%. From a computer science perspective, this research makes a notable contribution to the field of AI-driven decision support systems. MC-SVF not only addresses the limitations of existing feature selection methods but also sets a new benchmark in handling imbalanced datasets and multi-class classification. The study concludes by proposing two predictive treatment models that integrate MC-SVF, paving the way for adaptive algorithms and methodologies to encapsulate complex treatment outcomes more effectively. These innovations have the potential to advance AI in healthcare while contributing broadly to machine learning research. Future research directions include diversifying datasets, developing adaptive algorithms, and formulating methodologies to encapsulate complex treatment outcomes more accurately, paving the way for advancements in AI-driven decision support systems for breast cancer care

.

CHAPTER 1

INTRODUCTION

1.1 Overview

Breast cancer is a condition in which the cells of the breast begin to grow out of control. There are several types of breast cancer, the kind of breast cancer is determined by which cells in the breast become cancerous. Breast cancer can start in a variety of places in the breast. Lobules (glands that make milk), ducts (tubes that transport milk from the breast to the nipple), and connective tissue (fibrous and fatty tissue) are the three major components of a breast. Breast cancer usually starts in the ducts or lobules, it can also spread to other parts of the body via blood and lymph arteries. In situ or non-invasive cancer cells remain inside the basement membrane of the components of the terminal duct lobular unit and the draining duct. Breast cancer that has spread outside the basement membrane of the ducts and lobules into the surrounding normal tissue is known as invasive breast cancer and is considered to have metastasized (Holmes, Carter, and Metefa 2000; What Is Breast Cancer 2020a; What Is Breast Cancer 2020b).

In 2020, 2.3 million women were diagnosed with breast cancer with 685000 fatalities worldwide. As of the end of 2020, 7.8 million women alive have been diagnosed with breast cancer in the past five years, making it the world's most prevalent cancer (Breast cancer 2021). Malignant tumours are typically classified as positive in clinical terms, while benign tumours are classified as negative. Both cancers have subgroups that must be identified separately since each might have a different prognosis and treatment plan. Accurate identification of each subcategory is needed for proper diagnosis. Mammography, ultrasound, Magnetic Resonance Imaging (MRI), Computed Tomography (CT), Positron-Emission Tomography (PET), and microwave imaging are now used in the diagnosis of breast cancer (Dorrius et al. 2011; Friedewald et al. 2014;

Giger 2018; Jalalian et al. 2013; Sheth and Giger 2020; Winsberg et al. 1967). In medical images, there are two types of breast cancer manifestations: masses and calcifications. On appearance, benign tumours are often spherical, smooth, and transparent. Calcification has a coarser form, a granular shape, a popcorn shape, or a ring shape, and it has a greater density and a more scattered dispersion (Figure 1.1). The margins of typical malignant tumours are uneven and typically fuzzy, and the mass has a needle-like appearance. Calcification has a morphology that is typically sand-like, linear, or branching, with a variety of forms and sizes, and the distribution is usually dense or clustered in a linear pattern (Abdel-Qader and Abu-Amara 2008; Gallego-Ortiz and Martel 2016; Newell et al. 2010). The following section presents more information on breast cancer diagnosis.

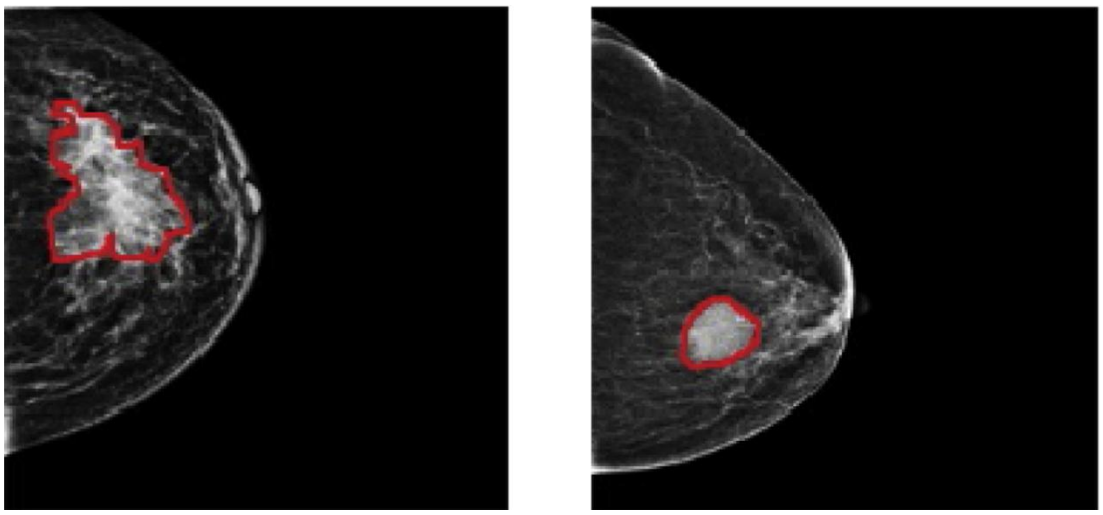

Figure 1.1 (Left) malignant and (right) benign breast cancer masses

1.2 Breast Cancer Diagnosis

Breast cancer is categorized into three subtypes based on specific biomarkers: Estrogen Receptor (ER), Progesterone Receptor (PR), and Human Epidermal Growth Factor 2 (ERBB2) gene amplification (formerly known as HER2). These subtypes include hormone receptor-positive (HR+), ERBB2-positive, and triple-negative. While HR+ and ERBB2+ subtypes have an average overall survival of five years, triple-

negative breast cancer is associated with a much shorter average survival of one year (Swain et al. 2015). Each of the three subtypes has its own set of risks and treatment options. The best treatment for each patient is determined by their tumor subtype, anatomic cancer stage, and personal preferences (Waks and Winer 2019).

There are several cancer staging methods in use right now. One method divides tumors into four stages: Stage 0, Stage I, Stage II, Stage III, and Stage IV, with further subcategories, where Stage IV denotes a metastatic distant cancer. TNM (Tumor, Node, Metastasis) is another cancer staging method that assigns stages based on the tumor, node, and metastases status (Edge and Compton 2010). Stage I breast cancer, defined anatomically as a breast tumor smaller than 2 cm and no lymph node involvement, have five years survival rate of at least 99% for HR+, at least 94% for ERBB2+, and at least 85% for triple-negative subtypes (Bardia et al. 2017).

1.2.1 Mammography

The US Food and Drug Administration (FDA) considers mammogram the best primary tool for breast cancer screening (Mammography: What You Need to Know 2021). X-rays are used to create pictures in screen-film mammography and full-field digital mammography, the picture is taken on film in screen-film mammography and is captured digitally in full-field digital mammography and then produced on film or interpreted straight from a computer monitor. By facilitating electronic transmission, storage, and retrieval, the digital acquisition process enhances logistics and workflow. Radiologists watching the image on a display may adjust the image's contrast and brightness, as well as enlarge regions, without exposing the patient to further x-rays (Elmore et al. 2005). To determine the morphology of a lesion on breast mammography, first determine if it is a mass, a localized lesion, or a non-mass-like enhancement. Mass and non-mass-like enhancement are defined in the Breast Imaging Reporting and Data

System (BI-RADS) lexicon as follows (Breast Imaging Reporting & Data System 2021):

- Mass: A mass is a three-dimensional space-occupying lesion that comprises one process, usually round, oval, lobular, or irregular in shape.
- Non-mass-like enhancement: Enhancement of an area that is not mass. This includes enhancement patterns that may extend over small or large regions, and whose internal enhancement characteristics can be described as a pattern discrete from normal surrounding breast parenchyma.

Breast cancer death rates have been demonstrated to be lowered by 30–70% when mammographic screening is used. The diagnostic accuracy of mammography depends on factors such as breast structure, density, and the radiologist's perception and level of experience (Bazzocchi et al. 2007). For single reading, almost 70% of all missed BCs are due to misinterpretation while 30% are overlooked lesions (Robyn L. Birdwell 2009). Mammograms are difficult to read, especially in the context of screening. Image quality and radiologist's skills have an impact on screening mammography sensitivity (Rangayyan, Ayres, and Leo Desautels 2007). Mammography, on the other hand, has the potential to miss cancer, particularly in women with dense breasts. Mammography alone may not be enough to detect breast cancer in women with a high risk of developing it. Newer imaging techniques for supplementary screening, such as Dynamic Contrast-Enhanced (DCE), MRI, Digital Breast Tomosynthesis (DBT), CT, PET, and Automated Whole Breast Ultrasound (AWBUS), have emerged in response to the need for more effective screening strategies to supplement mammography in these groups of women (Sheth and Giger 2020).

1.2.2 Magnetic Resonance Imaging

MRI uses magnetic fields to provide precise cross-sectional images of tissue structures, allowing for excellent soft tissue contrast. The mobility and magnetic environment of hydrogen atoms in water and fat that contribute to the measured signal that defines the brightness of tissues in the picture causes the contrast between tissues in the breast (fat, glandular tissue, lesions, etc.), this yields in images of the breast that mostly reveal parenchyma and fat, as well as any lesions that may be present. To reliably identify malignancies and other lesions, a paramagnetic tiny molecule gadolinium-based contrast agent is administered intravenously. Because the strong signal from enhancing lesions might be difficult to distinguish from fat, subtraction pictures or fat suppression, or both, are used to evaluate illness. MRI exams may consist of analyzing pictures at a single moment in time or, more commonly, collecting a pre-injection image followed by successive sets of images following contrast injection [DCE]-MRI. The appearance of lesions, as well as the uptake and washout pattern, when available, can be utilized to distinguish malignant illness from benign situations (Saslow et al. 2007). MRI is a helpful technique for detecting and identifying breast illness, determining the local extent of the disease, assessing therapy response, and guiding biopsy and localization. In some studies, the sensitivity of this modality in detecting invasive breast cancer has reached 100%, which is one of the reasons why breast MRI is crucial in preoperative staging. Breast MRI is now recommended as a complement to mammography for patients with a relative lifetime risk of higher than 20%, based on evidence from nonrandomized trials and observational studies (Lu, Li, and Chu 2017; Sheth and Giger 2020). The low-to-moderate specificity of breast MRI, which ranges from 37% to 97%, is a drawback as overtreatment might be caused by a lack of specificity (H. Cai et al. 2014; Roganovic et al. 2015). The high cost of MRI, as well as

the cost of invasive follow-up procedures associated with such low specificity, has limited its use as a screening tool for the public (Bluemke 2004; Newell et al. 2010). The Area Under Curve (AUC) of [DCE]-MRI is much lower than that of three-dimensional (3-D) MRI alone. [DCE]-MRI's kinetic patterns can aid in the early detection and categorization of breast lesions, but they should not be used as diagnostic criteria. Although the AUC achieved by combining DCE and 3-D MRI was like the one obtained by 3-D MRI alone, the specificity of a washout pattern without malignancy was improved. (Y. H. Huang et al. 2013). Breast MRI researchers have been working on developing quantitative diagnosis models. The goal is to create classifiers that offer the best diagnostic findings based on the lesion's morphological characteristics and dynamics (Newell et al. 2010).

1.2.3 Digital Breast Tomosynthesis

DBT is a type of imaging, or X-ray, in which multiple projections of the breast are obtained over a limited angular range to reconstruct a 3-D dataset of mammography images (Tomosynthesis: Cost & How It Compares to Mammograms 2021; Vedantham et al. 2015) that can be used to detect early indicators of breast cancer in patients who don't have any symptoms and can also be utilized to diagnose patients who are experiencing symptoms of breast cancer. Tomosynthesis is a form of enhanced mammography which was authorized by FDA in 2011. In comparison to FFDM, DBT detects 30–40% more cancers (Friedewald et al. 2014), the reading time has been doubled though and cognitive and perception errors still occur (Tagliafico et al. 2017).

1.3 Breast Cancer Treatment

There are two types of treatment for cancer, depending on the kind and stage of the disease: local and systemic. Surgery and radiation are considered local treatments

as they treat the tumor without harming the rest of the body. Systemic therapy, on the other hand, employs the use of medicines to combat disease. Drugs can reach cancer cells anywhere in the body and be administered directly into the bloodstream or orally. Systemic treatments include chemotherapy, hormone therapy, targeted medication therapy, and immunotherapy. Systemic treatment maybe preoperative (neoadjuvant), postoperative (adjuvant), or both (Breast Cancer: Symptoms, Risk Factors, Diagnosis, Treatment & Prevention 2020; Breast Cancer Treatment (Adult) (PDQ®)–Patient Version - National Cancer Institute 2020; Waks and Winer 2019). Most patients will have a combination of local treatments to control local disease and systemic treatment for any metastatic disease.

Machine learning (ML) models have demonstrated significant potential in predicting optimal treatment strategies for breast cancer, facilitating personalized and data-driven clinical decision-making. These models analyze complex, multidimensional datasets to forecast patient-specific responses to various treatment modalities, including chemotherapy, radiotherapy, surgery, and targeted therapies. For example, a study by (Braman et al. 2020) developed a deep learning model that predicts response to HER2-targeted neoadjuvant chemotherapy using pre-treatment dynamic contrast-enhanced MRI data, achieving an area under the curve (AUC) of 0.93 in validation cohorts. Additionally, (Gilad and Freiman 2022) introduced the PD-DWI model, which utilizes physiologically decomposed diffusion-weighted MRI data to predict pathological complete response to neoadjuvant chemotherapy, demonstrating improved AUC compared to conventional machine learning approaches. These advancements underscore the growing role of ML in enhancing treatment prediction accuracy, thereby contributing to more effective and personalized therapeutic strategies in breast cancer care.

1.3.1 Surgery

Breast conservation surgery (excision of the tumor with surrounding normal breast tissue) or mastectomy (removal of the entire breast) are two options for surgery (total removal of breast tissue). Because of their impact on local recurrence following breast-conserving surgery, some clinical and pathological variables may affect breast conservation or mastectomy choices. An inadequate initial excision, young age, the existence of a significant in situ component, lymphatic or vascular invasion, and histological grade are all factors to consider. Local recurrence is two to three times more probable in young individuals (under 35) than in older patients. While other risk factors for local recurrence are more common in young individuals, young age appears to be an independent risk factor (Association of Breast Surgery 2009).

1.3.1(a) Mastectomy

A mastectomy is a surgical procedure that removes the breast tissue and a portion of the underlying skin, which generally includes the nipple. A mastectomy should be paired with axillary lymph nodes surgery in some way. Lymph node ectomy is used for both diagnostic (determining the anatomic extent of breast cancer) and therapeutic purposes (removal of cancerous cells) (Waks and Winer 2019). About a third of locally advanced breast cancers are unsuitable for breast conservation surgery but may be treated with mastectomy. Some patients who are candidates for breast conservation surgery choose mastectomy instead. Until the demonstration of equivalent outcomes with mastectomy and breast-conserving surgery plus irradiation against the remaining breast in patients, mastectomy was the primary surgical treatment employed in the great majority of patients (Association of Breast Surgery 2009; Holmes, Carter, and Metefa 2000).

1.3.1(b) Breast Conservation Surgery

Breast conservation surgery may be done at any age, it consists of excision of the tumor with a 1 cm margin of normal tissue (broad local excision) or a more extensive excision of a complete quadrant of the breast (breast conservation surgery) (quadrantectomy). The extent of excision is the most critical factor that determines local recurrence following breast-conserving. Compared to grade II or III tumors, grade I tumors appear to have a 1.5-fold reduced recurrence rate. The lower the recurrence rate, but the poorer the aesthetic effect, the larger the excision. Although there is no size restriction for breast conservation surgery, adequate excision of lesions larger than 4 cm yields a poor aesthetic outcome. Hence most breast units limit breast-conserving surgery to lesions less than 4 cm (Holmes, Carter, and Metefa 2000; Møller et al. 2008; Senkus et al. 2015).

1.3.2 Radiotherapy

Breast cancer patients may get radiation treatment to the entire breast or a breast section (after lumpectomy), the chest wall (after mastectomy), and the regional lymph nodes. Whole-breast radiation after a lumpectomy is a standard part of breast-conserving treatment (Fisher et al. 2002). After breast conservation surgery, postoperative radiotherapy is strongly advised. Whole breast radiation therapy alone lowers the ten-year risk of any first recurrence (both local and distant) by 15% and the fifteen-year risk of breast cancer-related death by 4% (Senkus et al. 2015). Radiation to the chest wall, occasionally with a boost to the mastectomy scar and regional nodal radiation, is known as Post Mastectomy Radiation (PMRT). PMRT lowers the ten-year risk of recurrence (including locoregional and distant) by 10% and the twenty-year risk of breast cancer-related death by 8% in node-positive patients. The advantages of

PMRT are unaffected by the number of implicated axillary lymph nodes or the use of adjuvant systemic therapy (McGale et al. 2014).

1.3.3 Systemic Treatment

In the United States, 5.8% of breast cancer patients are metastatic, with a five-year survival rate of 29% (SEER*Explorer Application 2021). The best treatment for Metastatic Breast Cancer (MBC) remains a substantial therapeutic problem; the best medical therapy for each patient must be chosen based on breast cancer risk assessment, predictive indicators, toxicity risk, and patient preferences (Bernard-Marty, Cardoso, and Piccart 2004). Adjuvant systemic treatment should begin as soon as possible following surgery, preferably within two to six weeks (Senkus et al. 2015). There are a few broad guidelines to follow: in metastatic HR+/HER2 breast cancer, early treatment should be centered on endocrine therapy (ET). Patients are transitioned to chemotherapy (CT) after developing resistance to the available hormonal therapies (Cardoso et al. 2009). ET is still the most effective treatment for hormone-sensitive, non-life-threatening MBC. This systemic medication offers the benefits of effectiveness, low toxicity, and high quality of life (Bernard-Marty, Cardoso, and Piccart 2004). ET is used in clinical practice when the primary tumor or, if possible, a readily accessible metastasis is ER+, PR+, or HR+. When the danger of rapid disease development is modest, i.e., there is no life-threatening sickness, this sort of therapy is typically the first choice (Waks and Winer 2019). CT is presently the sole treatment option for women with endocrine resistant illnesses who are ER- and HER2- (Breast, Trialists, and Group 2008).

Neoadjuvant chemotherapy is used to treat localized early-stage triple-negative breast cancer (TNBC) to preserve the breast or for patients who are temporarily unable to undergo surgery. Chemotherapy in the neoadjuvant situation allows for a direct

clinical examination or imaging evaluation of the response (Lebert et al. 2018). TNBC has a more significant percentage of pathologic complete response (PCR) after neoadjuvant chemotherapy than HER2- illness (28% –30% vs. 6.7%) (Cardoso et al. 2009).

1.4 Breast Cancer Datasets

1.4.1 DDSM: Digital Database for Screening Mammography

DDSM is a resource for use by the mammographic image analysis research community (USF Digital Mammography Home Page 2021). The database contains approximately 2,500 studies, each one includes two images of each breast, along with some associated patient information (age at time of study, ACR breast density rating, subtlety rating for abnormalities, and ACR keyword description of abnormalities) and image information (scanner, spatial resolution, etc.). Images containing suspicious areas have associated pixel-level "ground truth" information about the locations and types of suspicious regions. DDSM is organized into "cases" and "volumes." A "case" is a collection of images and information corresponding to one mammography exam of one patient. A "volume" is simply a collection of cases collected for purposes of ease of distribution. All volumes are available on 8mm tape. Each case in this volume of cancer cases has at least one path-proven cancer. Some cases contain more than one cancer in one breast, a cancer in each breast, or a cancer along with other abnormal/suspicious regions. The outlines of all regions have been transcribed from markings made by an experienced mammographer. A case contains six to ten files, classified into four categories:

- "ics" file contains some information about the images, such as the age of the patient, the size of the mammograms, whether a file exists for the overlay of abnormality outlines or not, etc.
- "16-bit PGM" file: overview of the real mammograms.
- "ljpeg" file contains four image files that are compressed with lossless JPEG encoding.
- "overlay" files: give the keyword description for a given abnormality in each view, while normal cases will not have any overlay files.

The Curated Breast Imaging Subset (CBIS-DDSM) collection includes a subset of the DDSM data selected and curated by a trained mammographer. The images have been decompressed and converted to Digital Imaging and Communications in Medicine (DICOM) format. Updated Region of Interest (ROI) segmentation and bounding boxes, and pathologic diagnosis for training data are also included (Lee et al. 2017a).

1.4.2 MIAS: Mammographic Image Analysis Society

MIAS, an organization of UK research groups interested in understanding mammograms, has produced a digital mammography database (PEIPA, the Pilot European Image Processing Archive 2021). The X-ray films in the database have been carefully selected from the United Kingdom National Breast Screening Program and digitized with a Joyce-Lobel scanning microdensitometer to a resolution of 50 μm × 50 μm, a device linear in the optical density range 0-3.2 and representing each pixel with an 8-bit word. The database contains left and right breast images for 161 patients and is available on a Digital Audio File – Digital Data Storage (DAT-DDS) tape. There are 208 normal, 63 benign, and 51 malignant (abnormal) images. It also includes the radiologist's ground truth markings on the locations of any abnormalities that may be present. For each film, experienced radiologists give the type, location, scale, and other

helpful information. According to these experts' descriptions, the database concludes four kinds of abnormities (architectural distortions, stellate lesions, circumscribed mass, and calcifications). The database possesses an introduction file, which includes the following information:

- Type: type of abnormity.
- Sort: whether the abnormities are cancer or benign ones.
- Location and size: the original coordinates and diameters of the abnormities.

1.4.3 WBCD: Wisconsin Breast Cancer Dataset

WBCD was created by Dr. William H. Wolberg, a physician at the University of Wisconsin Hospital in Madison, Wisconsin, USA (UCI Machine Learning Repository: Breast Cancer Wisconsin (Diagnostic) Data Set 2021). To create the dataset, Dr. Wolberg used fluid samples taken from patients with solid breast masses and an easy-to-use graphical computer program called Xcyt, which can perform the analysis of cytological features based on a digital scan. Features are computed from a digitized image of a breast mass's Fine Needle Aspirate (FNA). They describe the characteristics of the cell nuclei present in the image. The program uses a curve-fitting algorithm to compute ten features from each of the cells in the sample, then it calculates the mean value, extreme value, and standard error of each feature for the image, returning a thirty real-valuated vector.

Attribute Information:

1) ID number
2) Diagnosis (M = malignant, B = benign)
3) Ten real-valued features are computed for each cell nucleus:
    a. Radius (mean of distances from the centre to points on the perimeter)
    b. Texture (standard deviation of gray-scale values)

c. Perimeter

d. Area

e. Smoothness (local variation in radius lengths)

f. Compactness: $\frac{Perimeter^2}{area} - 1.0$

g. Concavity (severity of concave portions of the contour)

h. Concave points (number of concave portions of the contour)

i. Symmetry

j. Fractal dimension ("coastline approximation" - 1)

1.4.4 INBreast Database

Mammography images of INBreast database were originally collected from Centro Hospitalar de S. Joao [CHSJ], Breast center, Porto (Moreira et al. 2012). INBreast database collects data from August 2008 to July 2010, which contains 115 cases with a total of 410 images. Among them, 90 cases were women with the disease on both breasts. There are four different types of breast diseases recorded in the database:

- Mass
- Calcification
- Asymmetries
- Distortions

The images of this database have two perspectives of Craniocaudal (CC) and Mediolateral Oblique (MLO), and the breast density is divided into four categories according to BI-RADS standards, which are Entirely fat (Density 1), Scattered fibro glandular densities (Density 2), Heterogeneously dense (Density 3), and Extremely dense (Density 4). Images were saved in two sizes: 3328 x 4084 or 2560 x 3328 pixels in DICOM. The main characteristic of this work is the carefully associated Ground

Truth (GT) annotations made by a specialist in the field and validated by a second specialist.

1.4.5 SEER Dataset

The Surveillance, Epidemiology and End Results (SEER) Program of the National Cancer Institute (NCI) is a trustworthy source of information on cancer incidence and survival in the United States. SEER now collects and publishes cancer incidence and survival data from community-based cancer registries covering about 47.9% of the US population. The SEER Program registries routinely gather data on patient demographics, initial tumor site, tumor shape and stage at diagnosis, the first course of therapy, and vital status follow-up (Surveillance, Epidemiology, and End Results Program 2021). The data is publicly available and can be obtained after signing a data use agreement. At the time of this research, its latest version covers identified cancer incidences from the years 1973 through 2018. In 2018 there were an estimated 3,676,262 women living with female breast cancer in the United States, the rate of new cases of female breast cancer was 129.1 per 100,000 women per year. The death rate was 19.9 per 100,000 women per year (Female Breast Cancer - Cancer Stat Facts 2022).

1.5 Problem Statement

Breast cancer is one of the most common forms of cancer faced by women in the whole world. Early detection can open various options for treatment and increase the chances of survival (Pe and Sipper 1999). Because of the complexity of breast cancer, as well as the challenges associated with identifying and treating it, it is necessary to utilize cutting-edge technology to improve clinical outcomes and survival chances of patients (Weinstein et al. 2021; Zhao et al. 2015; Zheng and Chan 2001). When breast cancer is not detected on time, it can become a cause of many major other

diseases, which may lead to the death of the patient (Bocchi et al. 2004). When it comes to achieving an accurate diagnosis of breast cancer, one of the most critical phases is the selection of appropriate features. Several factors, including the differentiation between benign and malignant tumors and the identification of distinct subtypes of cancer, are helpful in this regard. The precision with which breast cancer stages and types may be categorized, which is in turn impacted by successful feature selection, has an impact on the different treatment options available to patients as well as the results they experience (Bardia et al. 2017; Edge and Compton 2010; Swain et al. 2015; Waks and Winer 2019). Feature selection becomes even more pivotal in the context of multi-class classification problems, particularly when dealing with imbalanced datasets. Breast cancer datasets often suffer from class imbalance, where certain classes (e.g., specific subtypes of breast cancer) are underrepresented. This imbalance can lead to biased models that perform well on majority classes but poorly on minority classes, potentially resulting in inadequate treatment plans for patients with less common cancer subtypes (H. He and Ma 2013). In multi-class classification, where the goal is to accurately classify instances into one of several classes, the choice of features can significantly influence model performance. Effective feature selection helps in reducing dimensionality, eliminating noise, and enhancing model interpretability (Guyon and Elisseeff 2003). Furthermore, feature selection directly affects the prediction of treatment plans. Breast cancer treatment often involves a combination of surgery, chemotherapy, radiation therapy, and hormone therapy. The choice and sequence of these treatments can vary greatly depending on the cancer's stage, subtype, and other patient-specific factors. By selecting relevant features that accurately reflect these factors, machine learning models can provide more personalized and effective treatment recommendations (Parker et al. 2009). In summary, effective feature selection is crucial

not only for improving the classification accuracy of breast cancer stages and subtypes but also for ensuring that treatment plans are tailored to the individual patient's needs. Addressing class imbalance and selecting features that enhance model performance can lead to better clinical outcomes and more precise treatment strategies.

Breast cancer clinical diagnosis is a complex topic in the biomedical field due to imbalanced data distribution and imbalanced quality of the majority and minority classes, which leads to misclassification. Thus, the development of sophisticated multi-class classification models is of the utmost importance. These models need to be able to overcome the challenges that are provided by imbalanced datasets to correctly classify cancer subtypes in a timely way that is clinically useful. Although the majority class samples and their appropriate classification are more crucial to the classifier, cancer is diagnosed using samples from the minority class (cancer data class). While an incorrect diagnosis results in more clinical testing for non-cancerous patients, cancer patients pay the price with their lives. As a result, research into the problem of class imbalance is critical (Fotouhi, Asadi, and Kattan 2019; Majid et al. 2014a; Saini and Susan 2020).

Several Computer Aided Diagnosis (CAD) systems have been developed since the late 1960's. Computer-Aided Detection (CADe) systems and Computer-Aided Diagnostic (CADx) systems are the two types of CAD systems. The localization job (identification of a suspicious abnormality) is the focus of CADe, which acts as a second reader for radiologists and leaves patient care decisions to the radiologist. However, CADx classifies an abnormality that is detected by a radiologist or a computer, estimating the likelihood of an abnormality and classifying it as benign or malignant. The radiologist next evaluates if the anomaly deserves additional investigation and the clinical relevance of the finding (Giger 2018). CADe systems are hampered by high

false-positive rates, while CADx systems for mammography are not yet authorized for clinical usage, despite encouraging findings (Gardezi et al. 2019; Zou et al. 2019). Although the technical challenges of CAD in mammography have been significant, one cannot correctly evaluate these systems. According to our study of CAD literature, data sources and dataset sizes are inconsistent. Furthermore, because most evaluation datasets are not accessible, only a handful of the reported outcomes may be directly reproduced. It is hard to compare approaches rigorously without using common datasets, and thus mammography CAD research is impeded (Lee et al. 2017b). Additionally, owing to the ever-changing nature of breast cancer, which comprises several subtypes and stages, there is a need for prediction models that can develop tailored treatment regimens by making use of significant patient and laboratory data. In the present landscape of breast cancer treatment, which encompasses several choices ranging from cancer surgery and radiation therapy to systemic therapies such as chemotherapy and hormone therapy, there is an urgent need for predictive models that can make use of specific patient data in order to provide recommendations for the most effective treatment regimens. These models have the potential to substantially alter the treatment that is provided to breast cancer patients. By developing individualized treatment plans that take into consideration the particulars of each patient's ailment, they are able to maximize the efficacy of treatment and enhance the results for patients (Bernard-Marty, Cardoso, and Piccart 2004; Braman et al. 2020; Cardoso et al. 2009; Fisher et al. 2002; Gilad and Freiman 2022; Holmes, Carter, and Metefa 2000; McGale et al. 2014; Møller et al. 2008; Waks and Winer 2019). Existing research demonstrates the potential of machine learning (ML) in predicting breast cancer treatment outcomes, particularly through models focused on specific modalities like chemotherapy or targeted therapies. Although these studies highlight the promise of imaging-based

models in treatment prediction, their reliance on specialized imaging data limits their applicability in broader clinical settings, where such data may not be readily available. Furthermore, these works focus on single-task predictions, neglecting the complexity of integrating multiple treatment modalities into a cohesive recommendation framework.

The problem can be summarized as follows:

1. Effective feature selection is crucial for accurately categorizing breast cancer stages and types, which directly influences the variety of treatment options available to patients and their clinical outcomes. By selecting relevant features that reflect critical factors, machine learning models can provide personalized and effective treatment recommendations.
2. Medical datasets, particularly those related to breast cancer, are often imbalanced, leading to misclassification issues. This imbalance results in biased models that perform well on majority classes but poorly on minority classes, potentially leading to inadequate treatment plans for patients with less common cancer subtypes. Addressing class imbalance is essential for improving the accuracy and reliability of diagnostic models.
3. Clinical Decision Support (CDS) systems, including Computer-Aided Detection (CADe) and Computer-Aided Diagnostic (CADx) systems, are not yet widely implemented. Most research has been conducted in academic settings, focusing on specific stages of breast cancer. There is limited evidence supporting the performance of these systems in advanced breast cancer stages, and high false-positive rates further complicate their clinical adoption.

1.6 Research Objectives

Taking into consideration all that has been mentioned above, the primary objective of this thesis is to utilize cutting-edge ML techniques to address significant gaps in breast cancer treatment planning. The purpose of this research is to enhance the survival rates and quality of life of breast cancer patients worldwide. This will be accomplished by focusing on:

1. Enhancing feature selection procedures to improve the performance of ML models on imbalanced datasets, ensuring accurate classification and effective treatment recommendations.
2. Refining multi-class classification models to handle imbalanced datasets more effectively, addressing the challenges of misclassification and improving model performance.
3. Designing and implementing prediction models for individualized treatment planning, leveraging patient-specific data to provide tailored and optimal treatment strategies.

The goal of this project is to contribute to the advancement of breast cancer treatment techniques that are specifically optimized for each patient, thereby maximizing treatment efficacy and improving clinical outcomes.

1.7 Research Contribution

With fast-changing scientific findings, pharmacological approvals, and treatment recommendations, breast cancer radiologists are required to diagnose breast cancer early, and oncologists are required to tailor treatment plans. CDS systems that use artificial intelligence (AI) have the potential to help solve this problem. This

research contributes to the recommendations for an effective treatment plan in the following manner:

1. Design and implement an improved algorithm for feature selection.
2. Design and implement an improved algorithm for multi-class classification models for imbalanced datasets.
3. Design and implement treatment plan prediction models based on the properties of the diagnosed cancer and set of lab results.

By focusing on these key areas, this research aims to advance the field of breast cancer treatment, providing radiologists and oncologists with sophisticated tools to improve patient care and outcomes.

1.8 Research Scope and Limitations

Leveraging the SEER dataset to create and evaluate sophisticated ML algorithms and models with the goal of improving the feature selection process and prediction outcomes of multi-class classification models on imbalanced datasets, which will be used to predict treatment plans for breast cancer patients, with a particular emphasis on surgical procedures, radiation, and chemotherapy alternatives, is the primary objective of this project. By concentrating on this dataset, the research intends to make use of a diverse range of tumor features to develop models that are capable of properly predicting the treatment modalities that will be most beneficial for specific patients based on their profiles.

However, the research is limited by the intrinsic limits of ML multi-class classification algorithms when it comes to dealing with extremely imbalanced datasets. This may have an impact on the accuracy of predictions for outcomes that are less prevalent. In addition, the research is vulnerable to a few limitations that are intrinsic to

the SEER dataset as well as the complexity of cancer treatment prediction. Although the SEER dataset provides a comprehensive coverage of cancer incidence, it is possible that it does not provide specific information on all the factors that impact treatment decisions. The capacity of the proposed feature selection algorithm and the treatment prediction models to adequately reflect the intricacies of tailored treatment planning may be hindered because of this constraint. In addition, the emphasis placed on surgical procedures, radiation, and chemotherapy as treatment modalities, even though these are the principal alternatives that are accessible, does not take into consideration more recent and developing therapies like immunotherapy and targeted therapy. The difficulty of constructing a comprehensive prediction model that can adapt to the fast-changing environment of breast cancer treatment is brought to light by these constraints. A summary of the research scope and limitations is depicted in the Table 1.1.

Table 1.1 Research scope and limitations

Item	**Scope of Research**
Treatment prediction dataset	SEER dataset
Dataset type	Imbalanced dataset
Predicted treatments	Surgery, Radiotherapy, and chemotherapy

Despite these limitations, the research endeavors to make significant contributions to the field of ML and oncology by providing insights and tools that can assist in the selection of appropriate features to improve the performance of multi-class classification models and the personalization of breast cancer treatment.

1.9 Thesis Outline

The remainder of this thesis is organized into four chapters which are arranged as follows:

Chapter 2: presents a review on existing CAD systems for breast cancer diagnosis, CDS systems, basic concepts of machine learning, deep learning and classification, feature selection, and evolution of AI in breast cancer diagnosis. This chapter reviews related work in breast cancer diagnosis using supervised and unsupervised methods and then discusses the background of feature selection and imbalanced datasets. Finally, this chapter provides a critical analysis of the challenges that motivate this research.

Chapter 3: presents the research methodology with respect to the proposed approaches towards achieving the aim of this research. The overall framework of the study is discussed, the dataset preprocessing steps, and the prerequisites for the treatment plan prediction system.

Chapter 4: presents the methodology for the enhanced feature selection algorithm, the experiments performed to evaluate its capability, and comparisons with the state-of-the-art feature selection algorithms.

Chapter 5: presents the methodology for the proposed enhanced multi-class classification model for imbalanced datasets, the experiments performed to evaluate it, and comparisons with the common techniques used in multi-class classification models.

Chapter 6: discusses the experiments that reveal the successful performance of the proposed BC treatment prediction models. It also investigates the performance of the suggested combination of feature selection, and all introduced model decomposition approaches, as well as their potential to optimize classification task accuracy.

Chapter 7: the chapter summarizes the key findings for each of the three objectives, their limitations, and provides a conclusion and possible future work.

CHAPTER 2

LITERATURE REVIEW

This chapter presents a detailed and comprehensive background on CAD and CDS systems and the foundational concepts related to this research. It discusses the related works on CAD and CDS systems for breast cancer. Besides, it highlights the limitations of each solution, which motivates this research. The organization of the chapter is as follows: Section 2.1 provides background on CAD systems. Section 2.2 provides background on the CDS system. Section 2.3 provides an overview of AI in Breast Cancer. Section 2.4 describes the evolution of AI in Breast Cancer Diagnosis. Section 2.5 reviews feature selection, and the different methods used in the literature. Section 2.6 provides a review for multi-class classification on imbalanced datasets. Section 2.7 reviews the classification metrics used in classification models. Section 2.8 provides a critical analysis of the related studies. Finally, Section 2.9 summarizes the chapter.

2.1 Computer-Aided Diagnosis

Since the late 1960s, CAD systems for mammography have been under progress. Its major goal is to aid radiologists in detecting malignancies that might otherwise go undetected (Winsberg et al. 1967). CAD programs identify high-density regions and microcalcifications. In 1998, the US Food and Drug Administration approved the first CAD software for screening mammography, R2 Image Checker made by R2 Technologies (now known as Hologic) (Premarket Approval (PMA) 1998). Early findings were encouraging (R. L. Birdwell et al. 2001; Freer and Ulissey 2001), and by 2016 CAD has become extensively used in clinical practice, with roughly 92% of all mammography facilities in the United States adopting it (Keen, Keen, and Keen 2018).