Towards a Trustworthy Data-Driven Clinical Decision Support System: Breast Cancer Use-Case

by

Aya Farrag

A THESIS
SUBMITTED TO THE DEPARTMENT OF COMPUTER SCIENCE
AND THE FACULTY OF GRADUATE STUDIES
OF LAKEHEAD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE (SPECIALIZATION IN ARTIFICIAL INTELLIGENCE)

Supervisory Committee

Dr. Zubair Fadlullah
Supervisor
(*External Adjunct Professor, Department of Computer Science, Lakehead University, Thunder Bay, Ontario, Canada.*
Associate Professor, Department of Computer Science, University of Western Ontario, London, Ontario, Canada.)

Dr. Mostafa Fouda
Co-Supervisor
(*Assistant Professor, Department of Electrical and Computer Engineering, College of Science and Engineering, Idaho State University, Pocatello, ID, USA*)

Examination Committee

Dr. Muhammad Asaduzzaman
Internal Examiner
(*Assistant Professor, Department of Computer Science, Lakehead University, Thunder Bay, Ontario, Canada.*)

Dr. Al-Sakib Khan Pathan
External Examiner
(*Professor, Department of Computer Science and Engineering United International University (UIU), Bangladesh*)

ABSTRACT

Artificial Intelligence (AI) research has emerged as a powerful tool for health-related applications. With the increasing shortage of radiologists and oncologists around the world, developing an end-to-end AI-based Clinical Decision Support (CDS) system for fatal disease diagnosis and survivability prediction can have a significant impact on healthcare professionals as well as patients. Such a system uses machine learning algorithms to analyze medical images and clinical data to detect cancer, estimate its survivability and aid in treatment planning. We can break the CDS system down into three main components: the Computer-Aided Diagnosis (CAD), the Computer-Aided Prognosis subsystem (CAP) and the Computer-Aided Treatment Planning (CATP). The lack of trustworthiness of these subsystems is still considered a challenge that needs to be addressed in order to increase their adoption and usefulness in real-world applications. In this thesis, using the breast cancer use case, we propose new methods and frameworks to address existing challenges and research gaps in different components of the system to pave the way toward its usage in clinical practice.

In cancer CAD systems, the first and most important step is to analyze medical images to identify potential tumors in a specific organ. In dense prediction problems like mass segmentation, preserving the input image resolution plays a crucial role in achieving good performance. However, this resolution is often reduced in current Convolution Neural Networks (CNN) that are commonly repurposed for this task. In Chapter 3, we propose a double-dilated convolution module in order to preserve spatial resolution while having a large receptive field. The proposed module is applied to the tumor segmentation task in breast cancer mammograms as a proof-of-concept. To address the pixel-level class imbalance problem in mammogram screenings, different loss functions (i.e., binary cross-entropy, weighted cross-entropy, dice loss, and Tversky loss) are evaluated. We address the lack of transparency in current medical image segmentation models by employing and quantitatively evaluating different explainability methods (i.e., Grad-CAM, Occlusion Sensitivity, and Activation visualization) for the image segmentation task. Our experimental analysis shows the effectiveness of the proposed model in increasing the similarity score and decreasing the miss-detection rate.

Following the cancer diagnosis step, in Chapter 4, we propose a new framework for cancer survival prediction in CAP systems to precisely predict the estimated survival months of patients in order to facilitate treatment planning. We combine two main strategies in solving the cancer survivability prediction problem using Machine Learning techniques. In the first strategy, we model the survivability prediction task as a two-step problem, namely a classification problem to predict whether or not a patient survives for five years, and a regression problem to forecast the number of remaining months for those who are predicted to not survive for five years. The second strategy is to develop stage-specific

models, where each model is trained on instances belonging to a certain cancer stage in order to precisely predict survivability of patients from the same stage. We investigate the impact of adopting these strategies along with applying different balancing techniques over the model performance using breast cancer clinical data. The obtained results demonstrate the effectiveness of stage-specific modeling in both survivability classification and regression.

To incorporate the role of prognosis in determining the most suitable treatment plans for a cancer patient, in Chapter 5, we propose a novel survival-based framework for treatment planning. We employ the prediction models developed for stage-specific survival estimation to determine the best possible treatment plans for breast cancer patients in terms of their prognostic outcomes. The system generates an ordered list of all possible combinations of treatments associated with their predicted survival outcomes to offer more comprehensive treatment recommendations. To address the lack of explainability in current systems, we provide visualized explanations for the predicted survival outcome of different treatment plans. By integrating survival prediction models into treatment planning, healthcare providers can offer better patient care and help patients and their families make more informed decisions about the most appropriate course of treatment.

Experiments conducted in different chapters of this thesis demonstrate that the proposed AI-enabled techniques can improve the reliability and explainability of Clinical Decision Support Systems to help clinicians make patient-specific assessments and treatment decisions.

ACKNOWLEDGEMENTS

In the name of God, the most gracious and the most merciful. I would like to begin by acknowledging God's blessings and kindness that have enabled me to complete this thesis. I am deeply grateful for the strength and opportunities that He has bestowed upon me throughout this journey, and I would not have been able to make it without His generosity. I would like to express my sincere appreciation and gratefulness to Dr. Zubair Fadlullah, my supervisor, and Dr. Mostafa Fouda, my co-supervisor, for their continuous guidance and support throughout my master's study. I am also thankful to all of my colleagues in ACCESS Lab for the research discussions and support. I would like to extend my thanks to the examiners (Dr. Muhammad Asaduzzaman and Dr. Al-Sakib Khan Pathan) for their valuable comments and feedback that greatly helped me improve my thesis.

PUBLICATIONS

Parts of this thesis have been published:

- A. Farrag, Z. M. Fadlullah, and M. M. Fouda, **"A Two-Step Machine Learning Model for Stage-Specific Disease Survivability Prediction"** is published in the 2022 IEEE International Conference on Internet of Things and Intelligence Systems (IoTaIS), doi: 10.1109/IoTaIS56727.2022.9975966. (Chapter 4)

  Disclaimer: Chapter 4 is an IEEE publication and we are adhering to IEEE's copyright rules to report it.

Other parts of this thesis are submitted or to be submitted for publication:

- A. Farrag, G. Gad, Z. M. Fadlullah, and M. M. Fouda, **"An Explainable Medical Image Segmentation System With Preserved Local Resolution: Mammogram Tumor Segmentation"** is submitted to IEEE Access Journal. (Chapter 3)
- A. Farrag, Z. M. Fadlullah, and M. M. Fouda, **"Survival-Based Treatment Planning using Stage-Specific Machine Learning Models"** is to be submitted to IEEE Access Journal. (Chapter 5)

Apart from this, during my MSc. study, I also co-authored papers outside the scope of this thesis:

- G. Gad, A. Farrag, Z. M. Fadlullah, and M. M. Fouda,**"Communication-Efficient Federated Learning in Drone-Assisted IoT Networks: Path Planning and Enhanced Knowledge Distillation Techniques"** has been submitted for peer review in the 2023 IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC).
- G. Gad, A. Farrag, A. Aboulfotouh, K. Bedda, Z. M. Fadlullah, and M. M. Fouda, **"Communication-Efficient Federated Learning on Drone-aided LoRa network using Self-Organizing Maps and Knowledge Distillation"** is to be submitted to the IEEE IoT Journal.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

A Clinical Decision Support (CDS) System is a type of software system that uses patient data and medical images to assist healthcare providers in making clinical decisions [2]. The system can analyze patient data, provide diagnostic suggestions, and predict the likelihood of specific outcomes based on the patient's condition and medical history. Artificial Intelligence (AI) research has emerged as a powerful tool for health-related applications, with a growing number of studies and initiatives exploring its potential to transform healthcare. Hence, there is now an increasing trend toward developing CDS systems that leverage advanced technologies such as artificial intelligence, machine learning, and big data to enable the analysis of vast amounts of data and recognize patterns that are unobtainable by humans [3]. An end-to-end CDS system typically includes three main subsystems for diagnosis, prognosis and treatment planning. By combining these components, CDS systems can provide medical professionals with a comprehensive approach to personalized patient care in different steps of the clinical decision-making process.

Computer-aided diagnosis (CAD) is a component of CDS systems that uses algorithms and artificial intelligence techniques to analyze patient data and support medical professionals in making accurate diagnoses [4]. AI-enabled CAD systems can provide medical professionals with a data-driven approach to diagnosis, enabling them to make more accurate and timely decisions. By applying Machine Learning techniques, these systems can identify patterns and relationships that may not be immediately apparent to clinicians. Additionally, CAD systems can help reduce errors, improve diagnostic accuracy, and increase efficiency in healthcare delivery. This type of system can be used in a variety of medical fields, including radiology, pathology, and cardiology. There are various types of patient data integrated with CAD systems, such as medical history, diagnostic test results, and imaging data. However, imaging data is considered to be the most common type of patient data used in computer-aided diagnosis [5]. CAD systems can analyze various types of medical images, such as X-rays, computed tomography (CT) scans, magnetic resonance imaging

(MRI), and ultrasound scans, to support medical professionals in diagnosing various conditions. For example, by identifying and segmenting suspicious areas in mammograms, CAD systems can help radiologists detect and diagnose breast cancer at its early stages.

After a fatal medical condition is diagnosed, medical professionals need to determine the likely course of the disease and potential outcomes in order to make informed treatment decisions [6]. This is where computer-aided prognosis (CAP) can be particularly useful. The CAP system is the second component of clinical decision support (CDS) systems that uses predictive algorithms and artificial intelligence techniques to estimate a patient's survival outcomes based on their medical data. CAP systems can analyze patient data, such as medical history, lab results, and imaging data, and apply machine learning algorithms to identify patterns and relationships that may not be immediately apparent to human clinicians. These systems can provide medical professionals with valuable information, such as the likelihood of disease progression and the estimated length of survival [7]. It can also help patients and their families make more informed decisions about their medical care needs, financial needs, and other aspects of their lives.

In the context of clinical decision-making, treatment planning comes as a crucial step that determines the most suitable treatment regimen for a patient based on their clinical information and predicted prognosis. Treatment planning involves choosing from different treatment options, such as chemotherapy, surgery, radiation therapy, or a combination of these, to create a personalized treatment plan for each patient. Computer-Aided Treatment Planning (CATP) aims to help medical practitioners make more accurate and informed decisions, taking into account all the relevant clinical data available [8]. The treatment plan should take into account various factors, such as the stage and grade of cancer, the patient's age, overall health, and predicted survivability. By incorporating this information into CATP systems, medical professionals can provide more personalized care to their patients and make more informed decisions about the most appropriate course of action.

Despite the potential benefits of clinical decision support (CDS) systems, there is still a long way to go before they are fully trusted by medical professionals [9]. One of the main reasons for this lack of trust is the concern that CDS systems may not be accurate or reliable enough to make decisions that impact patient care. To optimize the effectiveness and adoption of these tools, it is crucial for researchers and developers to improve the performance of existing methods and understand the preferred mode of assistance by clinicians for each type of clinical task [5]. In addition, there is a need for greater transparency in how these systems are developed and how they generate their predictions so that healthcare providers can better understand and trust the recommendations made by these systems [10]. Moreover, to gain the trust of the medical community, CDS systems need to be integrated with medical experts rather than seen as a replacement for them. While CDS systems can provide valuable information and recommendations, they should be viewed as

tools to augment the decision-making capabilities of healthcare providers rather than as a substitute for their clinical judgment and expertise [11].

This dissertation extends previous research on CDS systems to contribute to improving the trustworthiness of these systems. Using the breast cancer use case, we investigate the development process of a trustworthy data-driven CDS system by solving some of the challenges existing in the different components of the system. In each component, we present new frameworks and techniques to improve the transparency and reliability of the system. Our objective is to enhance the reliability and transparency of these systems to push forward their integration into clinical practice. To achieve this, we explore various frameworks and techniques that have the potential to improve the performance of each component of the system. We also focus on improving transparency throughout the development of the system by promoting inherent explainability over methods that only approximate explainability. By following these general concepts, we aim to develop a trustworthy and reliable CDS system that can enhance clinical decision-making and ultimately improve patient outcomes. To sum up, the integration of AI-enabled CDS systems in clinical practice can facilitate healthcare providers as follows:

- Improving diagnostic accuracy and efficiency by analyzing vast amounts of patient data
- Helping to identify rare or complex conditions that may be overlooked by human clinicians
- Reducing the time needed for data analysis and interpretation
- Enhancing resource allocation by identifying patients who may require more intensive or specialized care
- Providing personalized treatment plans and shared decision-making with patients

Figure 1.1 outlines the organization and brief contributions in all chapters. As illustrated in the figure, this thesis investigates the development and viability of an AI-enabled CDS system, which can be combined with medical knowledge to make clinical decisions, including diagnosis, prognosis and treatment planning. As illustrated in the figure, the remaining chapters of this thesis are organized as follows:

Chapter 2 provides a general overview of clinical decision support systems and explains the motivation behind our study. We discuss the three components of the system considered in this thesis, namely, computer-aided diagnosis (CAD), computer-aided prognosis (CAP), and Computer-aided treatment planning (CATP). We explore the challenges associated with each component and review the literature to identify potential solutions that can address these challenges.

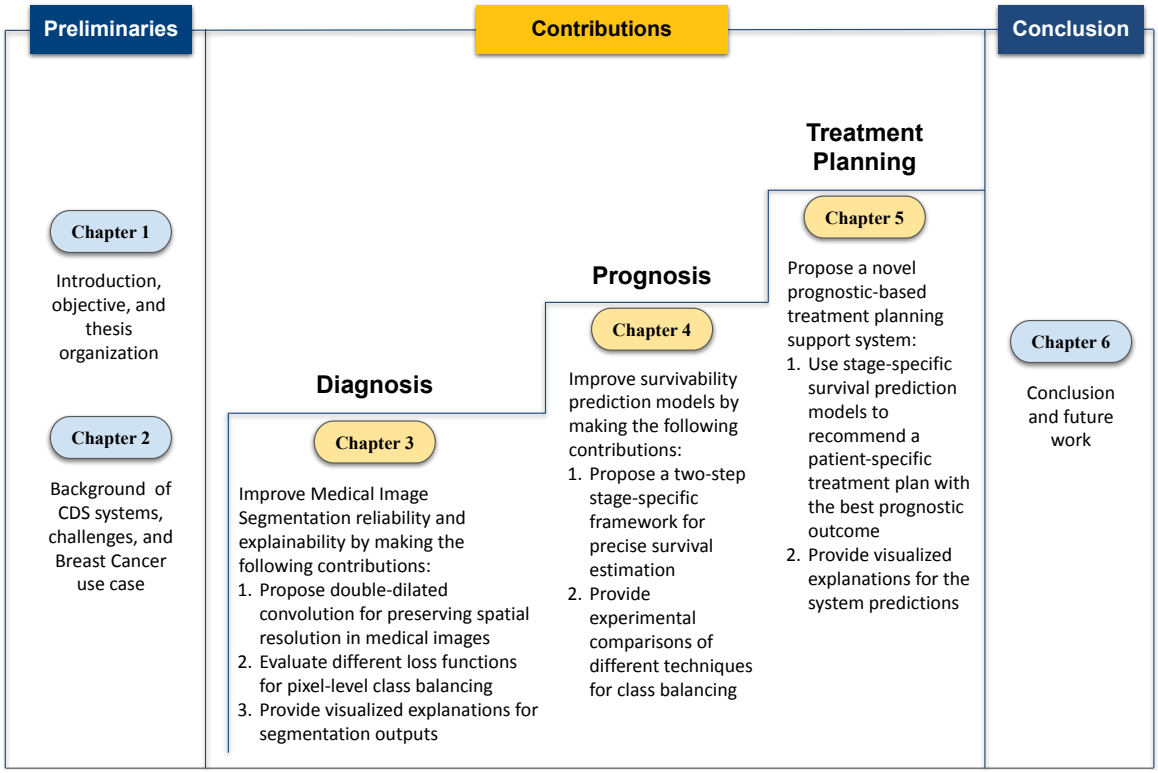


Figure 1.1: Organization of all the chapters of the thesis.

Chapter 3 aims to improve the reliability and transparency of CAD systems. We consider three research problems existing in medical image segmentation, which is a crucial task in CAD systems that helps identify regions of interest in medical images.

- First, to preserve local spatial resolution in medical images, we propose a double-dilated convolution module that perceives complex kernels with denser cores in order to eliminate the problem of decaying local resolutions, which occurs when applying existing CNN-based segmentation architectures. With the use of the state-of-art Deeplabv3+ network, we explain our simple implementation of the double-dilated convolution, which uses two dilation factors in parallel to replace the traditional dilated convolution layer used in the original network. To evaluate our proposed convolution method, we perform our analysis on the publicly available mammogram screenings provided by the INBreast dataset. The experimental results demonstrate the effectiveness of the proposed module in terms of both Dice similarity and Miss Detection rate when applied to the mass segmentation problem in mammograms.

- Second, to solve the pixel-level class imbalance problem existing in medical images, we compare using four widely-used loss functions in training our network to determine the best-suited method for this medical image segmentation. The results promote

employing the Tversky loss function as a balancing remedy in medical image segmentation.

- Finally, to overcome the lack of explainability in existing medical image segmentation networks, we apply explainability techniques that provide interpretable segmentation results. We quantitatively compare the performance of different explainability methods in terms of complexity and truthfulness. The results show the efficiency of Grad-CAM in visually explaining segmentation results.

In **Chapter 4**, we develop a computer-aided prognostic model for cancer survivability prediction. We propose a new framework to address the disease survivability prediction problem for patients suffering from multi-stage conditions such as breast cancer (our use case). Our approach to predicting breast cancer survivability with Machine Learning combines two key strategies. Firstly, we approach the task as a two-step problem, consisting of a classification task to predict whether a patient will survive for five years or not, and a regression task to estimate the remaining months of those predicted not to survive for five years. Secondly, we train stage-specific models for each cancer stage, rather than using all stages together, to predict survivability for patients within the same stage. We evaluate the impact of these strategies, along with different balancing techniques, on model performance using the SEER dataset from the National Cancer Institute. The results showed that our two-step stage-specific system improved the performance of survival estimation for breast cancer patients. Additionally, evaluating results for each summary stage separately highlighted performance differences between stages, demonstrating the importance of addressing survivability for each stage individually.

Chapter 5 discusses the development of an AI-enabled prognostic-oriented treatment planning system. The importance of prognosis in determining effective treatment plans has not been fully addressed in clinical decision support systems. To address this issue, this chapter proposes a novel survival-based treatment planning system to provide patient-specific treatment recommendations based on the estimated prognostic outcomes. The proposed system provides visualized explanations by employing the feature importance and the decision path followed to make the decision to ensure the transparency of the prediction model. By employing the SEER Plus dataset that provides treatment information for the female patients' breast-cancer incidence data in the US, we first re-compare different machine learning-based frameworks developed to predict the survivability of breast cancer patients after including information about the treatment history of the patient in order to ensure the superiority of the two-step stage-specific prediction. Then, we evaluate the proposed system for treatment planning that receives different combinations of possible options coupled with patients' clinical data to predict a ranked list of recommended plans based on the predicted survival outcome of each plan. The results show promising prediction

accuracy for the novel explainable survival-based system. Our analysis shows that both the disease stage and prognostic outcome are highly correlated with the treatment plan recommended by medical practitioners for a specific patient, which confirms the need of developing a stage-specific prognostic-based treatment planning framework for treatment recommendation.

Lastly, in **Chapter 6**, we summarize and conclude the thesis and put forward some future research directions.

Chapter 2

Background

In this chapter, we give an overview of clinical decision systems and highlight the motivation behind this research. Then, we describe different components of the system (i.e., diagnosis, prognosis and treatment planning) and discuss the existing challenges in each component along with the methodologies used in literature to address them.

2.1 Clinical Decision Support Systems

2.1.1 Overview

The development of the clinical decision support (CDS) system aims to improve healthcare delivery by incorporating health-related data such as clinical knowledge and patient information into medical decision-making [12]. A CDS system is typically a software tool that assists medical practitioners in their decision-making process by matching the patient's characteristics with a computerized clinical knowledge, and presenting patient-specific assessments or recommendations [3]. By combining clinicians expertise with computer-based

information and suggestions, these systems have been recognized for their potential to mitigate medical errors and enhance healthcare quality and efficiency [2].

Clinical decision support systems can be classified into two main categories: knowledge-based and non-knowledge-based, as shown in Figure 2.1. Knowledge-based systems (KB-CDSS) rely on human-based medical knowledge and rules to provide decision support. This is created by subject matter experts who use rules (IF-THEN statements) to program the system with guidelines, clinical pathways, and algorithms. When a clinician inputs patient data, the system uses these rules to make suggestions or provide recommendations based on the data input. Many of the earliest systems of this type were diagnostic decision support systems, where the system provided information to help the user make the diagnosis instead of coming up with the answer. One feature of these systems is that they included high interaction between the medical practitioner and the system as the user was expected to be active and to interact with the system, rather than just be a passive recipient of the output [13]. Non-knowledge-based CDS systems, on the other hand, only rely on a data source and employ computational methods to make data-driven decisions instead of following programmed medical knowledge. These computational methods include statistical approaches and Artificial Intelligence (AI) or Machine Learning approaches [3]. In this thesis, we focus on the Machine Learning-based CDS systems, as they proved to improve the overall accuracy of clinical decision-making, as demonstrated in previous studies [5].

Machine Learning (ML) involves the development of computational models that enable computer systems to improve their performance on a specific task by learning from data without being explicitly programmed. Traditional machine learning algorithms typically rely on a set of hand-crafted features that represent the patient's data. These features are fed into a learning algorithm, such as a Decision Tree [14], a Random Forest [15] or a Support Vector Machine [16], which learns to map the features to the target output. On the other hand, Deep Learning (DL) models learn to automatically extract relevant features from raw data without the need for hand-crafting. This is achieved through the use of Artificial Neural Networks (ANN) [17], which are composed of a large number of layers of interconnected neurons that can learn to represent increasingly complex features of the data.

In this thesis, we consider both traditional ML and DL algorithms in developing different components of our CDS system. Although DL algorithms have shown state-of-the-art performance in many tasks, traditional ML algorithms are still frequently used in a wide range of applications, especially where interpretability is important. Traditional machine learning models are considered more interpretable due to their reliance on direct feature engineering, which makes them easier for humans to understand [18]. In contrast, deep learning models involve the use of complex neural network architectures that can contain millions of parameters, which makes it challenging for humans to interpret the model's decision-

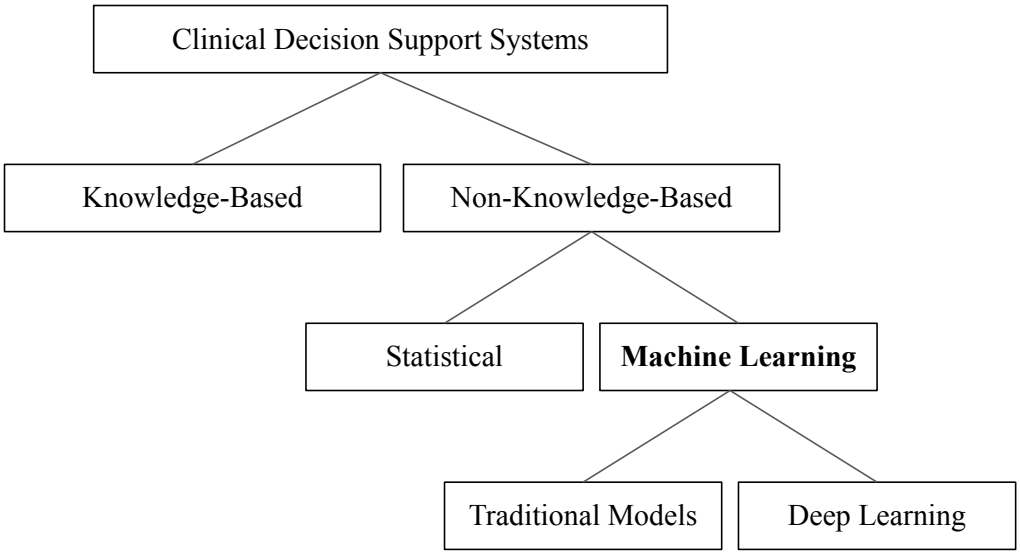


Figure 2.1: Different categories of clinical decision support systems.

making process. Therefore, traditional machine learning approaches can be preferred in healthcare-related applications that depend on tabular clinical data, such as cancer survivability prediction [19–22], where these models are able to achieve comparable (or sometimes better) performance than black-box DL models. However, this is not the case with applications that rely on medical imaging, where DL algorithms, in particular Convolutional Neural Networks (CNN), have rapidly become a methodology of choice for analyzing medical images [23]. This is because CNNs have been shown to outperform traditional ML models in image analysis by a large margin [24], and due to their ability to automatically learn relevant features from raw images. This is crucial in medical image analysis since it is unfeasible for humans to manually extract features from large numbers of medical images. Therefore, explainability methods are often needed with CNN-based models to overcome the accuracy-interpretability trade-off in CDS systems [18].

2.1.2 Challenges

With the increasing shortage of health specialists (e.g., radiologists, oncologists, etc.) around the world [25], developing a trustworthy data-driven CDS system for disease diagnosis and prognosis analysis can have a significant impact on healthcare professionals as well as patients. Such a system uses machine learning algorithms to analyze medical images and tabular clinical data to detect fatal diseases, estimate their survivability and aid in treatment planning. Despite significant research and development in the area of AI CDS systems, there is still a long way to go before these systems can be trustworthy and fully accepted

in clinical practice.

One main reason that hinders the adoption of AI-based CDS systems into clinical practice is the lack of transparency in these systems [26]. This means that the reasoning behind the system's recommendations is often unclear. This is a significant issue because physicians need to be able to understand the rationale behind the system's recommendations to make informed decisions for their patients. Additionally, if physicians don't understand how the system arrived at a particular recommendation, they may be less likely to trust the system and may only rely on their own judgment. The main reason for this problem is that many AI algorithms used in clinical decision systems are complex and difficult to interpret. These algorithms often rely on multiple layers of neural networks, which can make it challenging to identify which factors or variables are being considered in the decision-making process [27]. Furthermore, the data used to train these systems may also be complex and difficult to understand, making it hard to determine how the system has learned to make recommendations. In addition to this, the lack of reliability in current CDS systems also limits the integration of these systems into clinical practice [3]. This calls for revisiting the ML architectures and frameworks used in developing different components of the CDS system in order to further improve its accuracy and make it more reliable.

In this thesis, we try to address the trustworthiness problem of data-driven CDS systems from new perspectives with the objective of improving both the reliability and transparency of these systems in clinical practice. We experiment with different methods and frameworks that can potentially enhance the system's performance and increase reliability. In addition, we focus on inherently improving the transparency of different components of the system while incorporating explainable AI approaches. In order to improve the transparency within the system, we adopt the concepts derived from the work done in [28], which studied the transparency problem in medical AI from a multidisciplinary view, including technological, medical and patient perspectives. We summarize the general concepts followed throughout this thesis while developing different components of the AI-enabled CDS system as follows: 1) Promote inherent explainability by using traditional ML with tabular data. 2) Provide feature importance on two levels: model level and decision level. 3) Promote visualized outputs over black-box decisions. 4) Predict outcomes of different treatment options for patient-centred care.

To carefully examine the existing reliability and explainability challenges and find practical solutions for them, we break the CDS system down into three main components: a Computer-Aided Diagnosis (CAD) system, a Computer-Aided Prognosis (CAP) system and a Computer-Aided Treatment Planning system (CATP). In the following subsections, we analyze each component of the AI CDS system, survey the literature for the existing methodology, and identify the possible solutions that can improve its trustworthiness.

2.1.2.1 Computer-Aided Diagnosis (CAD)

Diagnosis is considered the first and most important component of the clinical decision-making process [29]. Computer-aided diagnosis (CAD) is a system that uses computer algorithms to assist clinicians in interpreting medical images and other diagnostic data in order to identify a disease. An AI-based CAD system uses Machine learning and Deep learning techniques to help diagnose various medical conditions, such as tumors, fractures, and lesions, by recommending potential diagnoses or highlighting areas of interest for further evaluation [5]. These systems are typically used in conjunction with medical imaging modalities like X-rays, CT scans, and MRI scans [5]. It has been shown in the literature how CAD systems can improve diagnostic accuracy and reduce the likelihood of errors, particularly in cases where the medical images are complex or difficult to interpret [30–32].

When we look at the research done in this area, we can identify two main tasks considered in the development of CAD systems: medical image classification and medical image segmentation. The classification task refers to assigning a label or a category to a medical image based on its content. In other words, classification models aim to provide ready-to-use diagnostic decisions for the medical team. Despite the high accuracy these classifiers reached in recent years, they still face resistance by many clinicians, i.e., radiologists and oncologists, due to the black-box nature of the state-of-the-art deep learning models behind them [26]. In addition to the lack of interpretability, some medical professionals have shown concerns that automated diagnostic classification systems may replace their jobs or reduce the demand for their services due to their ability to independently make diagnostic decisions [11]. On the other hand, the medical image segmentation task aims to partition a medical image into multiple regions or segments that represent different anatomical structures or potential lesions within the image. In this thesis, we dedicate all our focus to studying the medical image segmentation task in CAD systems by identifying the existing challenges and proposing new solutions. We argue that developing a reliable and explainable AI-based segmentation model that is able to accurately identify and partition suspicious regions in medical screening can be of significant importance for the following reasons:

- Visualization of the decision-making process: Medical image segmentation can help to visualize the decision-making process of AI clinical decision systems. By segmenting an image into different regions, the system can provide physicians with a visual representation of how it arrived at its recommendations. This can help physicians to better understand the rationale behind the system's recommendations and make more informed decisions.

- Integration with medical knowledge: Medical image segmentation can assure medical professionals that AI clinical decision systems are not intended to replace them as they only provide information that can complement medical practitioners' knowledge

to make the final decisions with higher accuracy in less time. This is crucial for the successful adoption and integration of these systems into clinical practice [4].

- Identification of errors: Medical image segmentation can also help to identify errors in AI clinical decision systems. If the system makes a recommendation that does not align with the segmented regions of an image, it may indicate an error in the system's decision-making process. This can help physicians to identify and correct errors in the system, improving the overall accuracy and effectiveness of the system.

- Improved communication with patients: Medical image segmentation can also help to improve communication with patients. By providing a visual representation of the clinical decision, patients can better understand the reasoning behind the recommended course of action. This can help to build trust between patients and physicians and improve patient outcomes.

Having stated the essential role that medical image segmentation plays in developing a transparent diagnostic model, it is important to mention that this task is considered one of the most challenging tasks in medical image analysis [33]. Medical images, such as CT or MRI scans, often depict complex anatomical structures that can vary significantly between individuals. The segmentation task is more difficult when the boundaries between different tissues or organs are not clearly defined, and when the appearance of the tissue changes due to factors such as inflammation or disease. We surveyed the literature for existing segmentation frameworks and identified three challenges that we try to address in our work:

1. Most of the work done in the area of medical image segmentation employed convolutional neural network (CNN) architectures that were originally developed for image classification. However, the authors in [34] raised the question of the suitability of these architectures for dense prediction problems such as the medical image segmentation. They argued that successive subsampling layers that reduce input resolution until a global prediction is obtained in classification networks can harm the performance of image segmentation. This is because the nature of dense prediction problems calls for multi-scale contextual reasoning combined with full-resolution output, which is not maintained in a typical CNN architecture. Hence, they emphasized the need for dedicated modules designed specifically to preform this type of tasks. Some solutions were proposed to solve this reduced resolution problem in image segmentation, including multi-scale piecewise training [35] and dilated convolution [34]. Although these methods managed to reduce the loss of resolution compared to traditional CNN networks, there is still room for improvement to further maintain the input image spatial resolution, which can be crucial in medical image segmentation problems. In chapter

3, we study the existing methods and propose improvements on the dilated convolution module to preserve local resolution and achieve better segmentation performance for medical images.

2. Pixel-level class imbalance is another common challenge in medical image segmentation, where the number of pixels belonging to different classes of interest (e.g., tumor and non-tumor) is highly imbalanced. This can lead to poor segmentation performance, as the model may learn to assign the majority class label to all pixels to optimize the objective function [36]. To address this issue, some proposed using an optimized batch size to include a balanced number of pixels from the majority and minority classes [37]. Another method employed sampled loss training where the loss is only calculated for some random pixels instead of the entire image [38]. However, the most popular solution for this issue is sample re-weighting, where a higher weight is given to pixels from the minority class during training. This can be controlled by the choice of the objective function used to calculate the loss during training. Many loss functions have been proposed for this purpose; however, it remains unclear which one achieves the best performance. Hence, while developing our medical segmentation model in chapter 3, we perform an experimental comparison between the most popular loss functions used in literature for medical image segmentation to identify the best performing one for this problem.

3. The lack of explainability is one of the main obstacles that hinder CAD adoption in clinical practice. While the explanation of disease classification networks have received attention in previous studies [39], little effort has been dedicated to improving the explainability of segmentation networks, which also use black-box architectures. Post-model explanatory analysis of segmentation networks can also aid in detecting overfitting and learning relevant features, leading to more robust performance. We address the lack of segmentation explainability in chapter 3 by evaluating the effectiveness of explainable AI models in medical image segmentation. Our proposed segmentation network utilizes explainable techniques to ensure transparent and trustworthy systems that can be integrated into medical practice.

2.1.2.2 Computer-Aided Prognosis (CAP)

Prognosis refers to the likely outcome of a medical condition or disease based on factors such as the patient's medical history, disease stage, and test results. A prognosis is often expressed in terms of the likelihood of survival over discrete time periods, but it can also be measured by the expected duration of the disease or the degree of recovery that is likely to be achieved. The term prognosis is used in this work to refer to survivability prediction since it is the most common indicator used in prognostic analysis [40]. In the process of

medical decision-making, once a diagnosis of a fatal disease is made, the prognosis can help guide treatment decisions [6] and inform patients and their families about the potential course of their condition. Computer-aided prognosis (CAP) is a relatively new field that builds upon the foundation of CAD by involving the use of computerized algorithms to help physicians predict disease outcomes and patient survival. Similarly, Machine learning and deep learning techniques have been applied to predict disease outcomes and patient survival. While the majority of research in the field of computer-aided medical decision-making has been focused on the CAD component, there has been some work done in the area of computer-aided prognosis for Idiopathic Pulmonary Fibrosis [41], Neuroblastoma [42], Prostat Cancer [7] and Breast Cancer [7, 43].

After reviewing previous CAP systems, we identified three structural shortcomings/research gaps in the current systems that can contribute to the lack of reliability in CAP frameworks:

1. Lack of stage specificty: In the case of fatal diseases with multiple stages, AI prognostic models are usually developed for patients from all disease stages together, which means that incidences from all stages are modeled together. This stage-agnostic modeling raises concerns that this can harm the performance of the system since different stages of fatal disease can largely vary in their prognostic patterns [44]. For example, in the case of breast cancer, the five-year survival rate for women diagnosed with localized-stage breast cancer was as high as 99%, whereas the rate was only 29% for those with distant stages during the same time period [45]. Hence, modeling all stages together can lead to inaccurate predictions and unreliable results, which contribute to the lack of overall trustworthiness. Moreover, this approach can make the predicted prognostic results ambiguous and difficult to interpret by medical practitioners since the important features that determine the outcomes of a disease can be different depending on the stage at which it was discovered [46]. Therefore, more work needs to be done to investigate the viability and effictiveness of prognostic stage-specific modeling of multi-stage diseases. This is studied in chapter 4.

2. Lack of prediction specificity: Most studies conducted for survivability prediction of fatal diseases aim to only perform survivability classification rather that regression. For example, the majority of cancer prognostic models focus on predicting whether or not a patient will survive for five years. This may not be sufficient for medical decisions, where precise estimation of patients' survival time plays a significant role in deciding treatment recommendations and personalized medicine [29]. To illustrate, a 5-year survival classifier can predict the same label for a patient who is likely to survive for only 1 month as well as the one who is predicted to survive for 59 months, as they both are predicted to not survive for 5 years. This lack of precise information can make the prognostic decisions hard to understand and rely on. Hence, providing

detailed predictions in a clinical decision system is important to improve CAP systems' reliability, improve accuracy and promote personalized care for the patients.

2.1.2.3 Computer-Aided Treatment Planning (CATP)

Treatment planning is a crucial step in the process of providing effective healthcare services to patients. It involves developing a plan of care that is typically done collaboratively between the healthcare provider and the patient, taking into consideration the patient's medical history, current health status, and the predicted outcome of the diagnosed disease. Computer-aided treatment planning (CATP) refers to employing computational methods to assist medical practitioners in determining the most appropriate treatment plan for a patient. The development of CATP systems is still a relatively new research area that has not received as much attention as CAD and CAP systems [8].

1. Lack of prognostic-oriented treatment planning: Although the role that prognosis can play in treatment recommendation has been greatly emphasized in medical research [6, 29], very little attention was given to the development of prognostic-based treatment planning models in CDS systems [47]. Prognosis can be an important indicator in making decisions regarding treatment plans, such as whether to pursue more aggressive therapies or focus on palliative care. This information can help clinicians give more personalized care to their patients while providing realistic expectations of the possible treatment options to patients and their families. Therefore, it is important to develop treatment recommendation systems where survival prediction is used to suggest pathways for treatment. In chapter 5, we propose a new framework for survival-based treatment planning. To ensure detailed and intuitive recommendations, we provide a list of all possible combinations of treatments associated with their survival prediction, instead of providing just one recommended treatment for the medical professional.

2. Lack of comprehensive treatment planning: Although many therapy approaches can be considered while deciding on a treatment plan (e.g. surgery, chemotherapy, radiation, etc.), most of the previous studies only considered one type of therapy in their CATP systems and developed models to optimize the parameters of this treatment method (e.g. radiotherapy parameters) [48–51]. Hence, it is important to develop treatment recommendation systems that account for all possible combinations of treatment options and recommend the best-suited one for patient-centred care.

3. Lack of transparent treatment planning: Many proposed models for treatment outcome prediction in cancer care have utilized black-box machine learning architectures [48, 50], which can make it difficult for oncologists and healthcare providers to un-

derstand and trust the automated decisions. Therefore, there is a growing need to develop more explainable models for treatment outcome prediction that can help to address the transparency concerns and improve the integration of CATP systems in clinical practice.

2.2 Breast Cancer Use Case

In this thesis, we propose new methods and frameworks to address existing challenges and research gaps in the three components of the CDS system to pave the way toward its usage in clinical practice. Although the proposed methods have the potential to be applied to many fatal diseases, we select the breast cancer use case to perform our analysis and develop different components of our CDS system for the following reasons:

1. Breast cancer can be the best-suited option to evaluate our double-dilated CNN-based segmentation network for preserving local resolution in medical images. This is because breast cancer screenings are known to be highly heterogeneous, meaning that they often include large areas of dense fibrous tissue, glandular tissue and fatty tissue [52]. As reported by the National Cancer Institute (NCI) in the US, around 40% of women have heterogeneously dense breast tissues, which makes it harder to find small masses in the breast tissue on a mammogram [53]. Improving the segmentation network architecture to maintain local spatial information of the mammogram images can help reduce the miss-detection rates by potentially identifying small masses in dense breast tissues, which is why we perform our analysis in Chapter 3 on mammogram screenings. This can serve as a proof-of-concept to show the effectiveness of the proposed idea so that it can be adopted in other medical image segmentation tasks where preserving local resolution can also be crucial.

2. Breast cancer is an ideal candidate to evaluate our proposed two-step stage-specific survival prediction framework for multi-stage diseases. This is because the stage of breast cancer is known to have a significant impact on a patient's survivability outcomes. For example, between 2011 and 2017, the five-year survival rate for women diagnosed with localized-stage breast cancer was 99%, whereas the rate was only 29% for those with distant stages [45]. In addition, breast cancer has a well-defined staging system that considers the size and location of the tumor, the extent of lymph node involvement, and the presence of metastasis to other parts of the body [54], which can benefit the development of our proposed survival prediction system in Chapter 4. Our proposed system can be potentially applied to other multi-stage diseases to improve the accuracy of survivability estimation.

3. Breast cancer is also an optimal choice for developing a computer-aided treatment recommendation system because of the availability of large and diverse electronic health records for breast cancer patients. For example, the last release of the Surveillance, Epidemiology, and End Results (SEER) database included 1,425,552 breast cancer incidences, which accounts for the largest number of records belonging to one cancer type (15.2% of all new cancer cases from 28 different types of cancer) [54]. Moreover, it has been established that the stage of breast cancer greatly affects the treatment options, which makes it a suitable use case for our stage-specific treatment planning system proposed in Chapter 5. However, our survival-based treatment planning system can also be adopted in other multi-stage diseases to recommend patient-specific treatment plans based on the projected survivability.

4. In addition to the above-mentioned reasons, breast cancer is the most common cancer in women worldwide, accounting for 25% of all cancer cases in women [55]. According to the World Health Organization (WHO), there were 2.3 million new cases of breast cancer and 685,000 deaths from breast cancer globally in 2020, corresponding to 16% or 1 in every six cancer deaths in women [56]. These rates can be reduced by improving the reliability of current CDS systems using new solutions to existing challenges in different components of the system, as illustrated in the next chapters.

To understand the workflow of this thesis, in the next few lines, we provide an overview of breast cancer diagnosis, prognosis and treatment planning. First, breast cancer can be diagnosed through several screening methods, including mammography, ultrasound, magnetic resonance imaging (MRI), and positron emission tomography (PET) may also be used, but they are less common. The most effective approach for the early detection of breast cancer is currently considered to be X-ray mammography [57]. In many countries, asymptomatic women are encouraged to undergo annual mammographic examinations to detect clinically unsuspected lesions in the breast. Mammography involves capturing two images of each breast: the craniocaudal (CC) view, which is a top-to-bottom view, and the mediolateral oblique (MLO) view, which is a side view. These images can be obtained using x-ray film, such as a film-screen mammogram, or in digital format using full-field digital mammography (FFDM) [58]. Radiologists look for suspicious lesions in the images, such as masses and calcifications, in order to diagnose breast cancer. A breast biopsy, which involves removing a small tissue sample for examination under a microscope, is often needed to confirm a positive diagnosis that was made based on mammogram screenings. Early detection is critical in the successful treatment of breast cancer. When breast cancer is detected in its early stages, treatment options tend to be more effective and less invasive, leading to higher chances of survival and better quality of life for the patient [59]. For this reason, CAD systems have emerged as a promising tool to help with the early detection of breast cancer.

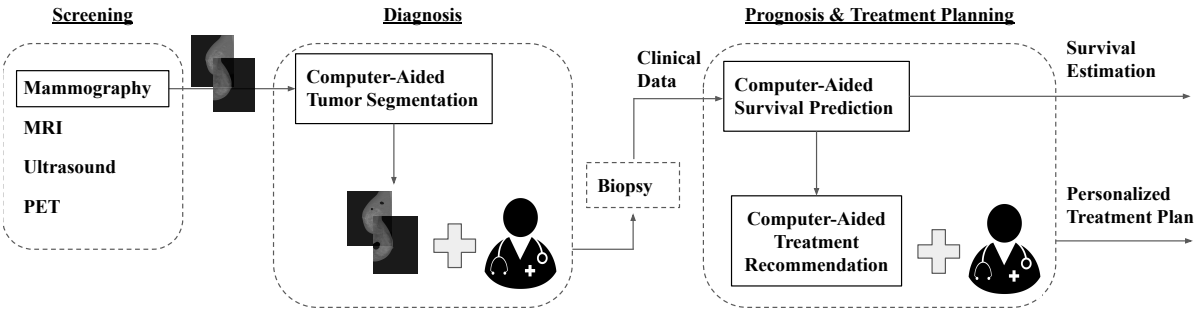


Figure 2.2: Breast Cancer Clinical Decision-Making Pipeline.

Although many factors can affect a breast-cancer patient's prognosis, including age and tumor characteristics, the survival rates for breast cancer greatly vary depending on the stage of the disease at the time of diagnosis. According to the American Cancer Society, between 2011 and 2017, the five-year survival rate for women diagnosed with localized-stage breast cancer was as high as 99%, whereas the rates were only 86% and 29% for those diagnosed at regional and distant stages, respectively [45]. Survival prediction for breast cancer patients is widely acknowledged as an important aspect of clinical decision-making. It helps oncologists and patients to make informed decisions about the most appropriate treatment plan based on the individual patient's prognosis. Treatment options for breast cancer depend on the type, stage and projected outcome of the disease and may include surgery, radiation therapy, chemotherapy, hormone therapy, and targeted therapy [60]. For this reason, CAP systems have arisen as a potential tool for forecasting the survivability of breast cancer patients and providing personalized treatment recommendations.

In this thesis, we develop computer-aided breast cancer CDS systems for diagnosis, survivability prediction and treatment planning using AI technologies. Figure 2.2 depicts an overview of the pipeline of the breast cancer decision-making system from screening to treatment planning while showing how our CDS is integrated with medical knowledge in this paradigm. First, out of different breast screening modalities, the 2-D mammogram screening is used in our analysis as it is the most common method used for early breast cancer diagnosis. Then, AI-based medical image segmentation is applied to identify potential masses in the input mammogram images. Our proposed semantic segmentation system is studied in detail in chapter 3. Next, the segmented images are examined by medical professionals to make the final diagnosis. If a positive diagnosis is made, a biopsy is performed to verify the diagnosis and verify the diagnosis and collect additional features of the disease. These features are then passed to the computer-aided survival prediction system developed in 4 to estimate the remaining survival time for a patient. These prognostic models are incorporated with possible treatment options used with previous patients to create a survival-based treatment planning system that predicts recommended treatment plans for

a specific patient, as proposed in chapter 5.

2.3 Integration With Healthcare Systems

In recent years, many healthcare providers have shifted from paper-based records and manual processes to Electronic Health Record (EHR) systems. EHR is a software technology that provides an electronic version of a patient's medical history that includes information such as diagnoses, medications, lab results, and other clinical data. It is designed to improve patient care by providing healthcare professionals with easy access to comprehensive patient data. There are many different EHR software options available in today's market, each with its own features and capabilities. Some of the most commonly used EHR software in Canada include Meditech, Epic, and Cerner [61]. These software solutions are used by hospitals, diagnostic labs, and other healthcare organizations to manage patient data and streamline clinical workflows.

Recently in April 2023, Microsoft and Epic announced that they will integrate the Microsoft Azure Open-AI Service with Epic's EHR platform. This integration aims to extend natural language queries and interactive data analysis to Epic's self-service reporting tool [62]. This announcement confirms the increasing interest in integrating AI tools into current healthcare systems, which suggests the need to address the trustworthiness challenges in previous AI-based Clinical Decision Support systems. Our CDS system proposes new solutions and frameworks that improve the reliability of the CDS system by preserving local resolution in medical image segmentation, ensuring stage and prediction specificity in survival prediction, and providing prognostic-based treatment planning. It also addresses the lack of transparency in different system components by providing visualized explanations for the system's automated decisions at each step.

To integrate the AI-based CDS system into current healthcare systems, we can think of two different approaches. The first approach is to adopt the proposed machine-learning models into the EHR systems by the EHR software companies, which often requires collaboration between these companies and the CDS system developers in order to implement these new technologies in their EHR systems. The advantage of this method is that the EHR system can locally use different components of the CDS system without the need to transfer patients' data over cloud connections with a remote system. Also, using this approach allows ML models to be trained on local datasets in each EHR system, which can enable the model to learn trends that are correlated to demographic factors [63]. However, the process of developing software-specific CDS tools can be time-consuming and requires many customized implementations of the CDS system to match the software requirements and frameworks used by each EHR software company.

The second approach is presented in Figure 2.3. This approach suggests developing

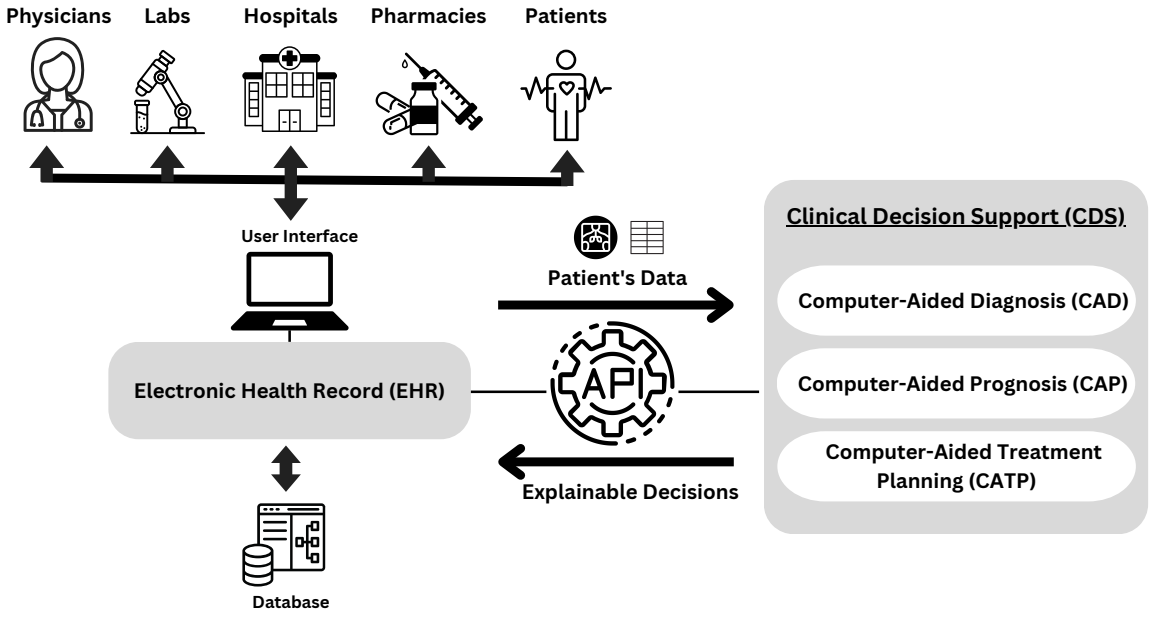


Figure 2.3: An integration paradigm for the proposed CDS system into current healthcare systems using an Application Programming Interface (API).

a generic Application Programming Interface (API) that can facilitate communication between the CDS system and current EHR systems. The API can have a standardized protocol that regulates receiving requests from any EHR software. The request includes the patient's medical information needed for the analysis, such as mammogram screenings and pathological data, and the API passes it to the requested component of the CDS system (i.e., CAD, CAP or CATP). Then, after the CDS makes a patient-specific prediction, the API returns a response to the EHR system, including the predicted decision, along with visualized explanations. In this approach, the CDS system can either be hosted on the cloud or locally in a private network of a healthcare organization. The local CDS system can ensure patients' privacy, whereas the cloud CDS systems can provide more generalized models that employ datasets from different sources in training. One solution to eliminate this trade-off is to incorporate Federated Learning techniques in CDS systems to allow distributed model training where each client can keep their data private and only share model updates [64]. This is one of our future research directions to enable a private and efficient integration of our CDS system into existing healthcare systems. Finally, although this paradigm can potentially enable seamless integration with different EHR systems, a collaboration with at least one of the EHR software providers is still needed to develop and test the proposed interface.

Chapter 3

Multi-dilation Convolution For Preserving Spatial Resolution in Medical Image Segmentation: Mammogram Use-Case

Medical image segmentation is a critical task in computer-aided diagnosis (CAD) systems that helps identify regions of interest in medical images. In this type of problem, preserving the input image resolution plays a crucial role in achieving good performance. The introduction of the dilated convolution module contributed to maintaining resolution across layers of a deep convolutional neural network by exponentially increasing the receptive field with a linear increase of parameters. One pitfall of dilated convolution is losing local spatial resolution by increasing the sparsity of the kernel in checkboard patterns. In this work, a double-dilated convolution module is proposed in order to preserve local spatial resolution in medical images while having a large receptive field in segmentation networks. The proposed module is applied to the tumor segmentation task in breast cancer mammograms as a proof-of-concept. In addition, the problem of pixel-level class imbalance problem in mammogram screenings is tackled by comparing different loss functions (i.e., binary cross-entropy, weighted cross-entropy, dice loss, and Tversky loss) to identify the best-performing function for the mass segmentation task. Finally, we address the black-box nature of the developed models by quantitatively evaluating our adopted Gradient weighted Class Activation Map (Grad-Cam) with other explainable models available for image segmentation. Experimental analysis is performed to compare the performance of lesion segmentation networks on mammogram screenings from the INBreast dataset [59] before and after plugging the proposed dilation module into one state-of-the-art deep convolutional neural network. The obtained results show the effectiveness of the proposed module in terms of both the

Dice similarity and the Miss Detection rate when applied to the mass segmentation problem.

3.1 Introduction

With the emergence and growing popularity of Convolutional Neural Networks (CNNs) in recent years, researchers have increasingly invested in the development of Computer-Aided Diagnosis (CAD) systems. Using CAD systems, radiologists can integrate their knowledge with computer output to make more accurate and timely diagnoses. These systems are typically used in conjunction with medical imaging modalities like X-rays, CT scans, and MRI scans [5]. Medical image segmentation is a crucial task performed in most CAD systems where a medical image is partitioned into multiple regions or segments, each of which corresponds to a specific anatomical or pathological structure. For example, in cancer CAD systems, tumor segmentation is a specific application of medical image segmentation that involves identifying and delineating the boundaries of a tumor within medical images. This can help to quantify various characteristics of a tumor, such as its size, shape, and volume. This information can be used to make the correct diagnosis, track changes in the tumor over time and evaluate the effectiveness of treatment.

Although its unquestionable impact on enhancing healthcare systems, medical image segmentation is still considered one of the most challenging tasks in computerized health analytics [33]. Some of the associated challenges are related to the task itself, as medical images often depict complex anatomical structures that can vary significantly between individuals. The segmentation task is more difficult when the boundaries between different tissues or organs are not clearly defined or when the appearance of the tissue changes due to factors such as inflammation or disease. Other challenges are related to the learning dynamics and lack of explainability of the Deep Learning (DL) models that are commonly used to perform this task. Similar to other CAD systems, a typical medical image segmentation framework consists of three participants that usually integrate to produce the final diagnostic output: a *model* learns from imaging *data* to guide *human* healthcare providers. In this work, we address three existing challenges in DL-based medical image segmentation systems that either exist in the underlying architecture of DL models or arise due to the integration of these models with data or human components of the system, as sketched in Figure 3.1. Specifically, these challenges are the resolution loss in CNN, the pixel-level class imbalance in medical image segmentation and the lack of explainability of DL models.

3.1.1 Resolution Loss in CNN

Image classification and semantic segmentation are two examples that show the promising performance of CNN-based architectures. Image classification aims to perform a sample-wise classification, while segmentation performs a pixel-wise classification to identify regions of interest. However, architectures originally designed to solve the first task are usually repurposed to solve the latter.

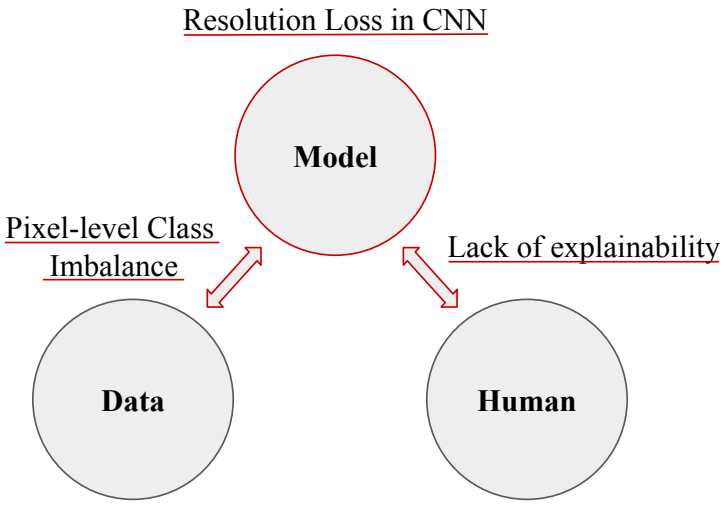


Figure 3.1: Challenges in medical image segmentation addressed in this paper.

Image classification networks use a sequence of pooling layers to increase the receptive field throughout the network and aggregate contextual information [65]. Adopting this approach in semantic segmentation networks [66, 67] came with the cost of losing resolution since the pyramid-shaped architecture ends with a global prediction in classification problems while semantic segmentation networks do not. Therefore, they use up-sampling layers to reconstruct pre-pooling resolutions. However, the authors in [34] raised the question of the suitability of pyramid-shaped architectures in dense prediction problems such as image segmentation. They argued that successive pooling layers that reduce input resolution in classification networks could harm the performance of image segmentation. This is because the nature of dense prediction problems calls for multi-scale contextual reasoning combined with full-resolution output, which is not maintained in a typical CNN architecture. Hence, they emphasized the need for dedicated models designed specifically to perform this type of task.

Atrous (or dilated) convolution, originally developed for the efficient computation of wavelet transform [68], was re-introduced as the dilated convolution module in [34] to be used in dense prediction models to partially solve this fading resolution problem by introducing an exponentially growing kernel's receptive field which reduced the need for pooling layers. However, for the dilated convolution to achieve this, sparsity in the kernel grows exponentially as well, which greatly harms local spatial resolution [69].

The complex nature of medical images calls for the need to maintain local spatial resolution in images while performing medical image segmentation. This work addresses this problem by proposing a modification to the existing dilated convolution module to have more control over the resolution of the kernel while still exponentially growing the receptive field. The proposed dilated convolution module has a multi-scale dilation parameter to

control inner and outer kernels. Figure 3.2 shows examples of different shapes of kernels performing dilated convolution, including the double-resolution dilation that can only be achieved using our proposed module. This example indicates the capability of our module to perceive novel and complex kernel shapes that give more control to enlarge the receptive field while eliminating the "gridding" problem [69], which occurs due to the increased sparsity of the kernels with large dilation factors.

3.1.2 Pixel-Level Class Imbalance

Pixel-level class imbalance is another common challenge in medical image segmentation, where the number of pixels belonging to different classes of interest (e.g., tumor and non-tumor) is highly imbalanced. For example, in the mass segmentation task, the number of pixels labeled as "mass" in an organ-specific screening is usually significantly lower than the number of pixels labeled as "normal". This large difference in size of classes to be segmented can negatively affect the Deep Learning model performance when it is integrated with medical imaging data [33].

The most popular solution for this issue is sample re-weighting, where a higher weight is given to pixels from the minority class during training [33]. This can be controlled by the choice of the objective function used to calculate the loss during training. Many loss functions have been previously proposed and incorporated into deep learning models for this purpose [70–72]. However, it remains unclear which one achieves the best performance with the image segmentation of highly imbalanced medical images. Therefore, in this study, we perform an experimental comparison between the most popular loss functions used in the literature for medical image segmentation to identify the best-performing one for this task.

3.1.3 Lack of Explainability

In medical applications, the explainability of DL models is crucial, as radiologists and medical professionals need to be able to understand the reasoning behind the model's predictions and decisions. With medical image segmentation, although the model does not provide ready-made diagnoses as the case with classification tasks, it is still crucial to provide transparency and a human-like explainability of the model performance to promote why and how it provides these annotations. In other words, if physicians do not understand why the system arrived at a particular output, they may be less likely to trust the system and may only rely on their own judgment. Therefore, the lack of explainability is considered one of the main obstacles that hinder CAD adoption in clinical practice [10].

Recently, researchers have been working on developing techniques to improve the explainability of deep learning models in medical applications. Many of these models have

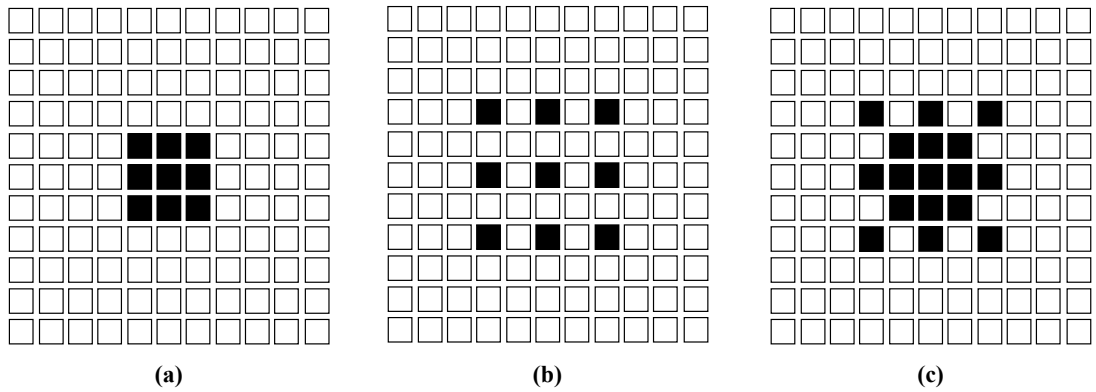


Figure 3.2: Our proposed dilation supports receptive field exponential expansion while retaining full resolution at the core of the kernel using multi-dilated convolution. (a) shows a 1-dilated 3x3 kernel. (b) shows a 2-dilated 3x3 kernel. (c) shows a combination of a 1-dilated 3x3 inner kernel and a 2-dilated 3x3 outer kernel. Both (a) and (b) kernels can be realized by traditional dilated convolution methods, whereas only our proposed method can realize the shape in (c).

been successfully adopted in disease classification networks, to justify the label predicted by black-box DL models [39]. However, very little attention has been given to improving the explainability of segmentation networks [73, 74] although they also suffer from black-box architectures. Moreover, post-model explanatory analysis can help computer science researchers uncover if the model is learning relevant features or if it is overfitting to the training images by learning spurious features. This allows us to adjust the model architecture and hyperparameters to achieve more robust performance that can be applied to real-life data [39]. Therefore, in this work, we address the segmentation explainability problem by quantitatively evaluating the performance of explainable techniques applied to medical image segmentation. We adopt explainable AI models in our proposed segmentation network to provide transparent systems that can be trusted and integrated into medical practice.

3.1.4 Mammogram Segmentation Use-Case

As for our use-case, the mammogram tumor segmentation problem is selected in this study. Breast cancer is the most commonly diagnosed cancer among female patients in the world and the second leading cancer-related cause of death [55]. For this reason, early detection and diagnosis of breast cancer is essential to decrease its associated mortality rate. The integration of Artificial Intelligence (AI) methods in AI systems plays a critical role in diminishing mammogram reading time for radiologists and improving their accuracy [75]. In

this study, we employ the publicly-available INBreast dataset [59] for mammogram screenings to evaluate the effectiveness of the proposed model in medical image segmentation. We aim to develop an explainable and efficient tumor segmentation model by 1) proposing a novel dilated convolution module to maintain the local spatial density of the input, 2) addressing the class-imbalance problem by employing the best-performing loss function for this task, and 3) adopting explainable AI models in image segmentation networks.

3.2 Related Work

In this section, we review some of the previous related work done in the area of medical image segmentation, including the different techniques used for class balancing and model explainability. Then, we summarize the recent progress of the dilated convolution and our proposed modification to this module.

3.2.1 Medical Image Segmentation

The process of identifying and partitioning the regions of interest (ROI) in medical images is significant in the diagnosis process. Therefore, many CAD solutions have been proposed in the literature for this task. Traditional algorithms included rule-based methods such as thresholding, boundary-based segmentation, region-based segmentation, and template matching, as reviewed in [76]. Deep learning-based techniques, especially the Fully Convolutional Neural Networks (FCNN) [77], have shown significant improvement in the image semantic segmentation task, compared to classical methods that require hand-crafted feature extraction [78–80]. This has made it possible to build large-scale trainable models that have the capacity to learn the optimal features which are required for segmentation.

In medical image processing, the anatomy of interest typically takes up a relatively small fraction of the overall image. Although these small anomalies are usually more significant, when a network is trained using such data, it frequently becomes biased toward the background as the model assigns the majority class label to all pixels to optimize the objective function [36]. To address this issue, some proposed using an optimized batching technique that includes a balanced number of pixels from the majority and minority classes in the training process [37]. The pitfall of this approach is losing geometrical information, which can be crucial in medical image segmentation. Another method employed sampled loss training where the loss is only calculated for some random pixels instead of the entire image [38]. The randomness of candidate selection for loss evaluation is the main disadvantage of this method, which limited its usage in existing image segmentation models [36].

On the other hand, sample re-weighting is known to be one of the most commonly-used remedies for class imbalance [81]. The idea is to give lesion pixels a higher weight when calculating the training loss. Authors in [82, 83] trained their models using the weighted

version of the cross-entropy loss function, which is largely used with classification tasks. Other works proposed using Region-based Loss functions such as the Dice loss [84] and the Tversky loss [72] in order to tackle the pixel-level class imbalance problem. However, it remains unclear whether or not there is a loss function that globally achieves the best performance in medical image segmentation tasks. Therefore, in order to overcome the pixel-level class imbalance in the mammogram images, we compare different loss functions, namely Binary Cross-Entropy, Weighted Cross-Entropy, Dice Loss, and Tversky Loss to identify the best-performing one for the mammogram tumor segmentation problem.

As for the explainability aspect, many explainable AI (XAI) models were proposed in the literature to overcome the black-box nature of deep neural networks. Many of these models, such as Local Interpretable Model Agnostic Explanations (LIME) [85], Deep Taylor Decomposition (DTD) [86] and Layer-wise Relevance Propagation (LRP) [87], were widely adopted in medical image classification tasks [26, 39, 88]. However, many of these methods implementations require a global classification layer at the end of the CNN network, which limits their applicability with pixel-level classification tasks such as medical image segmentation.

Although the vast majority of explainability work is focused on explaining CNN-based classification networks, we found few works that incorporated this important aspect in developing semantic segmentation networks. In [73], they employed the SHAP method to provide comprehensible explanations for oil slick segmentation models using coloured maps highlighting the input image areas that contributed to the model decision for a selected pixel or region. To provide explainable semantic segmentation for autonomous driving systems, the authors in [74] used the second-order derivative of neurons activations at the last encoding layer of their segmentation network to provide attention maps that visually explain the underlying network. In the area of medical image segmentation, we only found one work that provided explanations for segmenting tumors in liver CT images using activation maximization-based method [89]. To the best of our knowledge, our work is the first to address the explainability problem in mammogram tumor segmentation. We provide a qualitative assessment of the effectiveness of the adopted XAI techniques by evaluating their entropy, and the pixel-flipping graph similar to the work done in [88].

3.2.2 Dilated Convolution

Inspired by biological studies, a typical CNN adopts the pyramid-shaped structure where pooling layers succeed convlolutional layers to downsample the feature maps resulting from each convolution process. This has shown to be an efficient structure for processing digital images in many high-level computer vision applications such as face, object and digit recognition. However, applying the standard CNN architecture to the segmentation problems has inevitably resulted in a significant resolution reduction of the input image which can

affect segmentation performance, especially in medical applications [90].

Previous approaches addressed this problem by stacking a deconvolution network composed of deconvolution and up-pooling layers to reconstruct image resolution either from the encoded representation alone [91], or with combined respective scales in the convolution network [66]. Another approach creates multiple scaled versions of the input image, trains the network to predict the output for each, and uses attention to combine all outputs into one refined prediction [92]. However, it has been shown that the excessive adoption of these downsampling layers can be uncalled for in segmentation tasks in the first place [34]. Unlike high-level vision tasks where invariance of CNNs to local image transformations is advantageous, this invariance can hinder low-level tasks like semantic segmentation that require precise localization of spatial details along with a certain level of abstraction [93]. This motivated the introduction of dilated convolution [34].

In 2015, the dilated convolution module [34] proposed the use of a sparse kernel that grows exponentially to cover a bigger receptive field, reducing the need for pooling and downsampling layers. This was extrnsively used in several semantic segmentation models [69, 94]. However, the traditional dilated convolution framework has a fundamental issue referred to as "grinding" in [69]. Zeros are padded in a convolutional kernel at a fixed dilation rate, resulting in a receptive field that only covers an area with a checkerboard pattern. Consequently, only non-zero value locations are sampled, and neighboring information is lost. This problem becomes more severe as the dilation rate increases in higher layers where the convolutional kernel becomes too sparse to cover any local information because the non-zero weights are too far apart.

Some attempts were made to address the gridding problem. Authors in [69] proposed a hybrid dilation convolution where they stacked the standard dilated convolutions with different rates in a serial way. A similar technique was also adopted in many of the Deeplab family members, including Deeplabv2 [95], Deeplabv3 [1], Deeplabv3+ [96], and their variants achieving state-of-the-art results for the PASCAL VOC benchmark [97]. Although this approach achieved a better performance than traditional dilated convolution, it still limits the perceivable kernel shapes by only allowing checkboard patterns in dilated kernels. Recently, semi-dilated convolution [98] proposed a modified version of dilated convolution to better exploit the geometry of rectangular image (e.g. spectrograms and scalograms) by supporting exponential growth of the receptive field in only one dimension of the image.

In this work, we follow previous attempts and propose a modification to the dilated convolution to separate the dilation factor on the core of the kernel from the dilation factor on its edges while performing medical image segmentation tasks. Our contribution can be summarized as follows: (1) We introduce a simple implementation for a "double-dilated" convolution kernel that can assume more complex kernel shapes than previous dilated convolution approaches in order to have exploit local information in medical images. (2) We

integrate this new module in one state-of-art segmentation network with the appropriate modifications and evaluate its performance on the INBreast dataset [59] relative to using the traditional dilated convolution module.

3.3 Data Preparation

Among the available mammogram datasets, we select the INBreast dataset [59] for our analysis. There are many reasons for this. Unlike other public datasets that use digitized film-screen mammograms, the INBreast is the only publicly-available Full-field Digital Mammogram (FFDM) dataset. This enables it to have high-resolution images that are free from any inconsistency that may arise in the digitization process. In addition, this dataset provides radiologist-drawn pixel-level contours surrounding lesions, instead of providing only circles around ROIs as followed by most of the databases. This can be crucial in diagnosis since shape information is highly indicative of the malignancy of a mass [99]. The INBreast also provides images from both the craniocaudal (CC) view and the mediolateral oblique (MLO) view, which enables the development of generic CAD systems that are able to extract information from either view.

The database has 410 images in total, including both the CC and the MLO views, acquired between April 2008 and July 2010 using the MammoNovation Siemens FFDM aquisition equiment. The image size is either 3328×4084 or 2560×3328 pixels, depending on the compression plate used in the acquisition. Images contain different types of findings: normal, calcification, masses, asymmetries, multiple finding, and architectural distortions. In this study, we focus on the segmentation of masses, which are three-dimensional structures demonstrating convex outward borders, as defined by the Breast Imaging Reporting and Data System (BI-RADS). The number of images that included one or more mass lesions was 107 images.

As explained by the dataset documentation, the mammogram images were saved in the DICOM (Digital Imaging and Communications in Medicine) format whereas the annotations for all images were saved in the XML format. Each XML file includes the annotation information for all ROIs that are present in one image, with a list of contour points for each ROI using different tag names (e.g, Mass, Calcification, ..). Using Matlab, we prepared a script that reads an XML file, extracts annotation information of masses, and draws contours using Matlab's stroke() and imfill () functions. Then we save mask images in the PNG format as well as the original images format for visualization purposes. An example of the generated mask image and the corresponding DICOM image are shown in Figure 3.3. We load both the raw image files and the mask image files using Matlab's ImageDatastore and PixelLabelDatastore classes, respectively and resize all images to have a fixed-size input of 512x512 pixels. We use the 5-fold validation split to train and validate the models in all

our experiments.

Figure 3.3: A mammogram example from the INBreast dataset showing the craniocaudal (CC) view of a left breast image and the corresponding generated mask of existing masses. (a) shows the original image provided in DICOM format. (b) shows the mask generated by extracting masses annotation from the associated XML file using our Matlab script.

3.4 Proposed Method

In this section, we describe our proposed methods to perform tumor segmentation for the considered breast cancer use-case. First, we explain our proposed double-dilated convolution module and point out the differences with previous dilation modules. Then, we identify the methods considered for pixel-level class balancing. Finally, we present the experimental methodology followed and shed light on the explainability methods adopted in this work.

3.4.1 double-dilated convolution

In this subsection, we explain our proposed modification to the dilated convolution module introduced in [34] in order to improve local spatial resolution. The standard convolution operation is defined as:

$$y[i] = \sum_k x[i+k]w[k] \tag{3.1}$$

where y is the output feature map, x is the input feature map, and w is the kernel.

Dilated convolution generalizes the standard convolution operation to:

$$y[i] = \sum_k x[i+l.k]w[k] \tag{3.2}$$

where l is a dilation factor. Note that the kernel did not change to reflect the sparsity introduced by dilation. However, the operation itself now considers the dilation parameter by skipping a range in the input defined by the dilated factor.

Our modified dilated convolution can be expressed by the following piece-wise function:

$$y[i] = \begin{cases} \sum_k x[i+l_1.k]w[k] & k \leq r_1 \\ \sum_k x[i+l_2.k]w[k] & k > r_1 \end{cases} \tag{3.3}$$

While dilated convolution uses l to define the dilation factor and r to define the size of the receptive field as $(2l+1)^2$, our modified dilated convolution operation, called *double dilated convolution* or *double atrous convolution* uses r_1 and r_2 to define the sizes of two receptive

fields using two kernels: the inner (core) kernel has a size of $(2l_1 + 1)^2$ and a dilation factor of l_1, and the outer (edge) kernel has a size of $(2l_2 + 1)^2$ and uses a dilation factor of l_2 .

double atrous is the only variant of convolution operations that can build a kernel with different dilation factors for different locations of the kernel. The motivation for this modification is to solve the inherent problem of *gridding* [69] in conventional dilation: Due to the sparsity of weights, dilated kernel's receptive field is only covering an area with checkboard patterns, especially when using high dilation factors. In standard (1-dilated) convolution, the size of the receptive field is equal to its effective coverage (non-zero weights). However, as the dilation rate increases, these two concepts are no longer equal and the receptive field will exponentially increase at the cost of lower (weaker) effective coverage due to the introduced sparsity. This trade off is reasonable at the edge of the kernel since the benefit of covering larger area and constraining the number of parameters can overweight the loss of some neighboring information. However, losing local information at the core of the kernel due to using the same dilation rate as in edges is not justified because it limits feature extraction ability of the network and can be avoided by having a denser kernel core.

3.4.2 Pixel-Level Class Balancing

In all deep learning models, the goal is to minimize the loss function, which measures the difference between the predicted output and the actual output of a neural network. Loss functions can be broadly classified into two categories: distribution-based loss functions and region-based loss functions. While distribution-based loss functions measure the difference between the predicted and actual output probability distributions across the entire input space, region-based loss functions aim to measure the difference between the predicted and actual outputs for a specific region of the input space [100].

In problems with severe pixel-level class imbalance such as ours, using a standard loss function like binary cross-entropy can result in poor performance, as the model tends to be biased towards the majority class. This can lead to high accuracy for the majority class (normal) but poor performance for the minority class (tumor), which is the class of interest in our case. Since the choice of the loss function greatly depends on the task being addressed and the nature of the data being used, we experiment with both region-based and distribution-based loss functions in order to identify the appropriate loss function for training our model.

To determine which loss function handles the pixel-level class imbalance problem existing in our data, we surveyed the literature for the loss functions used for the segmentation task, and we selected the four most widely-used ones: binary cross-entropy loss, weighted cross-entropy loss, dice loss, and Tversky loss. As categorized by the survey done in [100], the first two are considered distribution-based functions, whereas the latter two are region-based functions. These functions were also shown to perform well with the skull segmentation

task performed in [100]. The definitions and equations used to calculate the loss functions employed in our study are briefly illustrated below. As for the notation used in this section, all the sums run over N pixels which represents the number of pixels in the batch. The predicted probability of a pixel being classified as a lesion is denoted as p_i, whereas the ground truth value for a pixel is denoted as g_i. We consider g_i as 1 for pixels that are labeled as part of a lesion and 0 otherwise.

3.4.2.1 Binary Cross-Entropy

Binary Cross-entropy (BCE) [101] measures the difference between two probability distributions for a given random variable and it is frequently used for classification objectives with binary labels. Since image segmentation is performing classification on a pixel level, BCE is widely used for segmentation tasks as well [100, 102]. Although BCE does not consider sample reweighting, we include it in our comparison to act as a baseline performance, especially that it is still extensively use in training DL models. The equation for the Binary Cross-Entropy loss function is defined as:

$$L_{BCE} = -\frac{1}{N}\sum_{n=i}^{N}(g_i log(p_i) + (1 - g_i) log(1 - p_i)) \tag{3.4}$$

3.4.2.2 Weighted Cross-Entropy

Weighted cross entropy (WCE) is a variant of the binary cross-entropy loss, in which the positive examples get weighted by some coefficient to compensate for the unbalanced ratio of the training data [70]. It can be defined by the following equation:

$$L_{WCE} = -\frac{1}{N}\sum_{n=i}^{N}(\beta * g_i log(p_i) + (1 - g_i) log(1 - p_i)) \tag{3.5}$$

, where β is a hyper-parameter that can be adjusted to give more weight to the positive examples by assigning it to values larger than 1.

3.4.2.3 Dice Loss

Based on the Dice coefficient, a popular metric for measuring similarity between two images, the Dice Loss has been proposed in [71] specifically for segmentation tasks. The Dice coefficient is also equivalent to the F1 score, which measures the harmonic average of precision and recall to evaluate the performance of binary classifiers. For binary segmentation problems, the function is defined as:

$$L_{Dice} = \frac{1}{N}\sum_{n=i}^{N} 1 - \frac{2p_i g_i}{p_i^2 + g_i^2} \tag{3.6}$$

3.4.2.4 Tversky Loss

The Tversky loss (T) [72] can also be seen as a generalization of the Dice loss, where the parameters α and β are added to control the weights of false positives and false negatives, respectively. The function of Tversky loss is defined as:

$$L_{Tversky} = \frac{1}{N}\sum_{n=i}^{N} 1 - \frac{p_i g_i}{p_i g_i + \alpha p_i(1-g_i) + \beta(1-p_i)g_i} \tag{3.7}$$

When we set $\alpha+\beta = 1$, the Tversky index produces a set of $F\beta$ scores, which can be used to adjust the tradeoff between precision and recall [72]. When $\alpha = \beta = 0.5$, the Tversky loss simplifies to the Dice loss. In order to place more emphasis on minimizing false negatives, we can vary the value of β in the range [0.5,1] to reach the optimal performance for the lesion segmentation task. It has been shown in the literature that adjusting the parameters of the loss functions used during the training step helps the network to generalize and perform well in highly imbalanced data. Hence, we perform fine-tuning for the hyper-parameters used in all functions considered in this study.

3.4.3 Experimental Methodology

In this section, we illustrate the systematic approach used to implement our proposed methods and explain the steps followed in different experiments. All implementations were done using Matlab R2022b.

3.4.3.1 Baseline Model

To choose the baseline model for our analysis, we first considered two state-of-the-art segmentation architectures, namely U-Net [103] and DeepLabV3+ [96], similar to the work done in [104] for brain tumor segmentation. While U-Net [103] uses a standard classification convolutional neural network as its architecture block, DeepLabV3+ [96] uses a backbone of convolutional neural network followed by the Atrous Spatial Pyramid Pooling (ASPP) module, where the dilated convolution is heavily used. Although its structure was not expressly built for medical image segmentation, the DeepLabV3+ network with a ResNet18 backbone achieved the best performance for our dataset, when compared to the U-Net network in a pilot experiment. Hence, DeepLabV3+ with the default cross-entropy loss function was used as the baseline model for our experiments.

The DeepLabV3+ network adopts the encoder-decoder structure which is justified in their paper as a method to exploit multi-scale features from the encoder part and recovers the spatial resolution from the decoder part [96]. Since our proposed dilation module is applied to the convolution layers in the encoder components, in figure 3.4, we show the encoder path of the original Deeplabv3+ network when considering an output stride of 8 (the ratio of input image spatial resolution to final output) and the ResNet network as the backbone model. As illustrated in their work [1], atrous convolution with various rates is inherited in both cascaded modules and spatial pyramid pooling to enlarge the receptive field and incorporate multi-scale context without excessively sacrificing the image resolution. This enables the network to generate output maps with a resolution down-sampled only by a rate of 8 instead of 256, as the case in a typical pyramid-shaped convolution network.

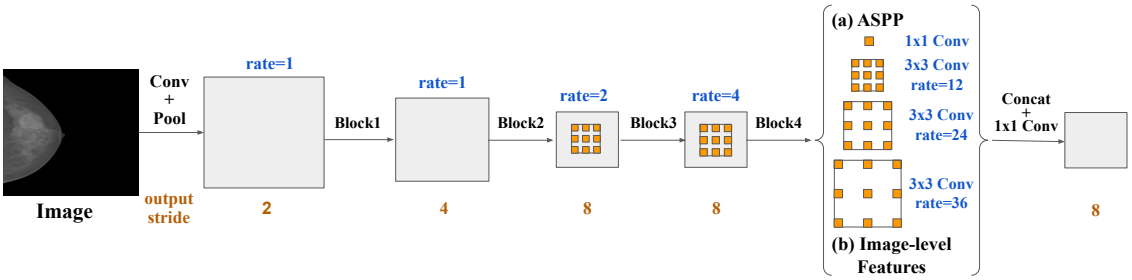


Figure 3.4: The original Deeplabv3+ encoder architecture used in our study. Dilated Convolution is inherited with different rates in the last 2 blocks of the ResNet backbone network and the parallel modules in the Atrous Spatial Pyramid Pooling (ASPP) module. The output of the ASPP is augmented with image-level features to produce the output feature maps. The figure is modified from [1] to show the dilation rates used with output stride=8.

3.4.3.2 Double-Dilated Convolution

There are many ways to implement the proposed double-dilated convolution. Methods can vary in memory footprint, speed, and ease of development. In this work, we provide a simple implementation of the proposed double-dilated kernel in a way that makes use of the available most efficient convolutions modules adopted by many deep-learning frameworks. As shown in Figure 3.5, the idea of applying a single kernel with two different dilation rates on a given input can be viewed as applying two different kernels, each with its own dilation rate, on the same input and then summing up the results of both convolution processes. This approach was inspired by the distributive property of the convolution process, which states that for any three discrete functions $h_1[n]$, $h_2[n]$ and $x[n]$, we can say that:

$$x * h_1 + x * h_2 = x * (h_1 + h_2) \tag{3.8}$$

The pitfall of this method is that it doubles the number of single-dilated convolution

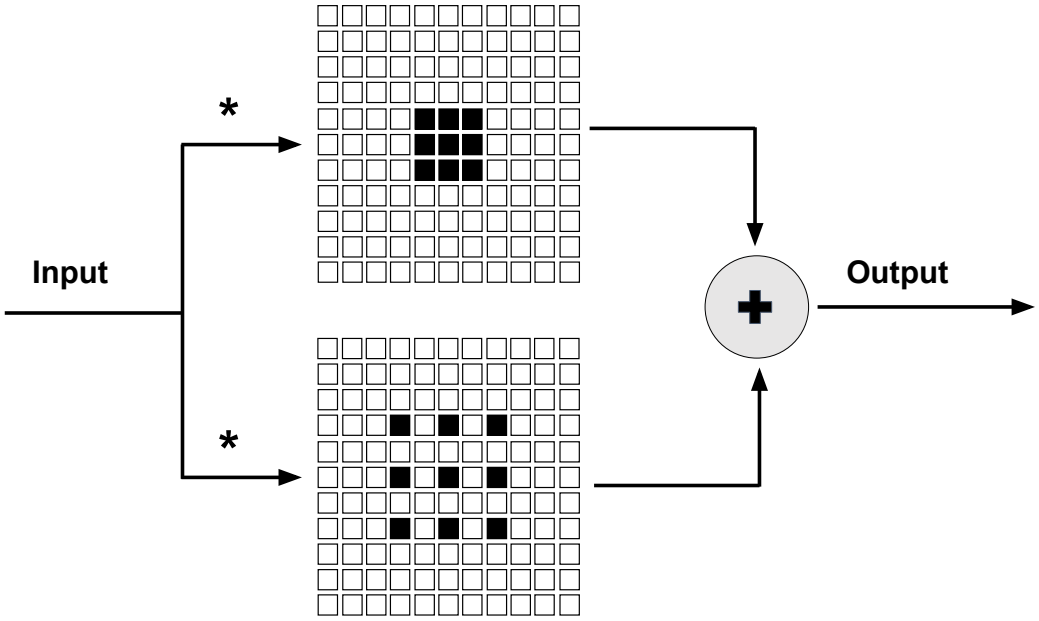


Figure 3.5: A simple implementation to achieve a double-dilated convolution with 1 and 2 dilatation rates using two parallel convolution branches: one with an inner dense kernel (rate=1), and the other for the sparse outer kernel (rate=2). The feature maps generated by the two processes are summed to produce an output equivalent to applying the double-dilated convolution.

layers in the network, which can affect the computational speed. However, this implementation does not require any additional modifications in the convolution process itself, which makes it easy to adopt in any state-of-art architecture using the existing optimized convolution modules in deep learning frameworks. This can serve as a proof-of-concept to evaluate the effectiveness of double-dilation in semantic segmentation tasks.

To measure the performance gain of the double dilation module, we plugged it into the Deeplabv3+ model to compare it with the original network. First, we inherited the Deep Learning Toolbox Model for Deeplab V3+ with Resnet18 as the backbone model and with an output striding factor of 8 for denser feature maps. Then, we applied our modification by replacing every dilated-convolution layer in the original network with two parallel dilated-convolution layers at different rates: one was fixed at rate = 1 to represent the dense core, and the other was kept the same as the dilation factor of the original layer, as showin in Figure 3.6. We then created a two-input addition layer to perform an element-wise addition on the two feature maps generated by the parallel convolution layers to generate the output of the double-dilated convolutions. This is done in the sequential blocks of the backbone model as well as the ASPP module which we now call Double Atrous Spatial Pyramid Pooling (DASPP) after applying our double-dilation technique.

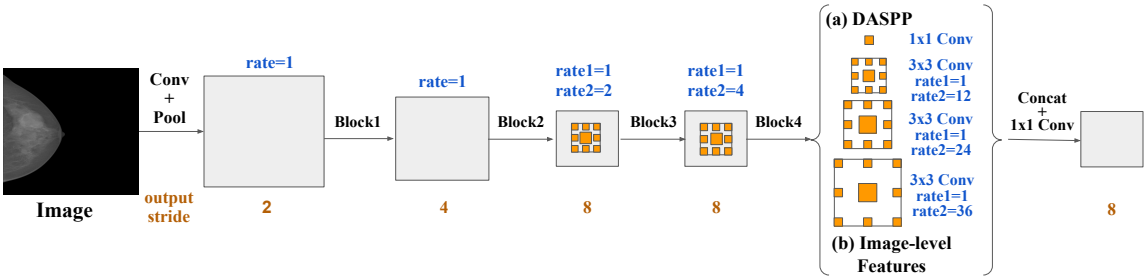


Figure 3.6: The modified Deeplabv3+ architecture after plugging the double-dilated convolution module. Each dilated convolution from the original network is replaced with 2 parallel convolutions: one undialtated convolution (rate1) and one dilated convolution with the same rate used in the original layer (rate2).

The modified network was trained following a similar protocol as the original training protocol mentioned in [1]. The training was done using an initial learning rate of 1e-2, with a mini-batch size of 12 images and with a maximum epoch number of 30. The data was shuffled after each epoch to increase the generalization capability of our model.

3.4.3.3 Pixel-Level Class Balancing

In this section, we explain the experimental approach followed in order to compare the different loss functions considered in this study. For each loss function, we created a different version of our model where the pixel-level classification layer at the end of the network is employing the corresponding loss function. If the function has hyper-parameters (e.g, class weights), the model is first fine-tuned using the 5-fold validation method to select the best combination of parameters for this loss function.

Table 3.1: Parameters used in different loss functions to address the pixel-level class imbalance problem. Class1 is the non-lesion class while class2 represents a mass lesion. Alpha and beta are used in Tversky loss to control the weights of false positives and false negatives, respectively.

	Weighted Cross-Entropy	Tversky
Parameters	[class1_weight, class2_weight]	[alpha, beta]
Range	[1, 10], [1,20].. [1,100]	[0.1, 0.9], [0.2,0.8], .. [0.5,0.5]

The hyper-parameters of different loss functions and their sets of values considered in this study are shown in Table 3.1. First, for the binary cross-entropy (CE) loss function, we used the default classification layer provided in the original Deeplabv3+ network without any modifications. In order to select the parameters range for the weighted cross-entropy (WCE), we first calculated the ratio between the number of pixels of the non-lesion class

(class 1) to the number of pixels classified as masses (class 2) in our training set. We found that the non-lesion pixels were around 70 times more frequent than the mass pixels. Hence, we varied the weights of the minority class from 10 to 100 with an incremental step of 10 by modifying the ClassWeights parameter in the original classification layer of the model, to evaluate different weighting ratios and select the best-performing one. As for both the Dice and Tversky Losses, since the latter is the generalized version of the first, we defined one custom pixel classification layer that implements the Tversky index and computes the loss based on it. This layer has two parameters: alpha and beta, which are both set to 0.5 when the Dice loss is used. On the other hand, to tune the parameters of the Tversky loss in a way that assigns more weight to the minority class, we consider different combinations of alpha and beta where the value of beta is in the range [0.5,1] and the value of alpha = 1-beta. We report the performance of all models using different loss functions on the 5-fold validation set.

3.4.3.4 Explainable AI Methods

To provide explanations for the output of our segmentation network, we incorporated explainable AI methods into our developed models. First, we implemented a simple technique to inspect our model by visualizing the activation of the last feature extraction layer in the segmentaion network. Although its simplicity, this method, which we call Activation Visualization in this paper, can help us discover which features the network learns by comparing areas of activation with the original image. We selected the last 6 channels of the final deconvolution layer in Deeplab V3+ as they provided comprehensive visualization of the decision-making step. Then, we employed two of the popular XAI methods, namely, Gradient-weighted Class Activation Mapping (Grad-CAM) and Occlusion Sensitivity, which both use visualization techniques to provide explanations to black-box architectures such as our CNN-based segmentation network.

Grad-CAM [105] is a visualization technique that highlights the important regions in an input image that contributed the most to the model decision by computing the gradients of the output with respect to the feature maps of the last convolutional layer, and then weighting the feature maps based on the strength of their gradients. This produces a heatmap that shows which regions of the input image were most important for the model's decision. On the other hand, the Occlusion Sensitivity [106] involves systematically occluding different parts of an input image and observing the impact on the output of a CNN. By measuring the change in the model's output as different parts of the input are removed, it highlights which regions of the input image are most important for the model decision.

We adopt both techniques in our networks to produce heatmaps of the image regions that participate in segmenting a mass in a mammogram screening. This is done in both the original Deeplabv3+ network and the modified double-dilated network to analyze explana-

Algorithm 1 Pixel Flipping Similarity Scores

input: Input image: img, Ground truth image: gt, Explanation map: map, Segmentation Network: net
output: Similarity scores list for different percentages of pixel flipping: $scores$
procedure PIXELFLIP(img, gt, map, net)
    $img_{segmented} \leftarrow net(img)$
    $score_{initial} = similarityScore(imgsegmented, gt)$
    $scores = []$
    $idx_{sorted} \leftarrow$ Get indices of sorted elements in map
    $idx_{start} \leftarrow 1$
    for `i in 1:10` **do**
        $idx_{end} \leftarrow idx_{start} + i/100 * size(img)$
        $img(idx_{sorted}(idx_{start} : idx_{end})) \leftarrow radndom()$
        $img_{segmented} \leftarrow net(img)$
        $scores(i) \leftarrow similarityScore(img_{segmented}, gt)$
    end for
    $scores_{normalized} = scores/score_{initial}$
    return $scores_{normalized}$
end procedure

tions for the segmentation made in both cases. We then provide quantitative evaluation of these models along with the activation visualization method by measuring the image entropy and plotting the pixel flipping graph similar to previous works [88, 107, 108]. While the entropy evaluates the complexity of the explainable method by measuring the randomness (uncertainty) in the generated explanation map, pixel flipping curves determine whether the removal of the features highlighted by the explanation as being the most relevant, results in a significant reduction in the prediction capabilities of the network. We use the Matlab's built-in function *entropy* to calculate the image entropy values for different explanation maps generated for all images in the validation set then we average these values to have a single numeric metric of complexity. To plot the pixel-flipping graphs of different XAI methods for the segmentation task, we implemented the algorithm explained in Algorithm 1 to calculate the similarity scores achieved at different percentages of pixel flipping. The pixel-flipping process involves an iterative removal of input features, starting from the most relevant and moving towards the least relevant until 10% of the image pixels are flipped, while tracking the changes in the segmentation network output. The resulting decay in the similarity scores are then plotted as a curve, with a faster decrease indicating a more reliable explanation method that aligns with the decision of the neural network. These curves are computed and averaged over the entire validation set to obtain a comprehensive evaluation of the faithfulness of the explanation algorithm being studied.

3.5 Performance Evaluation

In this section, we start by explaining the metrics we utilized in this study to evaluate our models. Then, we report the results of different experiments and discuss the implications of the those results.

3.5.1 Segmentation Evaluation Metrics

In order to provide a comprehensive analysis, we measured the performance of the tumor segmentation task on two levels: pixel-level and lesion-level. While the first provides a measurement of how well the model predicts each pixel of the image, the latter provides high-level metrics to indicates how well the model predicts a lesion in the image.

3.5.1.1 Pixel-Level Evaluation

As previously established in semantic segmentation tasks, the pixel-level similarity between the segmented image and the reference image determines the quality of the segmentation model [109]. The two widely-used similarity indicators in validating medical volume segmentation are Dice similarity [110] and Jaccard similarity [111]. While the first calculates the ratio of matches to mismatches, the latter computes the ratio of matches to the overall membership. We measured both similarity scores for each class separately to have a better understanding of the model performance for detecting the lesion class as well as the non-lesion class. However, as pointed out by [112], including both of the metrics as validation metrics does not provide additional information since they both measure the same aspects and provide the same system ranking. Therefore, in this work, we chose the Dice similarity as our main pixel-level evaluation metric since it is the most used metric in validating medical volume segmentations, and it is equivalent to the F1 score in binary segmentation problems [112].

In addition to the similarity metric, we reported the validation accuracy as a standard metric for CAD analysis, which measures the ratio of correctly classified pixels to the total pixels in the input image. This can be used to compare results with similar studies performing mass segmentation on the INBreast dataset. However, the accuracy can suggest overestimated results since our dataset is significantly skewed towards the non-lesion class.

3.5.1.2 Lesion-Level Evaluation

Considering that the task at hand is a tumor segmentation task, we found that measuring metrics at the level of mass lesions provided highly intuitive indicators of the prediction quality. As suggested by the authors who provided the INBreast dataset, there are two metrics recommended to evaluate segmentation models developed for this type of imaging

data: the miss detection rate and the false-positive rate [59]. The miss detection rate is the percentage of reference masses that were not detected by the algorithm, whereas the false positive rate is the percentage of automatically detected masses that do not correspond to actual masses. Both metrics provide eye-level indicators of the model performance, which can be easily interpreted by humans (e.g., radiologists).

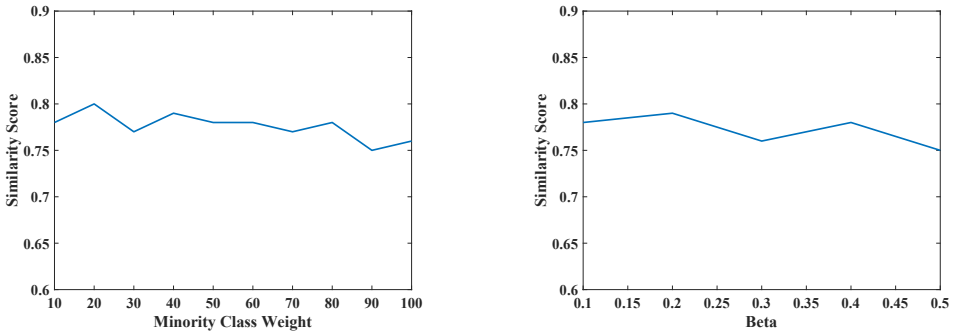
To calculate the two lesion-level metrics, we first determine whether a detected mass is correctly classified or not. This is done by measuring the overlap between the detected mass and the manually-annotated mass, and the detection was considered correct if the overlap (i.e., intersection over union) between the detected and true lesions is > 0.5. Then, the number of miss-detected lesions was calculated as the number of undetected reference masses and the number of false positive lesions was calculated as the number of automatically detected masses minus correctly classified masses. To normalize both values, we divide them by the number of actual masses present in the validation set, resulting in the lesion-level rates of miss-detection and false positivity, respectively. This approach has been frequently used in CAD research for evaluating object detection algorithms in medical imaging [59].

3.5.2 Results and Discussion

In this section, we evaluate the performance of our proposed methods for tumor segmentation in mammogram screenings. First, we present the results of the experiments conducted to tackle the pixel-level class imbalance using the baseline model. Then we show and discuss the performance of the proposed dilation module for convolution. Finally, we display the generated explanation heatmaps for selected mammogram segmentation results and discuss the performance of the adopted XAI methods. All results are reported as the average of the 5 validation sets generated using the standard k-fold validation method.

3.5.2.1 Pixel-Level Class Balancing

First, we show how the performance of the baseline model changes when varying the hyper-parameters used in loss functions. Since the WCE and T loss functions are the ones that include parameter tuning, we show the 5-fold validation sensitivity plotted against different parameters of both functions in Figure 3.7. As shown in Figure 3.7(a), it can be noticed that increasing the weight of the minority (lesion) class does not always improve the performance. In our case, the model hit the best similarity at class weights = [1,20] for the non-lesion and lesion classes. This emphasizes the necessity to fine-tune the class weights when considering the employment of the WCE loss in order to select the best-performing hyper-parameters for the problem at hand. Similarly, ranging the beta parameter in the Tversky loss formula to inherently add more weight to the less-frequent class while training our model resulted in a fluctuating validation similarity as in Figure 3.7(b). This also indi-



(a) Weighted Cross-Entropy Loss (b) Tversky Loss

Figure 3.7: Plot of the 5-fold validation sensitivity against the loss function hyper-parameters. (a) shows how the performance of the WCE loss-based model changes when varying the minority class (lesion) wight from 10 to 100. (b) shows the performance of the Tversky loss-based network when tuning the Beta parameter in the range [0.1,0.5].

cates that increasing beta does not guarantee better performance in highly skewed datasets, which suggests the need to perform parameter searching when using the Tversky loss. Given our task and our data, the baseline network configured with Tversky loss achieved the best sensitivity coefficient when alpha and beta were set to 0.2 and 0.8, respectively.

Table 3.2: Results of 5-fold validation of the baseline model using different loss functions with tuned parameters to address the pixel-level class imbalance problem. The considered loss functions are Cross-Entropy (CE), Weighted Cross-Entropy (WCE), Dice (D) and Tversky (T). Both lesion-level and pixel-level evaluation metrics are reported.

	CE	WCE [1, 20]	Dice	Tversky [0.2, 0.8]
Miss Detection Rate	0.17	**0.08**	0.33	**0.08**
False Positive Rate	0.29	0.29	0.23	**0.20**
Similarity	0.74	**0.79**	0.75	**0.78**
Accuracy	0.98	0.99	0.99	0.99

Then, we examine the results obtained by the original Deeplabv3+ model when trained using all different loss functions with the tuned parameters, which are shown in table 3.2. The results are averaged over all validation folds. Although the standard CE loss resulted in a similar False Positive Rate as the WCE loss, the latter managed to significantly reduce the number of miss-detected masses and enhance the overall similarity coefficient. On the other hand, the Dice loss function managed to slightly reduce the number of unmatched detected lesions, compared to the Entropy losses, at the expense of missing a significantly larger number of actual lesions. We can see that both the Weighted Cross Entropy loss

and the Tversky loss achieved almost the same results in terms of the pixel-level similarity coefficient and the miss detection rate. However, adopting the Tversky loss function resulted in a 9%l less False Positive Rate than the WCE loss. Hence, we employ the Tversky loss as the network's loss function in the following experiment.

3.5.2.2 Double-dilated Convolution

In Table 3.3, we compare the results of the original Deeplabv3+ model with the modified double-dilated model. We can see that introducing the modified convolution module with double dilation rates preoved to be effective in enabling the network achieve better performance in terms of both the Dice similarity and the Miss Detection rate. This improvement can be attributed to the increase of the resolution of the inner kernels by the double-dilated module to encode multi-scale context information existing in the input image. By reducing the number of miss-detected masses, early diagnosis can be facilitated and death rates can be lowered. Figure 3.8 shows a snapshot of the radiologist-annotated mammogram images along with their corresponding automatically-segmented images detected by the modified model for a validation set of 22 mammograms.

Table 3.3: Results of the 5-fold validation of the original Deeplabv3+ model and the modified model with the double-dilated module. Both lesion-level and pixel-level evaluation metrics are reported.

	Deeplabv3+	Double-Dilated Deeplabv3+
Miss Detection Rate	0.08	**0.04**
False Positive Rate	0.20	0.20
Similarity	0.79	**0.81**
Accuracy	0.99	0.99

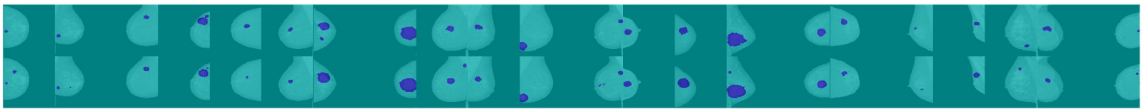


Figure 3.8: A snapshot of segmentation results of 22 validation mammograms generated by the proposed network after plugging the double-dilated convolution module. Images include CC or MLO views from left or right breasts. The top images show the mass annotations made by radiologists while the bottom images display the segmentation done by our CAD model.

On the other hand, the proposed model still predicted a number of false (unmatched) lesions at the same rate as the single-dilated network (20% of the actual number of tumors), which was still quite high. This aligns with the remarks made by previous studies where

CAD systems were reported to result in relatively high false positive rates [113–115]. It was suggested by [116] that the superimposition of tissues in 2D digital mammography contributed to this high rates of false positives, which encourages more integration of 3D mammograms in CAD systems in the future.

To have a better look at where the modified model outperforms the original one in terms of miss-detection rate, in Figure 3.9, we show four mammograms from one validation set where we can analyze this behavior. It is noticeable that the masses that were not detected by the original network in these screenings are relatively small in size. The double-dilated network, on the other hand, was able to partially segment many of these small masses due to preserving the local resolution of the input image throughout the CNN network by employing a denser kernel at the core of the convoluion window. This trend was consistent in different screenings using different validation sets. At the same time, some masses were too small to be detected in both networks, such as the bottom lesion segmented by radiologists in screening (c). We can also see that in screenings (a) and (d), this decrease in the number of missed masses came at the cost of increasing the falsely detected ones. That means that some other anatomical structures in the screenings were mistakenly classified as suspicious lesions by the modified network. However, as reported in Table 3.3, the overall average false positive rate was the same in both networks, which still gives the advantage to the double-dilated network due to its lower miss-detection rate.

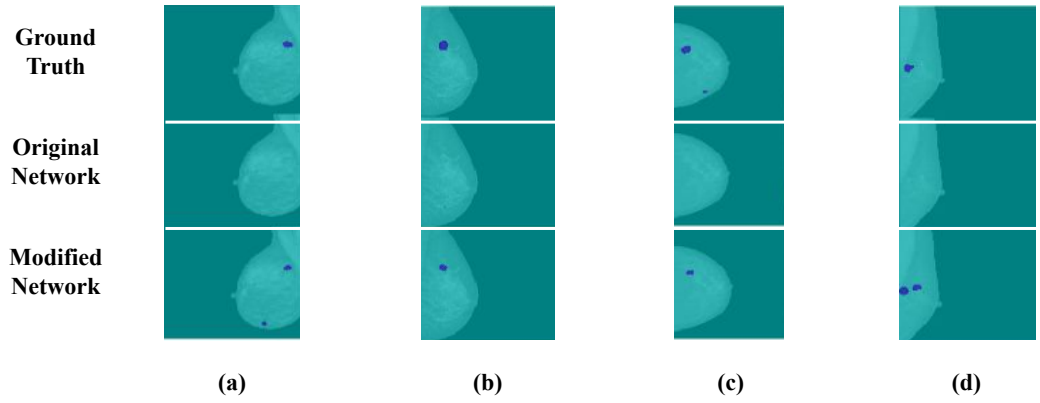


Figure 3.9: Selected segmentation results for four validation mammograms generated by both the original network and the modified one compared to the ground truth segmentation.

3.5.2.3 Explainablity via Visualization

Since it is difficult to visualize the results of applying different explanation models on all images in the validation set, in Figure 3.10, we show selected examples of the heatmaps generated for one mammogram screening using the three considered explanation techniques:

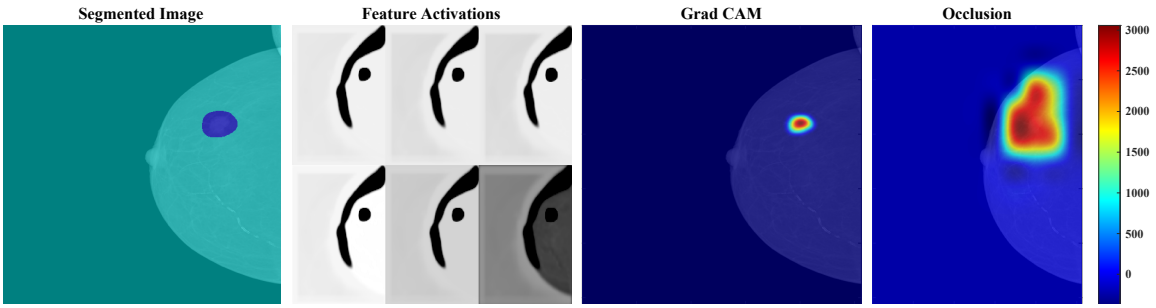


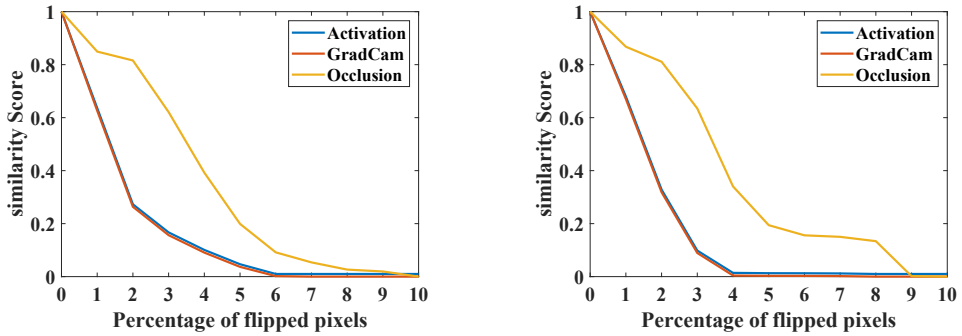
Figure 3.10: Examples of segmented mammogram images generated by the modified network shown along with explanations of the output segmentation. Different explanation maps are shown using Activation Visualization, GradCAM, and Occulsion Sensitivity.

Activation Visualization, Grad-CAM and Occlusion Sensitivity. However, similar remarks were made when visualizing other segmented mammogram explanation results. First, by simply visualizing the activation maps from six channels at the last feature extraction layer, we are able to see how the network arrived at a final segmentation decision that aligns with the extracted features. This, although might not be a very compelling explanation for radiologists, can assure us as researchers that the network is not overfitting to irrelevant attributes of the image while classifying a certain region as a mass. Comparing the visual results of the two adopted XAI methods, we observe that Occlusion Sensitivity tends to generate explanation maps that are more heated than the corresponding heatmap generated by GradCAM, which indicates that more regions in the image are assigned high positive relevance values for segmenting existing masses. This may be attributed to the big difference in the underlying technique used in the two methods. In Occlusion, different regions are removed to measure their contribution to the output decision. This might make it easier for the model to attribute higher weights to pixels that are not in the tumor area only because their removal affected the segmentation results. Whereas in the GradCAM maps, the red areas are much more constrained in the mass region since this method weights the gradients of the activation maps at the last layer of our network, which are highly correlated to the segmentation output itself.

Table 3.4: Image entropy results for explanation maps generated by different explainable methods with the original and double-dilated segmentation networks. The table shows the average results for all images in the validation set.

	Activation	Grad-CAM	Occlusion
Original	3.154	**0.119**	2.526
Double-Dilated	3.505	**0.139**	1.445

In order to quantitatively compare the complexity of different explanation methods, we first calculated the image entropy of all explanation maps generated for the mammogram screenings in the validation set. For the activation visualization method, we considered the activation neurons of the last channel in our calculations. Table 3.4 shows the average entropy results for different XAI methods when applied on images segmented by both the original Deeplab V3+ network and the double-dilated network. The results show that the Activation Visualization had the highest average entropy value, indicating a very high level of randomness in explanation. The Occlusion Sensitivity had less but relatively high average entropy, with values distributed over a wide range (e.g., [0.113, 2.089] with the modified network). In contrast, the Grad-CAM method achieved the lowest complexity with an average value of around 0.12 in both networks, indicating the least randomness in explanation mapping. We can also notice that the proposed network resulted in slightly more complex explanations than the original network when using both Grad-CAM and Activation maps. This can be due to the added complexity of the double-dilated kernels, which are reflected in the feature maps employed for explanation mapping in these two methods. However, this was not the case with the Occlusion maps, since the algorithm used is not dependent on the activation values.



(a) Original Network (b) Modified Network

Figure 3.11: Plot of pixel flipping graphs for different explanation methods. (a) shows the performance of different explainable models when applied to the original DeeplabV3+ Segmentation Network. (b) shows the performance with the double-dilated network.

Then, in Figure 3.11, we display the pixel-flipping graphs for different explanation methods when using the original and the modified networks. The plotted graphs are the averaged graphs for the whole validation set. We can see in the figure that the Grad-CAM and the visualized activation methods yielded the same average similarity scores for different percentages of pixel flipping, which can be understandable since both methods mainly rely on the activations of the last convolution layer of the network. Moreover, the Grad-CAM method achieved a higher decay rate than the Occlusion Sensitivity method in all cases, which

suggests that Grad-CAM explanations are more truthful compared to the other method. Although this trend is the same in both original and modified networks, we can observe that the Grad-CAM curve decayed faster with the double-dilated Deeplab V3+ network than with the original network, which indicates a higher level of truthfulness in Grad-CAM explanations of the proposed segmentation network. This proves how the network structure can affect the performance of explanation methods as discussed in [117]. Overall, the pixel-flipping results agree with the entropy results and our visualized analysis, which all suggest that the Grad-CAM explainability technique is able to provide truthful and comprehensible explanations for mammogram mass segmentation results. This encourages the adoption of this simple yet powerful tool in medical image segmentation networks to improve their transparency and promote their integration into clinical practice.

3.6 Conclusion

Computer-aided Diagnostic (CAD) modeling of cancer is crucial to identify the disease at an early stage and achieve better outcomes. In this work, we addressed three challenges in medical image segmentation systems. First, we proposed a double-dilated convolution module that perceives complex kernels with denser cores in order to eliminate the problem of decaying local resolutions in medical images which occurs when applying existing CNN-based segmentation architectures. With the use of the state-of-art Deeplabv3+ network, we explain our simple implementation of the double-dilated convolution which uses two dilation factors in parallel to replace the traditional dilated convolution layer used in the original network. To evaluate our proposed convolution method, we performed our analysis on the publicly available mammogram screenings provided by the INBreast dataset. Second, to solve the pixel-level class imbalance problem existing in the data, we compared using four widely-used loss functions in training our network to determine the best-suited method for this task. Finally, we adopted explainability techniques to provide interpretable segmentation results and quantitatively compared their performance in terms of complexity and truthfulness.

Based on our experiments, it may be concluded that double-dilated convolution can achieve promising results by increasing the similarity scores and lowering the miss-detection rate when adopted in CNN-based networks performing medical imaging segmentation. In addition, evaluating different loss functions The Tversky loss function showed the best validation results compared to the other functions. It also emphasized the importance of selecting the optimal loss function with the fine-tuned hyper-parameters to avoid poor performance on the underrepresented classes. Finally, the explainable AI results show the effictiveness of Grad-CAM in explaining CAD segmentation results to provide medical professionals with interpretable and trustworthy decisions.

In the future, we will experiment with adopting the proposed methods in large datasets with different medical image modalities to verify the effectiveness of our segmentation and explainablity techniques. Also, similar to the double-dilated convolution module implemented in the study, the concept of multi-resolution dilated convolution can be extended to develop an N-dilated convolution module which employs a kernel with N sparsity factors on different scales. Moreover, our research direction can explore the ability of our proof-of-work in this study to generalize for other non-medical dense prediction and object detection tasks. Finally, we intend to investigate the phenomenon of high false positive rates associated with CAD systems in order to spare patients from the negative psychological impact and unnecessary biopsies.

Chapter 4

Two-Step Stage-Specific Machine Learning Model for Breast Cancer Survivability Prediction

Disclaimer: This chapter is an IEEE publication and we are adhering to IEEE's copyright rules to report it [118].

Following the diagnosis of cancer with the aid of medical image segmentation (Chapter 3), medical professionals need to perform prognostic analysis to make informative decisions about the appropriate course of actions for a specific patient. While traditional medical informatics focus primarily on disease classification problems, the development of Computer-Aided Prognosis systems for patients suffering from multi-stage conditions, such as breast cancer, surprisingly remains an overlooked research topic. In this work, we address the survivability prediction problem for breast cancer patients due to the importance of survivability analysis and prediction for healthcare providers to make informed decisions on recommended treatment pathways for different patients. Then, we combine two main strategies in solving the breast cancer survivability prediction problem using Machine Learning techniques. In the first strategy, we model the survivability prediction task as a two-step problem, namely 1) a classification problem to predict whether or not a patient survives for five years, and 2) a regression problem to forecast the number of remaining months for those who are predicted to not survive for five years. The second strategy is to develop stage-specific models, where each model is trained on instances belonging to a certain cancer stage, instead of using all stages together, in order to predict survivability of patients from the same stage. We investigate the impact of adapting these strategies along with applying different balancing techniques over the model performance using the National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) dataset. The obtained results demonstrate that the proposed methods prove effective in both survivability

classification and regression.

4.1 Introduction

After diagnosing a cancer patient using the computer-aided medical image segmentation system proposed in Chapter 3, medical professionals confirm the positive diagnosis and collect clinical information about the developed disease typically by performing a biopsy procedure [59]. This involves removing a tissue sample from the tumor to examine it under a microscope. The collected information can help determine the stage of the disease and can then be used to predict survivability and plan appropriate treatments. With the recent advancement of biomedical imaging, Internet of Medical Things (IoMT), and medical cyber physical systems, computer-aided disease classification have gained much popularity to complement the caregivers in diagnostic decision making and significantly reduce their burden. However, disease survivability prediction remains a significantly overlooked area for numerous diseases, particularly with multiple stages such as congestive cardiac disorders, various cancer types, chronic kideny disorder, diabetes, and so forth. Fig. 4.1 depicts the research gap and the need for survivability analysis and prediction in terms of survivability regression models. In this paper, we address this issue by envisioning an appropriate framework to combine the disease survivability classification and regression tasks in a seamless manner. Among these numerous multi-stage chronic disorders, we continue to consider the

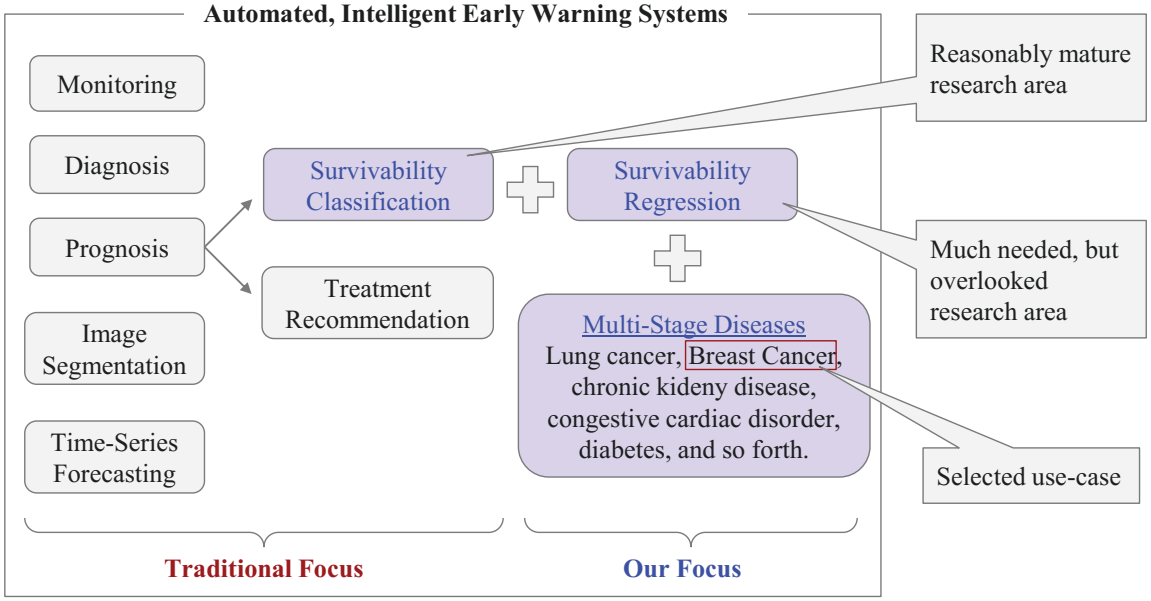


Figure 4.1: The research gap in the current state of research in health-related intelligence systems, illustrating the need for survival time estimation for multi-stage diseases in next-generation, automated and intelligent early warning systems for chronic diseases.

breast cancer use-case, similar to previous chapter, as it is well-suited for the objective of this study (refer to Chapter 2. Moreover, breast cancer is the most commonly diagnosed cancer among female patients in the world [55]. According to the Canadian Cancer Statistics in 2021, 1 in 8 females (12%) is expected to be diagnosed with breast cancer in their lifetime [55]. With emerging biomedical technologies allowing better prognostic factors to be measured and recorded, cancer survival prediction has become a popular research interest in the last decade. Accurately predicting patients' survivability may provide medical teams with appropriate treatment recommendations and help prescribe personalized medicine [119]. However, most cancer survivability studies aim to only predict patients' five-year survivability, which may not be sufficient for medical decisions. For instance, if a patient is predicted to not survive, the survival time of the patient remains unknown. Therefore, it is important to develop an accurate regression model for survival time prediction of breast cancer patients to provide more precise information for medical decision making [120].

The Surveillance, Epidemiology, and End Results (SEER) repository is the most comprehensive publicly-available source of information on cancer incidence and survival in the United States [54]. It has been used in many studies on breast cancer, mostly for prognosis analysis. SEER cancer records are assigned a phase, referred to as the "summary stage", which can be defined as the most basic way of categorizing how far a cancer has spread from its point of origin [121]. The stages associated with malignant tumors in the SEER

database are Localized, Regional, and Distant stages. The stage at which the patient is first diagnosed greatly affects the rate of survivability among breast cancer patients. For example, between 2011 and 2017, the five-year survival rate for women diagnosed with localized-stage breast cancer was as high as 99%, whereas the rate was only 29% for those with distant stages [45]. Most breast cancer survivability prediction studies, however, tend to model incidences from all stages together, while only providing the stage as an input attribute to the model.

In this paper, we aim to develop a novel prediction system for breast cancer survival time that can serve as a proof-of-work for generalization toward survivability prediction in other multi-stage diseases. As depicted in Fig. 4.2, our proposed method combines two main strategies in solving the breast cancer survivability prediction problem using Machine Learning (ML) techniques. Our first strategy formulates the breast cancer survivability prediction task as a two-step problem: 1) a classification problem to predict whether or not a patient survives for five years, and 2) a regression problem to predict the number of survival (remaining) months for the patients who have been predicted not to survive for another five years. Our second strategy consists of stage-specific ML models, where each model is trained on instances belonging to a certain summary stage, instead of simultaneously exploiting all stages, in order to predict survivability of patients from the same stage. Then, we investigate the effect of adopting our proposed strategies on both classification and regression performances. We also evaluate the prognostic value of the Lymph Node Ratio (LNR), which is defined as the ratio between positive lymph nodes and the examined ones, when employed as an input to the proposed model. Moreover, we compare applying different balancing techniques, including under-sampling, over-sampling, and cost-sensitive learning, to overcome the class-imbalance problem in the data. To the best of our knowledge, this is the first study that explores breast cancer survivability prediction using a two-step stage-specific framework.

The remainder of the paper is organized as follows. We survey the relevant research work in section 5.2. Next, the dataset preparation is described in section 5.3. Our proposed strategies, model selection, balancing techniques, and experimental methodology are described in section 4.4. The performance our proposal is presented in section 4.5. Finally, section 4.6 concludes the paper and provides future research directions.

4.2 Related Work

In this section, we provide the recent research work on breast cancer classification and survivability prediction. Different methods emerged in the literature for performing breast cancer prognostic analysis. For instance, Delen *et al.* [119] employed the SEER data collected in the years 1973-2000 to compare the performance of Artificial Neural Networks

(ANN), Decision Trees (C5) [14] and Logistic Regression in predicting breast cancer five-year survivability. The C5 algorithm was found to be the best of the three models evaluated in terms of accuracy, sensitivity, and specificity. The study also proved that for this specific problem, Multi-layer Perceptrons (MLPs) [122] were more suitable than other ANN architectures, such as Radial Basis Function (RBF), Recurrent Neural Network (RNN), and Self-Organizing Map (SOM). Similarly, the C4.5 [14] decision tree algorithm was found to achieve a better accuracy in [123] when compared to Naïve Bayes, and Neural Networks for five-year survival prediction, and also in [124], when compared to the Naïve Bayes and Logistic Regression in predicting patients' ten-year survival status. In a different direction, Hussain *et al.* [125] experimented on performing dimensionality reduction on the SEER data (from 1973-2010) by employing the Principal Component Analysis (PCA). Their work revealed that reducing the 14 cancer-related SEER attributes to only five components obtained by PCA that captures 98% of total variance, only resulted in a 0.2% drop in the five-year survival classification accuracy of the Logistic Regression model.

Although the vast majority of the research conducted in this area only focused on the survival classification problem, i.e., predicting whether or not a breast-cancer patient will survive for a certain amount of time (typically 5 years), some researchers also studied the survival time estimation problem for breast cancer patients. For instance, the authors in [126] developed a web-based decision support system that predicts both the five-year survivability and the survivability period that a breast-cancer patient could survive. They employed decision trees and Generalized Linear Models (GLMs) for the classification and regression tasks, respectively. However, these models worked separately without having a sequential setting. In other words, the output of the classification task was not taken into consideration when performing the regression task. Next, Teng *et al.* [20] proposed a Bayesian prognostic model with age stratification to predict survival time in months for breast-cancer patients. The effectiveness of their model was demonstrated via Concordance statistic and compared with the classical Cox model and other ML approaches.

In spite of the large number of models proposed in the literature for breast cancer survival prediction, the following shortcomings still exist and need to be addressed:

- Limited attention has been given to the survival time estimation for breast-cancer patients, particularly those who are unlikely to survive for five years. Predicting the remaining amount of time for a non-surviving patient is necessary for medical service providers to allocate resources and decide recommended medication.

- Almost all studies employ instances from all cancer stages combined to train the model, which can negatively impact the model performance for two reasons. First, the survival rates for different stages are greatly different. Second, the importance of the features used to predict survivability can vary from one stage to another. Hence,

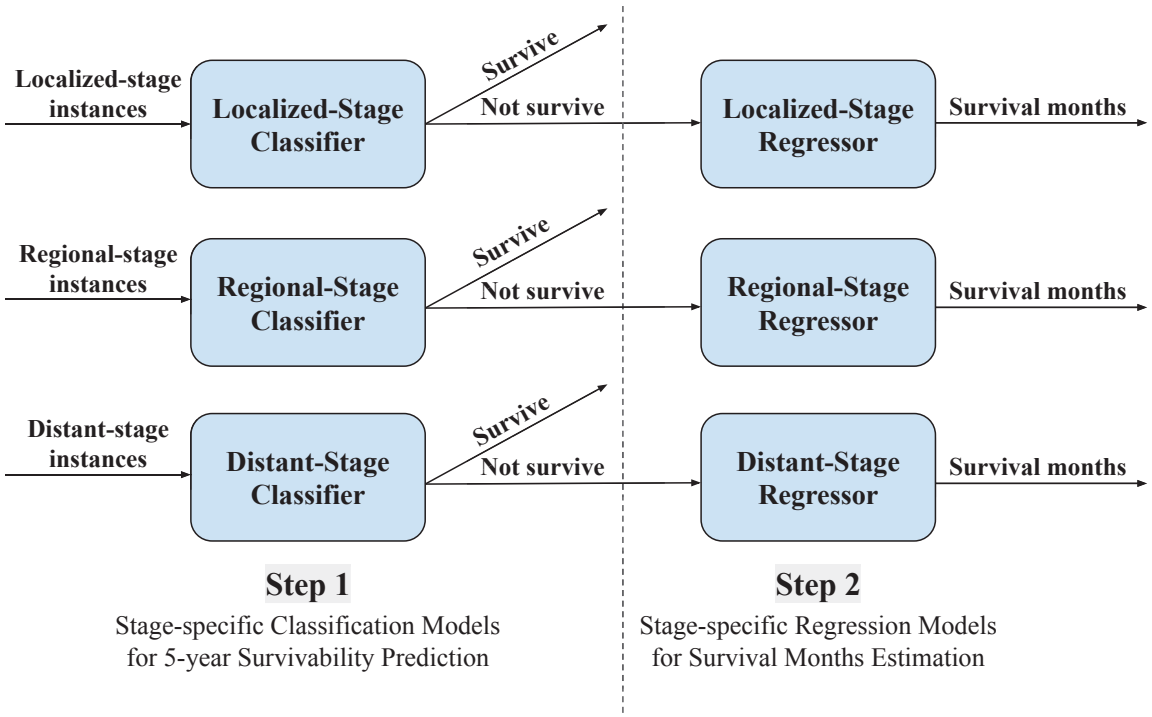


Figure 4.2: The proposed two-step stage-specific framework for breast cancer survivability prediction in inference time. The regression models in the second step are only trained on instances of patients who died within 5 years of diagnosis.

the trends or patterns existing in the stage with the largest number of instances tend to dominate the behavior of the ML model adopted for all the stages.

- Although it is known that the breast cancer data suffer from a high class imbalance, most studies use accuracy to evaluate their models that can be misleading. For instance, when the survival rate of breast cancer for all stages combined was known to be 90% in 2017 [45], developing a model which predicts all instances to survive, could easily attain a 90% classification accuracy, when it did not actually learn anything. This would result in having an incredibly high number of false positives (when considering the positive class as the surviving class).

- The train-test splits used in the previous research work mostly adopt conventional methods, such as k-fold validation split or 80:20 fixed train-test split. Hence, such models are not tested against instances coming from different time periods than the ones in the training set. This is a drawback as it has been demonstrated that data-driven knowledge for breast cancer survivability is not persistent over time [46]. This suggests the need for over-time validation before these models can be clinically applied.

In the remainder of this paper, we aim to address the above-mentioned shortcomings while

utilizing and re-purposing the findings from the previous research work.

4.3 Data Preparation

In this section, we describe the dataset selection and preprocessing phases that are required for the machine learning model development.

Table 4.1: The SEER dataset attributes used in our analysis and their corresponding data types.

#	**Attribute name**	**Type**
1	Age recode with $<$ 1 year olds	Categorical
2	Year of diagnosis	Numeric
3	ER Status Recode Breast Cancer (1990+)	Categorical
4	PR Status Recode Breast Cancer (1990+)	Categorical
5	CS tumor size (2004-2015)	Numeric
6	Breast - Adjusted AJCC 6th T (1988-2015)	Categorical
7	Breast - Adjusted AJCC 6th N (1988-2015)	Categorical
8	Breast - Adjusted AJCC 6th M (1988-2015)	Categorical
9	Breast - Adjusted AJCC 6th Stage (1988-2015)	Categorical
10	Grade (thru 2017)	Categorical
11	Summary stage 2000 (1998-2017)	Categorical
12	Regional nodes examined (1988+)	Numeric
13	Regional nodes positive (1988+)	Numeric
14	Survival months	Numeric
15	Vital status recode (study cutoff used)	Categorical
16	COD to site recode	Categorical

List of notations/acronyms used:
ER: Estrogen Receptor, PR: Progesterone Receptor, CS: Collaborative Stage, AJCC: American Joint Committee on Cancer (6th Edition), T: Extent of Tumor, N: Spread to nearby Lymph Nodes, M: Metastasis, COD: Cause Of Death.

4.3.1 Dataset selection

The Surveillance, Epidemiology, and End Results (SEER) repository contains cancer incidence data from population-based cancer registries covering approximately 47.9% of the US population. We used the latest version of the SEER database (November 2021), which covers de-identified cancer incidences from years 2000 through 2019 comprising a total of 8,721,474 cancer incidences including 1,425,552 breast cancer incidences (both malignant and in-situ). The database was accessed through the SEER* Stat software (version 8.4.0.1) [127] after signing a data use agreement. Since the attributes provided by SEER are not necessarily persistent throughout the years, we extracted data for females diagnosed with malignant breast tumors between 2004 and 2015, after ensuring that all the columns-of-interest were collected during those years. We selected the cancer-related features that were commonly utilized in the literature for this purpose and were demonstrated to be highly correlated to survival prediction [19, 20, 44, 119, 123–126, 128–130]. Table 4.1 lists the attributes used in our study. All the listed attributes were used as features for the predictive models except for the last three which were only used to define the target variables for both regression and classification tasks.

4.3.2 Preprocessing

We applied some preprocessing steps to clean the data, perform some feature engineering, and prepare the labels for the classification problem. All the records that had missing values along any attribute were removed. Then, a new feature was created to represent the Lymph Node Ratio (LNR), which was found to improve the model ability to predict survivability [20]. However, instead of developing a whole new model to estimate LNR prior to feeding it to the main prediction model as considered by Teng *et al.* [20], we derived the values for LNR from the data by dividing the number of positive lymph nodes over the number of examined lymph nodes. We noticed that the difference between using the derived LNR and the estimated one was negligible in terms of the model performance. However, the estimated LNR adds an unnecessary complexity to the system and requires more inference time than the calculated LNR. All instances that had positive lymph nodes more than the examined ones, or had zero examined nodes were excluded. In our experiments, we found that adding LNR resulted in approximately 0.1 improvement in the model average F1-score. Next, we defined the output attribute in both classification and regression tasks. As for the regression, the survival months attribute was directly used as the dependent variable. However, the five-year survivability (survival) was labeled based on three attributes, namely Survival Months (months), Vital Status Recode (status), and Cause of Death Recode (COD). The steps of Algorithm 2 explain the logic used to determine whether a breast cancer patient in the SEER dataset survived or not. This method overcomes one

drawback found in previous studies [44, 123], as the patients who died due to breast cancer after living more than 60 months since diagnosis were ignored in those studies according to their logic, although they should be classified as five-year survivors. Our final dataset included 404,576 instances for female patient cases with malignant breast tumors. We used the instances from 2004 through 2012 for training and the remainder for testing. Thus, we reserved approximately 22% of the entire dataset for testing the model performance. This split was intended to evaluate the generalization ability of the model through time, especially when survivability rates and trends tend to change over the years [46].

Algorithm 2 Logic used to define the five-year survival.

```
if status = "alive" and months < 60 then
    Drop instance
else if status = "dead" and COD ≠ "breast" then
    Drop instance
else
    if months ≥ 60 then
        survival ← True
    else
        survival ← False
    end if
end if
```

4.4 Proposed Method

In this section, we describe our proposed machine learning models for the joint survivability classification and prediction for the considered breast cancer use-case. First, we delineate the ML model selection among the candidate classifier and regressors. Then, we identify the best class balancing method for different stages of the disease. At the end, we present the proposed experimental methodology for the survivability classification and regression tasks.

4.4.1 Model Selection

Although choosing the best ML algorithm for the problem was not the main objective of this work, we conducted a pilot experiment to systematically identify the best performing classifier and regressor for our dataset out of the popular methods used in previous research work. The classification experiment included K-Nearest Neighbor (KNN), Decision Tree (C4.5), Support Vector Machine (SVM), Naive Bayes, Logistic Regression, Multi-Layer Perceptor (MLP), and Random Forest (RF). The macro-average F1 Score was chosen as the evaluation metric, because the dataset is significantly imbalanced and the F1-score can

serve as a real number evaluation metric that combines both recall and precision, and is commonly used for making decisions when dealing with skewed datasets. Also, we opted to use the macro average instead of the weighted average because the latter can be misleading as it assigns more weight to the majority class, when misclassifying instances from the minority class can be equally (or even arguably more) important in our case. We found that the Random Forest model achieved the highest average F1-score, hence it was used in all the following experiments. As for the regression task, the MLP Regressor was chosen since it achieved the lowest Root Mean Square Error (RMSE) when compared to KNN, SVM, and RF regressors. All initial experiments were conducted using built-in methods provided by the Scikit-learn [131] library of Python with the default settings.

4.4.2 Balancing Techniques

Due to the high survivability rate of breast cancer, our dataset is highly skewed towards the positive class, except for the distant summary stage where the opposite is true. Many balancing techniques were proposed to solve this problem. We experimented up-sampling the minority class using the Synthetic Minority Oversampling Technique (SMOTE) [132] and its variants; BorderLine1, BorderLine2 [133] and Adaptive Synthetic Sampling approach (ADASYN) [134]. Random down-sampling, and Cost-Sensitive Learning were also included in the comparison to select the method that yields the best performance for our problem. In the cost-sensitive learning, a weight is assigned to the minority class to increase the penalty of misclassifying it in the training step compared to classifying the majority class [135]. We employed the oversampling methods provided by the Imbalanced-learn library [136] at this step, while other techniques were implemented using the available Scikit-learn functions. For the upsampling methods, we introduced a range of upsampling rates commencing from the original ratio between minority and majority classes until reaching 1 (balanced case) with a step of 0.1. For example, when training the joint model that used instances from all stages combined, the ratio between the negative class and the positive class samples was approximately 0.1. In that case, we employed SMOTE and its variants using different upsampling strategies that increased this ratio from 0.2, 0.3, .., 1. The same concept was used for the random downsampling and cost-sensitive learning techniques, in order to determine the suitable balancing method for our prepared dataset.

To find the best balancing method for different stages, we performed the experiment using data instances from each stage separately. Table 4.2 demonstrates that the cost-sensitive learning method was able to achieve the best average F1-score for all datasets. On the other hand, the ADASYN technique failed to generate any synthetic samples in the distant-stage data space, as the algorithm only duplicates data points which are located outside homogeneous neighborhoods [134], and none were found in the distant stage. Therefore, to have a consistent approach when modeling all stages, we used the cost-sensitive learning as our bal-

ancing technique in all the following experiments. The same approach was also successfully utilized in [44] for breast cancer survivability prediction.

Table 4.2: Results of applying different balancing techniques on the data used to train the joint, localized, regional and distant models. The table compares the **macro-average F1-score** obtained by the RF classifier with default settings when using SMOTE, BorderLine1 *BL(1)*, BorderLine2 *BL(2)*, ADASYN, Random Down-sampling *RDS*, and Cost-sensitive Learning *CSL*.

	SMOTE	BL(1)	BL(2)	ADASYN	RDS	CSL
Localized	0.583	0.582	0.568	0.578	0.577	**0.598**
Regional	0.694	0.687	0.712	0.678	0.677	**0.723**
Distant	0.633	0.637	0.633	-	0.634	**0.646**

4.4.3 Experimental Methodology

4.4.3.1 Classification

To predict whether or not a breast-cancer patient will survive for five years, we built and compared two types of models: Joint and Stage-specific. The traditional joint model was trained on instances from all the stages combined, whereas the stage-specific model was exclusively built for one of the three summary stages, i.e., "Localized", "Regional", or "Distant". Other than the fact that survivability rates are greatly different for different cancer stages, the idea of developing separate classifiers based on the SEER summary stage was inspired by the work done by Kate *et al.* [44], where they concluded that the prognostic values of features are different from one stage to another. We also attempted to separate models based on other important features, such as the Grade, but no improvement in performance was achieved in this case. To develop each of our classification models, a randomized search using a three-fold cross-validation was executed to fine-tune the Random Forest hyper-parameters and to choose the class weights suitable for the cost-sensitive learning in each case. To compare the performance of the two model types, we calculated the macro-average F1-score as our main evaluation metric for test samples from each stage separately. We also report the accuracy achieved by each model for each of the three subsets.

4.4.3.2 Regression

For this task, we conducted two types of experiments. Similar to the classification task, the first experiment was to compare the performance of the joint regression model to the

stage-specific ones, when predicting the number of months left for a patient. In the second experiment, we evaluate the improvement in the regression results obtained by a two-step system that first predicts whether a patient is likely to survive for five years or not, and then estimates the number of months remaining for those who are classified as non-survivors, using a regressor that was trained on patients who died within five years from diagnosis. The two-step prediction approach was also adopted in [19] to predict the survivability of comorbid cancer cases. All models were developed using Scikit-learn Multi-Layer Preceptors, where the network size, the learning rate, and the regularization parameter were fine-tuned using randomized grid search, similar to the classification task. The maximum number of iterations used was 10,000, and the default values were employed for all other parameters for all the networks. The RMSE results of the two-step system and the traditional one-step regression with and without stage-specificity were compared for each of the three summary stages.

4.5 Results and Discussion

In this section, we evaluate the performance of our proposed ML models for breast cancer survivability classification and regression tasks, respectively.

4.5.1 Survivability Classification Results

As demonstrated in Table 4.3, the stage-specific models achieved better results than the joint-stage model for all stages. The improvement was the most significant for distant-stage instances for which the average F1-score increased by almost 11%. The regional stage did not have a large change in any of the two reported metrics, which could indicate that the model was already able to predict regional instances using patterns learned from all stages combined. However, it is worth-mentioning that the stage-specific model uses less training data, hence, requires less computational resources than the joint model, and can be more efficient in inference time when using tree-based algorithms such as Random Forest.

It can be noticed from the results how using the accuracy for evaluation is not suitable for highly-imbalanced data such as our case. The reason is that it can overestimate the classification ability of the model due to having the vast majority of instances belonging to the same class. For instance, the accuracy achieved by the joint model for localized test samples was as high as 96% while the model had a low recall score for the negative class (minority class), causing the F1-score to be significantly low.

Table 4.3: Results for joint and stage-specific models when applied to test instances of each of the three stages. The main evaluation metric is the macro-average F1-score. The corresponding accuracy is reported for reference.

	Metric	Joint	Stage-specific
Localized	**F1-score**	0.525	**0.579**
	Accuracy	0.961	0.952
Regional	**F1-score**	**0.671**	**0.669**
	Accuracy	0.833	0.832
Distant	**F1-score**	0.517	**0.625**
	Accuracy	0.705	0.665

4.5.2 Survivability Regression Results

In Table 4.4, we report the regression results for the traditional one-step joint model as the baseline model, and demonstrate the effect of introducing the stage specificity and the two-step strategies to the system. For the sake of comparison, we only show the results obtained by each of the compared systems for the correctly-classified negative test instances. Looking at the results, we clearly notice that having stage-specific models resulted in better performances for all stages. Furthermore, the RMSE dramatically dropped by integrating the classification model as a pre-regression step, to identify patients who are unlikely to survive for five years and provide them with additional information. It can also be noticed that the localized-stage instances were the hardest to predict in the one-step models, whereas patients from the distant stage had slightly higher prediction errors than other stages in the proposed two-step stage-specific models. Although the proposed method was able to achieve a significant improvement in the results, the regression errors for all models were still quite high, even after using the two-step framework. Similar findings were previously reported for comorbid-cancer patients' survivability prediction in the work conducted by Liu *et al.* [19]. This can be attributed to the fact that predicting the exact number of months left for a cancer patient is considered as a hard problem to be solved even by human domain-experts.

Table 4.4: The Root Mean Square Error (RMSE) obtained by Multi-Layer Perceptrons (MLPs) for breast cancer survival months estimation using three different systems. The results are reported for all models when applied on the same set of test samples from different summary stages.

	1-Step Joint	1-Step Stage-Specific	2-Step Stage-Specific
Localized	42.193	37.531	**14.015**
Regional	32.681	28.272	**13.939**
Distant	23.381	19.481	**15.027**

To further understand why the two-step, stage-specific model was able to obtain better estimation results, we plotted the estimated Probability Density Function (PDF) for the number of months survived by the patients who died within five years of diagnosis. The total number of training instances that did not survive was 28187, and we plotted the PDF for instances from different stages separately. As demonstrated in Fig. 4.3, the mean and variance for the distribution change from one stage to another, as the patients who are diagnosed at earlier stages tend to have higher chances of living longer. This confirms the need to model different stages separately, instead of just providing the cancer summary stage as a feature to the ML model. Another explanation for the above results is that including the samples that survived for more than 60 months, while training the one-step regression model, appears to confuse the model when it estimates the survival months remaining for the test patients. The number of survival months when considering all patients in the training set varies from zero to 191 months, which makes it harder for the model to learn the necessary trends observed for patients who die within 60 months. On the other hand, adding the classification step enables the model to better estimate the number of months left for a patient within five years, after learning the survival patterns from patients who also died within five years from diagnosis. This way, more precise prediction can be provided for critical cases who are not likely to survive for more than five years.

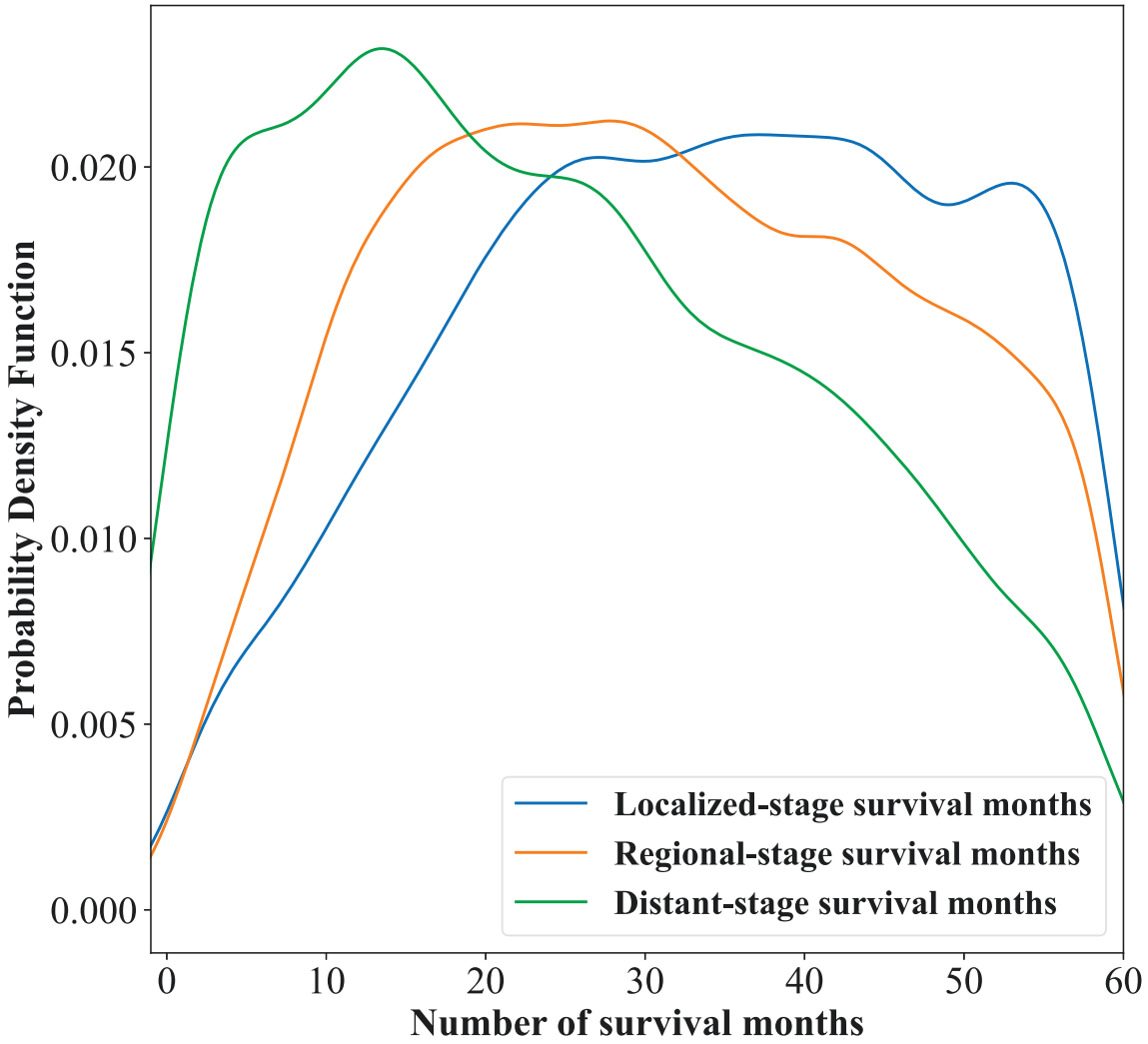


Figure 4.3: Probability Density Function of survival months for non-surviving training instances from different summary stages.

4.6 Summary

Computer-aided prognostic modeling of cancer survivability is important to help oncologists predict disease outcome and patient survival. In this paper, we developed machine learning-based predictive models to estimate survivability of breast cancer patients within five years from diagnosis. We performed our analysis on the female patients' breast-cancer incidence data in the US from the publicly available SEER repository. We split the data to train and test datasets based on the year of diagnosis, in order to investigate the generizability of our developed models over time. Unlike previous research work, our proposed model considers the survivabilty problem as a two-step problem, while ensuring cancer stage-

specific modeling at the same time. This means that the five-year survivability status of a patient is predicted at the first step; then the second step predicts the remaining lifespan of the patient if they are estimated to not survive. It also means that all models in both the classification and regression steps are stage-specific models so that predicting the survival of a patient diagnosed with a particular summary stage is performed by a model exclusively trained with incidences of the same summary stage.

Based on our experiments, it may be concluded that the two-step stage-specific system enhances the overall performance of the survival estimation for breast cancer patients. Moreover, evaluating the results for each summary stage separately revealed the differences in performance between stages, confirming the need to address the survivability problem for each stage separately. In the future, we will consider investigating possible methods that can further improve the survival time estimation accuracy for breast-cancer patients. By including more informative features, utilizing feature selection methods, employing different learning algorithms for different stages, and building hybrid deep neural networks, we will systematically investigate how the model performance may further improve. Also, as part of our future research endeavor, we will extend and generalize our proof-of-work in this paper for survivability classification coupled with survivability prediction for other multi-stage chronic diseases.

Chapter 5

Survival-Based Stage-Specific Treatment Planning Using Machine Learning Models

Despite the crucial role of prognosis in determining the most suitable treatment plans, the development of survival-based treatment planning models in clinical decision support systems has been insufficiently addressed. To overcome this issue, the chapter builds on the survival prediction methods introduced in Chapter 4 and proposes a novel framework for survival-based treatment planning, which predicts a ranked list of possible combinations of treatments associated with their estimated survival outcome. This approach aims to provide medical professionals with more detailed and intuitive treatment recommendations, enabling them to make more informed decisions about the most appropriate course of action. By incorporating the stage-specific survival prediction models into treatment planning, we first re-conduct the experiments that were previously performed on the SEER Research data for survival prediction using the treatment-inclusive SEER Research Plus data to 1) evaluate the performance gain after including different treatment fields (i.e., surgery, chemotherapy, radiation) and 2) check if the conclusions derived from previous analysis still hold on the new set of data. Second, to provide prognostic-oriented aid for the treatment planning step, we use the developed survival prediction models to design an inference treatment planning system that receives the patient's data while considering the treatment fields as variables. Then, the survival prediction model tests all different combinations of the treatment fields and ranks the predicted survivability outcomes for this patient given different treatment plans. To provide more understandable decisions for the medical team, the system provides interpretable outcomes by visualizing different features' importance and the decision path followed by the model for a specific patient.

5.1 Introduction

In clinical decision support (CDS) systems, the term computer-aided treatment recommendation refers to the use of computer algorithms to assist healthcare providers in making decisions about treatment options for patients. This technology can help clinicians to evaluate patient data, such as medical histories, laboratory results, imaging scans, and other diagnostic tests, to identify potential diagnoses and suggest appropriate treatment plans . These systems are also capable of analyzing large amounts of data from clinical trials, medical literature, and other sources to inform clinicians about the latest research findings and treatment options. The use of computer-aided treatment planning has the potential to improve patient outcomes by providing more accurate and evidence-based treatment recommendations.

AI can be used to provide real-time, data-driven clinical decision support to help doctors make better treatment decisions. By analyzing patient data and treatment outcomes, AI can help doctors identify patterns and predict which treatments are most likely to be effective for a particular patient. However, it is important to note that in order to improve the integration of these systems in medical practice, these systems should not replace human judgement and expertise and should only be used as a tool to support clinical decision-making, as its name suggests [11].

The medical community has recognized the significance of survivability prediction in determining appropriate treatment plans [6]. However, the development of survival-based

treatment planning models in clinical decision support (CDS) systems has been largely neglected in previous studies [48–50]. Considering survivability as an important prognostic factor in making treatment decisions, it can help doctors determine whether to opt for aggressive therapies or focus on palliative care [29]. This approach can help healthcare providers deliver personalized care to patients while also providing realistic expectations of the available treatment options to patients and their families. Therefore, it is crucial to design treatment recommendation systems that utilize survival prediction to suggest treatment pathways.

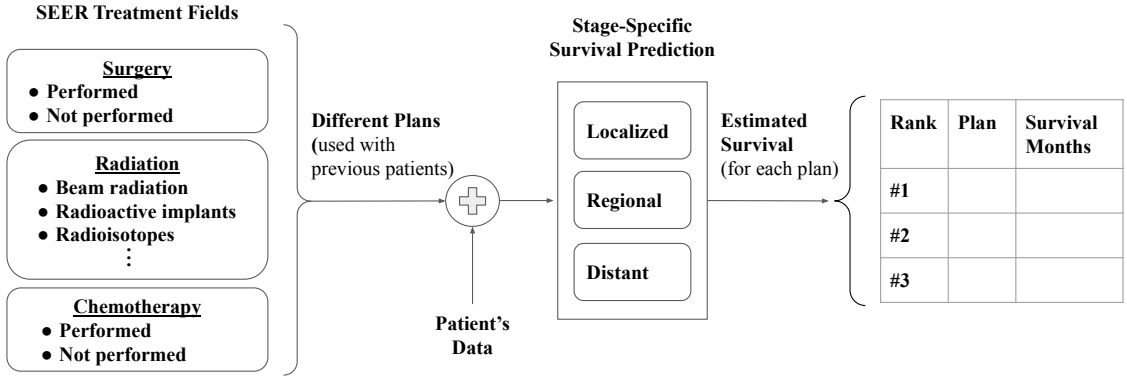


Figure 5.1: The proposed system for prognostic-based treatment planning using stage-specific survival prediction models.

In this chapter, by employing the SEER Research Plus cancer data [54] and the survival prediction models introduced in Chapter 4, a new framework for survival-based treatment planning for breast cancer patients is proposed to address this issue. As shown in Figure 5.1, the system takes the patient's data and different possible sets of treatment combinations as inputs and estimates the survival outcome of each plan using our stage-specific survival prediction models. The input treatment options are predefined by extracting all possible treatment plans used with previous patients in the training dataset. Then, instead of only providing a single recommended treatment for medical professionals as done in previous treatment recommendation systems [48–50], our system generates an ordered recommendation list of possible treatment plans with the best survival outcomes for a specific patient as its output. This approach is aimed at providing more detailed and intuitive prognostic-based recommendations, which can assist medical professionals in making more informed decisions regarding the best course of treatment for their patients. By incorporating survival prediction into treatment planning, medical professionals can provide better patient care and help patients and their families be involved in making decisions about their treatment options.

5.2 Related Work

Treatment planning is a crucial step that can determine the disease outcome for cancer patients. Recently, many studies investigated the usage of Artificial Intelligence techniques in recommending patient-specific cancer treatment plans. Since radiation therapy is one of the most common treatments for cancer, many researchers investigated the use of machine learning (ML)-based models to predict the optimal radiation parameters to use for a certain patient based on their available data [48–50]. On the other hand, [51] developed different machine learning models, including linear models, support vector machines, tree-based models, and neural networks to predict quality assurance (QA) outcomes in intensity modulated radiation therapy (IMRT) planning. This QA is important to validate the dose calculation accuracy and verify that the plan can be delivered as intended on the treatment machine.

Another objective that is usually considered during the treatment planning process is predicting the response of the patient's body to a specific treatment plan. Based on our review of the literature, we found some studies that applied ML-based models on medical data from different modalities for treatment response prediction. The authors in [137] applied Random Forest classifiers (RF) on magnetic resonance (MR) scans to predict the treatment response using multivariate delta-radiomic features for locally advanced rectal cancer (LARC) patients treated by neoadjuvant chemoradiation therapy (nCRT). In [138], however, they used clinical data, including patient characteristics, mutations, and laboratory findings, from the electronic medical records to develop an AI–based clinical decision support algorithm that predicts if PD-1 inhibitors therapy results in complete, partial or no clinical response in lung cancer treatment. By employing both MR imaging and tabular clinical data, the authors in [139] created a framework for predicting therapeutic outcomes of transarterial chemoembolization using ML techniques.

Despite having several studies and research projects focused on the use of AI in cancer treatment planning, we identified some of the limitations existing in current decision support systems for cancer treatment in order to address them in our work:

- First, although many therapy approaches can be considered while deciding on a cancer treatment plan (e.g. surgery, chemotherapy, radiation, etc.), all of the above-mentioned studies only considered one type of therapy in their treatment outcome prediction models.
- Moreover, including the survival prediction aspect in cancer treatment planning was greatly overlooked in previous CDS systems, which calls for the need to design a new survival-based system.
- In addition to that, most of the proposed treatment outcome models used black-box

architectures which made the automated decisions hard to understand or rely on by oncologists and healthcare providers.

In this work, we propose the usage of survival prediction ML models on breast cancer clinical data for patient-specific treatment planning while considering combinations of different methods of therapy. We also provide additional explanatory information along with the decision generated by the model to make it interpretable by the medical team. To the best of our knowledge, this is the first work to approach the cancer treatment recommendation task through survival prediction ML models.

5.3 Data Preparation

In this section, we describe the dataset selection and preprocessing phases that are required for devloping the machine learning models and evaluating the treatment planning framework.

5.3.1 Dataset Selection

Similar to the work done in the previous chapter, we continued employing the clinical data provided by the Surveillance, Epidemiology, and End Results (SEER) repository [54] for breast cancer incidence collected by cancer registries in the US. However, in this chapter, we requested access to the SEER Research Plus data package, which includes all cancer treatment information available for the patients to be able to carry out the analysis required in this chapter. After our request was approved and the data use agreement was signed, we imported the November 2021 version of the SEER Research Plus database using the SEER* Stat software (version 8.4.0.1) [127].

To provide consistent analysis with the work done in the previous chapter, we followed the same approach in extracting the cancer incidents of females diagnosed with malignant tumors. In addition to the features selected from the basic dataset as listed in 4.1, we added the therapy-related attributes that were only available in the Research Plus version. The appended features are displayed in Table 5.1 along with their corresponding possible values and the count of incidents of each value in our dataset. As explained by the database documentation, *Reason no cancer-directed surgery* states whether or not a surgery was performed to treat the cancer along with the reason if a surgery was not performed. *Radiation recode* indicates the method of radiation therapy performed as part of the first course of treatment, whereas *Chemotherapy recode* simply records whether chemotherapy was given or not. The features used to calculate the target variable for the survivability prediction model were the same as explained in Chapter 4.

Table 5.1: The treatment-related attributes from the SEER Research Plus database used in our analysis and their possible values. These attributes are added on top of the attributes listed in 4.1, which were available in the SEER Research data. All names are listed as provided by the SEER Repository.

Attribute name	**Values**	**Count**
Reason no cancer-directed surgery	Surgery performed	371783
	Not recommended	4357
	Recommended but not performed, patient refused	314
	Recommended but not performed, unknown reason	173
	Not recommended, contraindicated due to other cond	139
Radiation recode	Beam radiation	196633
	None/Unknown	162687
	Radioactive implants (includes brachytherapy) (1988+)	9650
	Refused (1988+)	4451
	Radiation, NOS method or source not specified	2562
	Combination of beam with implants or isotopes	622
	Radioisotopes (1988+)	161
Chemotherapy recode	No/Unknown	203589
	Yes	173177

5.3.2 Preprocessing

For the pre-processing stage, we first followed the same cleaning steps explained in the previous chapter to remove the records with missing values in the cancer-related attributes employed in both chapters as well as the surgery-related field. However, we did not remove the records containing Unknown values in the radiation or chemotherapy fields since for each of these fields, patients who did not receive the treatment, as well as the ones whose status is unknown, are all assigned the same category named *None/Unknown*. Hence, it is impossible to differentiate between the incidents where the treatment was not given from the unknown ones. Removing records with unknown values, in that case, would inevitably remove valuable records that represent an important category in our analysis (i.e., not giving a certain treatment).

Algorithm 3 Feature engineering of cancer-related treatment fields.

input: Treatment attributes: 1)Reason no cancer-directed surgery: *surgery*, 2) Radiation recode: *radiation*, and 3) Chemotherapy recode: *chemotherapy*
output: Performed and recommended treatment fields.
if *surgery* = "Surgery performed" **then**
    $surgeryPerformed \leftarrow True$
    $surgeryRecommended \leftarrow True$
else if *surgery* = "Recommended but not performed, patient refused" **or** "Recommended but not performed, unknown reason" **then**
    $surgeryPerformed \leftarrow False$
    $surgeryRecommended \leftarrow True$
else
    $surgeryPerformed \leftarrow False$
    $surgeryRecommended \leftarrow False$
end if
if *radiation* = "None/unknown" **then**
    $radiationPerformed \leftarrow False$
    $radiationRecommended \leftarrow False$
else
    $radiationPerformed \leftarrow True$
    $radiationRecommended \leftarrow True$
end if
if *chemotherapy* = "Yes" **then**
    $chemotherapyPerformed \leftarrow True$
    $chemotherapyRecommended \leftarrow True$
else
    $chemotherapyPerformed \leftarrow False$
    $chemotherapyRecommended \leftarrow False$
end if

Then, we performed feature engineering steps on the treatment fields to create abstract treatment fields to be used in training and testing our survival-based treatment planning framework. Since some of the three original treatment fields have a wide range of values and some of these values indicate that a treatment is recommended but not performed, we prepared three features indicating whether or not a treatment was recommended and another three features indicating whether or not a treatment was performed. We call these features: performed treatment fields and recommended treatment fields, respectively, until the end of this chapter. The performed treatment fields are employed in training and testing our survival prediction models, whereas the recommended treatment fields are intended to evaluate the accuracy of the treatment plans recommended by our framework. Algorithm 3 explains how different treatment fields were mapped to the corresponding engineered fields. In the next section, we explain how these attributes are used to develop and evaluate our proposed system. Our final preprocessed dataset included 376,766 instances, and they were

split to training and testing subsets based on the year of diagnosis similar to chapter 4.

5.4 Methodology

In this section, we describe our proposed methods for cancer treatment recommendation using survival prediction ML models. We designed a treatment planning system using our stage-specific survival prediction models to provide treatment recommendations ranked by their survivability outcomes. In Figure 5.2, we show the pipeline we followed to design our survival-based treatment planning system, including the development, inference and evaluation phases.

5.4.1 Development

First, in the training phase, the survival prediction models are developed using our training dataset extracted from the SEER Research Plus data. Although our proposed stage-specific survival prediction framework proved to outperform other traditional frameworks for breast cancer survival estimation in Chapter 4, we still need to ensure that this behavior holds after including the performed-treatment fields. In other words, before directly plugging the proposed survival prediction system we need to make sure that it is the best-performing model with the new set of treatment-inclusive features. Moreover, we aim to investigate how the inclusion of treatment information affects survival prediction performance in different frameworks (i.e., whether or not there is a performance gain from including these features).

Therefore, we re-run our previously-performed experiments after including the *performed-treatment fields* prepared in the preprocessing step based on the treatment information provided by the SEER Research Plus data, as shown in section 5.3.2. We adopted the same workflow used in Chapter 4 to develop different frameworks for survival prediction with the same fine-tuning techniques. However, we only incorporated cost-sensitive learning as our balancing technique since it proved to be the most efficient method in all stages. We experimented with applying stage-specific models as well as joint models in the classification step. Then, we trained and tested three different types of regression systems: one-step joint, one-step stage-specific, two-step stage-specific. We evaluated the results using the same evaluation metrics as the previous chapter to compare the survival prediction performance with and without treatment information.

In addition to that, the training set is also used in the development phase to prepare a list of all valid combinations of the three performed treatment fields that were used with training instances. This is done to avoid considering invalid combinations of treatment methods that are not used in real life. For demonstration purposes, let's assume that the valid options of treatment plans are four plans named A, B, C and D, as shown in 5.2. This list is passed to the inference phase to be used in the decision-making process. In

our training dataset, there were eight different combinations of treatment methods used with patients in different stages of breast cancer. Hence, our predefined list included eight treatment plans.

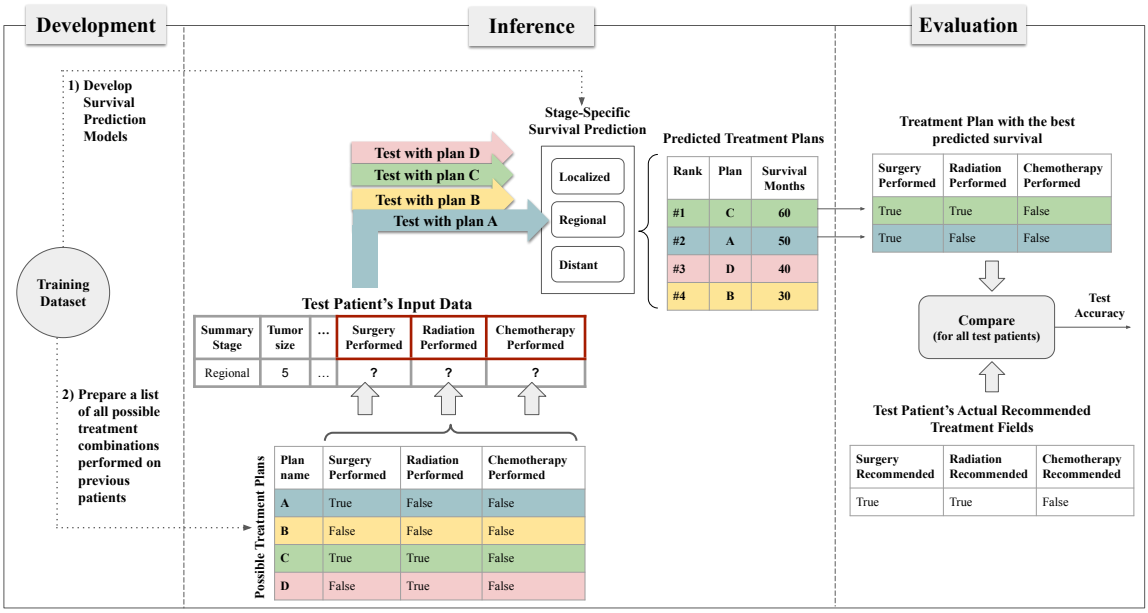


Figure 5.2: Different phases of the designing process of our proposed survival-based treatment planning system.

5.4.2 Inference

To predict the treatment plan that has the best prognostic outcome for a test patient, we employ this patient's clinical data and consider the treatment fields as variables. Then, the predefined list of treatment options, prepared in the previous phase, is used, where each plan option in the list is combined with the test patient's data to be passed as input to the survival prediction model. For example, for the four valid plans shown in Figure 2.2, the survivability model is applied four times on the patient's data, each time with a different set of values inserted in the corresponding *performed-treatment fields* for this patient. In each inference run, the survival prediction model estimates an answer to the following question: "Assuming that this treatment plan is used with this particular patient, how long will this patient live?". These four inference runs result in four survival estimates, one for each plan. The plans are then ranked based on their survival outcomes and displayed associated with their estimated survival months to aid the medical professional in making the final decision.

In addition to that, our survival prediction models provide additional output attributes related to the model decision-making process, along with the survival months predicted for a certain plan. This is intended to improve the transparency of the CDS outcomes for

clinicians and help them understand how the model arrived at a particular decision. We used the Random Forest's explanation attributes, namely, the decision path and feature importance, for this purpose. We show examples of the provided outputs in the results section.

5.4.3 Evaluation

In order to evaluate the accuracy of our survival-based treatment planning system, we employ the recommended-treatment fields prepared in the data preprocessing step from the SEER Research Plus data, as shown in section 5.3.2. For each test instance, we first compare all the values in the surgery, radiation and chemotherapy fields of the first-recommended plan generated by our system with the corresponding fields in the actual recommended treatment plans in the test dataset. If all the values are the same, we consider the system-recommended first plan as correct. Then, we follow the same steps to evaluate the accuracy of the second recommended plan to determine how close the automated recommendations are to the expert-recommended plans for all test patients. This also aims to evaluate whether or not considering more than one plan as the recommended output of the CDS system helps improve the reliability of the automated recommendation. Finally, to analyze the performance in different stages, we calculate the accuracy of our treatment planning system for test patients diagnosed with breast cancer in localized, regional and distant stages separately.

5.5 Results and Discussion

In this section, we first evaluate different classification frameworks and different regression frameworks when applying them to the SEER Plus data that includes treatment features in order to 1) check if it matches the results from the previous experiments and 2) evaluate the performance gain by including the treatment information. Then, we examine the performance of the treatment planning system in inference time by comparing its output with the treatment recommended for a patient. Finally, we present examples of the visual explanations provided for the predictions made by our proposed system.

5.5.1 Survival Prediction

First, to examine the performance of each of the different frameworks for survival classification, we present their results in table 5.2. We can see that stage-specific models still prove superior to joint models in performing survival classification after adding the treatment features in both localized and distant subgroups while achieving almost the same results for the regional-stage test instances. This aligns with the trend observed in the previous

chapter, where similar findings were reported in table 4.3. The reason we considered the macro-average F1 Score as the main evaluation metric for the survival classification step is that the dataset is significantly imbalanced, and the accuracy can be misleading, as shown in the results. The F1 score serves as a real-number evaluation metric that measures the harmonic average of both recall and precision. The precision metric indicates the percentage of positive identifications that were actually correct, whereas the recall measures the rate at which the positive instances were correctly classified. By calculating the macro-average F1-score, the precision and recall of both the majority and minority classes are equally evaluated, which can efficiently indicate how well the model predicts the overall survivability for surviving and non-surviving patients. The results show that the stage-specific classifier was able to outperform the stage-agnostic one with the SEER Research Plus data.

Table 5.2: Results for joint and stage-specific models with the use of treatment features when applied to test instances of each of the three stages. Results are reported in terms of the macro-average F1-score and accuracy.

	Metric	Joint	Stage-specific
Localized	**F1-score**	0.552	**0.579**
	Accuracy	0.957	0.953
Regional	**F1-score**	**0.691**	**0.685**
	Accuracy	0.824	0.839
Distant	**F1-score**	0.583	**0.635**
	Accuracy	0.695	0.678

As for the regression task, the averaged results obtained by the survival regression models are compared when using different frameworks (i.e., 1-step joint, 1-step stage-specific, and 2-step stage-specific) and with different sets of data attributes (i.e., SEER and SEER Plus)in Figure 5.3. From the graph, it can be observed that similar to the classification task, moving from joint modeling to stage-specific modeling also helped achieve better regression results when applied to the new dataset (red graph). Moreover, adding the 2-step design significantly lowered the rmse of the regression system. On the other hand, when we compare these results with previous results obtained from the SEER data (plotted in blue), we can see that the introduction of the treatment fields made a slight enhancement in the overall performance of the regression models in both types of the 1-step system. However, it can be noticed that no improvement was achieved by adding these attributes to the 2-step regression framework. This observation may suggest that treatment attributes were an

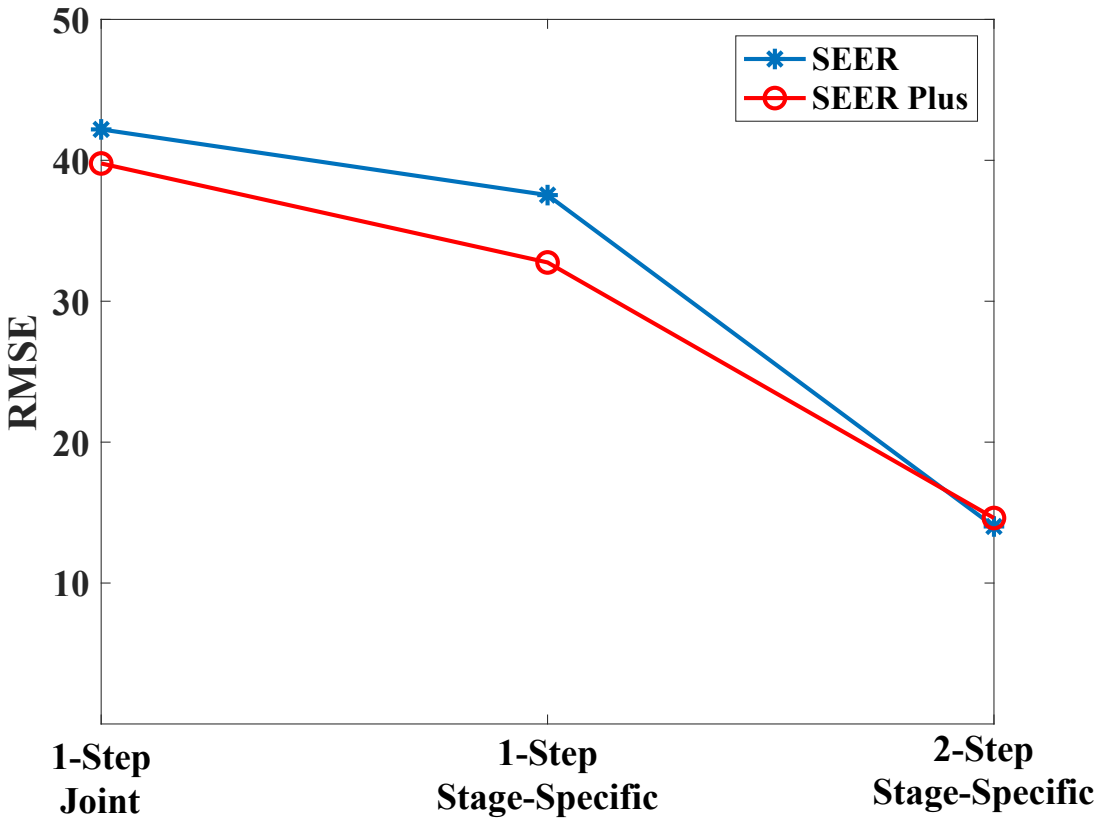


Figure 5.3: Survival prediction results using different frameworks, including the proposed two-step stage-specific model when applied to the SEER data and the SEER Plus data. Each line graph shows the regression results obtained by one framework in terms of root mean square error (rmse). One marker on the line plot represents the rmse error resulting from one of the three compared models when used with one of the two employed datasets.

addition to the survival estimation task only in learning patterns that differentiate between prognostic outcomes of patients on a large scale, such as the case in the 1-step model, where patients who are surviving for more than five years are combined with those who are not surviving. On the other hand, in the two-step regression model, there was no additional gain from those attributes when the learning instances were all from the same class (the non-surviving class).

5.5.2 Treatment Planning

In this subsection, we show the evaluation results of the treatment planning system by comparing the outputs of the system with the expert-recommended treatment plans as explained in section 5.4.3. The average results for applying the proposed system on the 82,300 patients in our test dataset, are shown in Table 5.3. First, we can see that considering the first

two plans significantly improved the accuracy of the system output for all stages of breast cancer. The accuracy for the first two plans slightly varies among different stages, with the highest two-plan accuracy achieved for localized-stage patients at a rate of 88.3%. On the other hand, the recommendation accuracy of the first plan was the highest (63.4%) for patients diagnosed with regional-stage breast cancer. This can indicate that the survivability outcomes of regional-stage patients are highly correlated with the recommended treatments for these patients, which allowed the model to correctly predict these plans as they were estimated to achieve the longest survival time. This was not the case with distant-stage patients, where the treatment recommendation performance was only reasonable when the first two plans were considered in calculating the accuracy. On average, we can conclude that for around 84% of the test patients, the first two patient-specific plans recommended by the system included a plan that was actually recommended by the medical professionals for that patient.

Table 5.3: The average test accuracy of the first two treatment plans recommended by the proposed survival-based treatment planning system for breast cancer patients. The results are averaged over all patients in the test set and reported for different stages separately.

Stage	First plan	First two plans
Localized	59.62%	**88.33%**
Regional	63.42%	**80.64%**
Distant	42.95%	**84.71%**

Since this work is believed to be the first to provide survival-based treatment planning while considering multiple types of treatment options, we were not able to compare our results with any previous work. However, it is clear from the results that the system's overall prediction accuracy is significantly higher than a "random guess" classifier accuracy, which is estimated to be around 0.125 since we have eight possible classes (i.e., eight valid treatment plans existing in our dataset). The accuracy is also higher by a huge margin than the accuracy of a model that always predicts the majority treatment plan, which is 0.178 in our case.

To understand how employing stage-specific survival prediction was able to achieve the above results, we analyzed the frequency of different treatments used with patients who survived for more than 5 years at different stages of the disease. We found that the frequency of different treatment approaches was greatly impacted by the stage at which the disease was diagnosed. For example, as shown in Figure 5.4, the majority of localized-stage cancer patients who survived were not reported to receive chemotherapy, whereas this was not the case with both the regional and distant stages, where more than 73% of the surviving

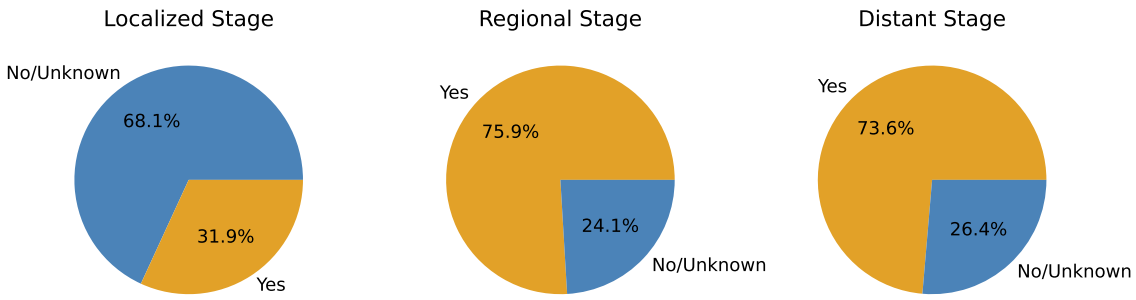


Figure 5.4: The frequency of using Chemotherapy with breast cancer patients who survived for more than five years. The frequency is plotted for patients diagnosed at localized, regional and distant summary stages.

patients received chemotherapy. This emphasizes that stage-specific models can be essential to achieve better recommendation accuracy compared to stage-agnostic prediction models.

5.5.3 Explainability Via Visualization

In this subsection, we shed light on the explainability measures we employed in our system to improve the transparency of the predicted outcomes. First, we provide a bar chart that shows the importance of each feature in predicting the 5-year survivability of patients from a certain stage. We used the Gini impurity-based importance in the Random Forest model, which is computed as the normalized reduction of the classifier accuracy brought by that feature [131]. The importance values are percentage values that add up to 1. For example, figure 5.5 shows the importance of different features in the 5-year survivability classifier trained on breast cancer patients from the distant stage. The figure shows that the tumor size, age, and Lymph Node Ratio are the three most important features considered by the distant-stage survivability classifier. This method provides a model-level explanation for the survival results, which can help the clinician understand which attributes are considered the most important by the CDS system.

Although the feature importance plot can give a human-like understanding of the model in general, it does not provide a patient-specific explanation for the prediction made for a certain patient. Therefore, we also provide a visual explanation for the decision path followed by a decision tree in our model to predict the survival outcome of a specific patient. This is done by randomly selecting and visualizing one of the decision trees in our RF classifier that contributed to the output class (majority-voted) for a certain test instance. Then, using the *decision˙path* attribute provided by Scikit Learn, we highlight the tree nodes that were visited by this test instance to arrive at the predicted 5-year survival outcome.

In figure 5.6, we show an example of a decision path for predicting the 5-year survivability of one patient from the distant stage. For demonstration purposes, we only employ

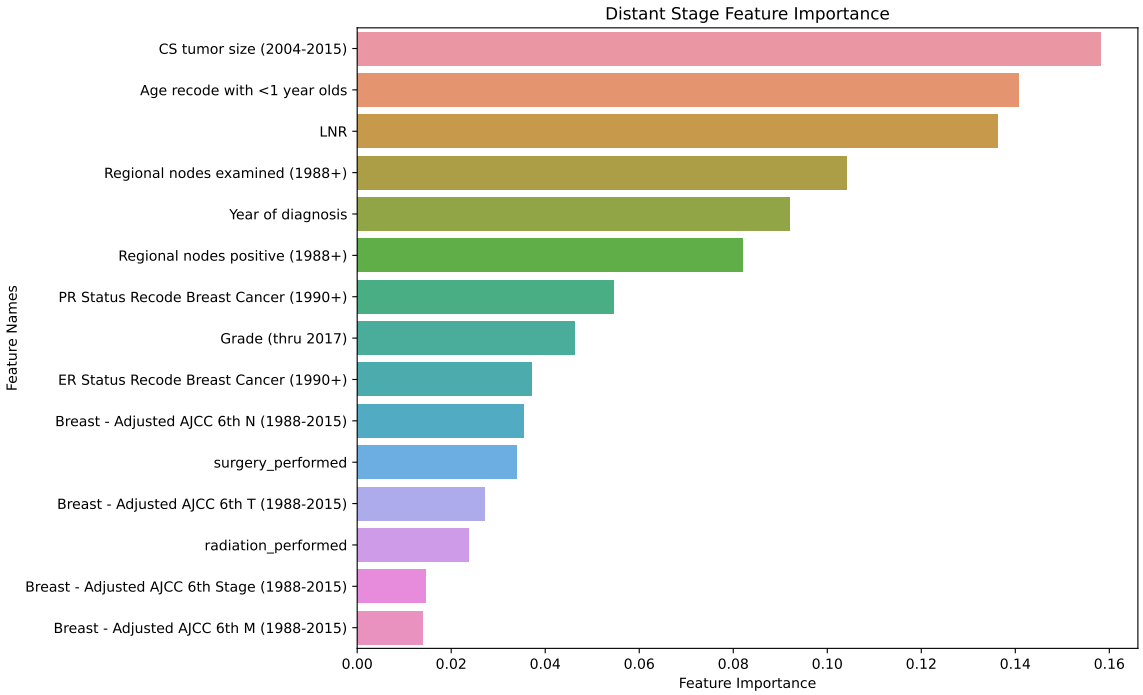


Figure 5.5: The proposed system for prognostic-based treatment planning using stage-specific survival prediction models.

the top three important features in the distant-stage classifier identified from the feature importance chart, along with the three treatment fields used in our analysis. Each node contains information about the split condition, the number of training samples belonging to each class, the Gini impurity, and the majority class at this node. In this example, the decision path shows how this patient was predicted to survive for five years based on their tumor size, age, Lymph Node Ratio (LNR) and no surgery performed. This explains why the first two recommended treatment plans generated by the system did not include performing surgery. This can help medical professionals decide which plan to proceed with based on the rationale provided by the decision path and the number of surviving training instances that shared similar features and followed the recommended treatment plan. It can also help identify when the model goes against established medical knowledge by tracing the decision path followed for a specific patient.

5.6 Summary

In conclusion, Computer-aided treatment recommendation for cancer is important to help oncologists decide which treatment plan to follow with a specific patient. In this chapter, we proposed a survival-based treatment planning system to provide patient-specific treatment

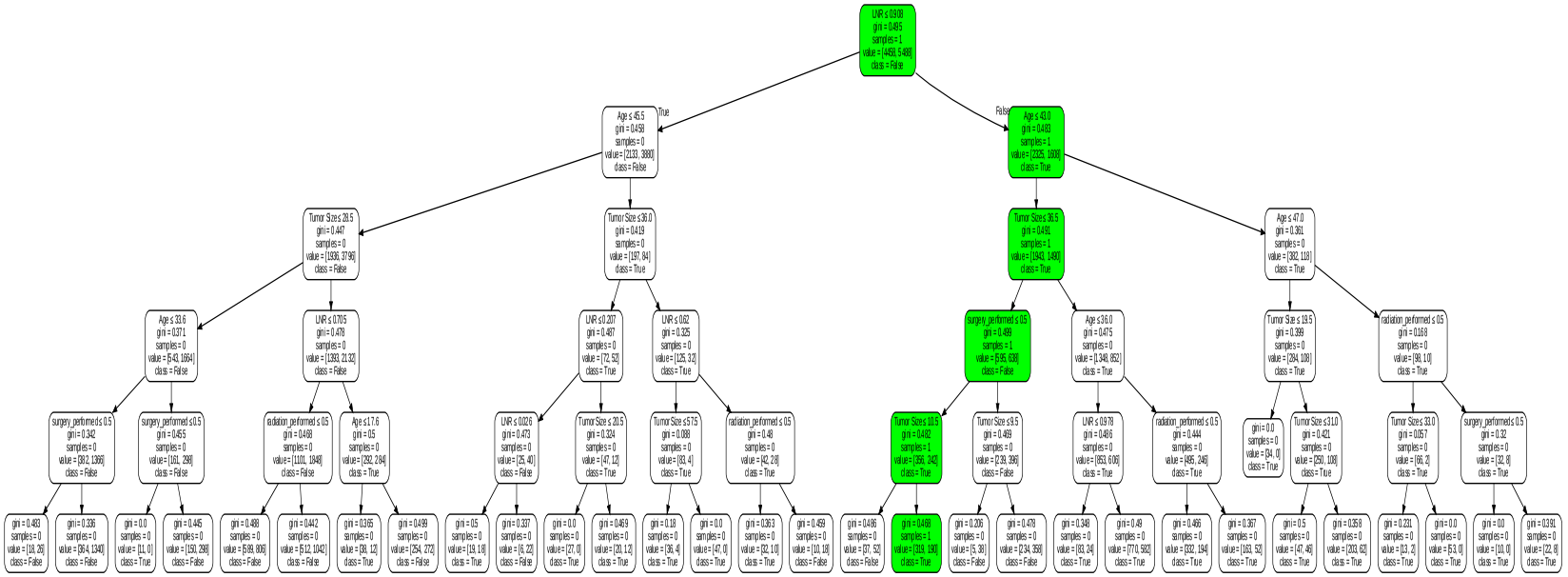


Figure 5.6: A snapshot from a decision tree showing the decision path followed for predicting the 5-year survivability of a test patient diagnosed in the distant stage. The tree is generated using the three most important features for a distatnt-stage classifier: age, tumor size and Lymph Node Ratio (LNR). Each node contains multiple indicators (*gini*: the Gini impurity score, *samples*: the number of test samples considered in calculating the decision path, *value*: the number of training samples belonging to each of the 5-year survival classes, *class*: the 5-year survival class for the majority of training samples in this node.)

recommendations based on the estimated prognostic outcomes. We employed the SEER Plus dataset that provides treatment information for the female patients' breast-cancer incidence data in the US. First, we compared the different machine learning-based frameorks developed to predict the survivability of breast cancer patients after including information about the treatment history of the patients to ensure the superiority of the two-step stage-specific prediction. Then, we proposed a new inference system for treatment planning that receives different combinations of possible options coupled with patients' clinical data to generate a ranked list of recommended plans based on the predicted survival outcome of each plan.

Based on the conducted experiments, first, it can be concluded that stage-specific modelling performs better than the traditional models in performing survival classification and regression regardless of the presence or absence of treatment information. Moreover, adding the treatment features is found to generally enhance the performance of the predictive model when estimating the number of survival months remaining for a patient. As for the proposed treatment recommendation system, the results showed to be in favour of the designed system as both the disease stage and prognostic outcome were shown to be highly correlated with the treatment plan recommended by the medical practitioners for a specific patient. This confirms the viability of using a stage-specific prognostic-based treatment planning framework to provide detailed information about recommended treatment plans associated with the projected prognosis.

Chapter 6

Conclusions and Future Works

In this thesis, we address challenges existing in different components of Clinical Decision Support systems that contribute to the lack of their trustworthiness. This chapter summarizes the contributions of the dissertation work and manifests potential future research directions.

6.1 Contributions

In this thesis, we present novel approaches to address challenges and research gaps in diagnostic, prognostic and treatment planning components of a CDS system to improve its trustworthiness and make steps toward its integration into clinical practice. The breast cancer use case is chosen as it is well-suited for different components of our analysis, and it is the most prevalent cancer among women and the second leading cause of cancer death.

In Chapter 3, to address the problem of decaying resolution in traditional networks used for medical image semantic segmentation in CAD systems, we propose a double-dilated convolution module to preserve spatial resolution and improve the performance of mass segmentation. We evaluate different loss functions to address the pixel-level class imbalance problem in mammogram screenings. To overcome the lack of explainability in existing segmentation networks, we employ and evaluate different explainability methods and provide visualized explanations for the segmented outputs. Experimental analysis shows the effectiveness of the proposed module in increasing the similarity score and reducing the miss-detection rate.

In Chapter 4, to improve the precision of predicted survivability of breast cancer patients, we propose a new framework for survival prediction in CAP systems. We model the survivability prediction task as a two-step problem and develop stage-specific models to learn survivability patterns based on the stage at which the disease was diagnosed. We investigate the impact of adopting different strategies for balancing techniques on model

performance using breast cancer clinical data. Experimental results show that the two-step stage-specific system enhances the overall performance of the survival estimation for breast cancer patients. Moreover, evaluating the results for each summary stage separately revealed the differences in performance between stages, confirming the need to address the survivability problem for each stage separately.

In Chapter 5, to account for the prognostic role in treatment planning, we propose a novel survival-based framework for treatment planning that employs prediction models developed for stage-specific survival estimation to determine the best possible treatment plans based on the prognostic outcomes. The system predicts a ranked list of possible treatment combinations based on their predicted survival outcomes to aid medical professionals in making informed treatment decisions. Finally, we provide visualized explanations for the predicted outcomes to ensure transparent decision-making.

Our experiments demonstrate that the proposed AI-enabled techniques have potential applications in future Clinical Decision Support Systems to assist clinicians in making patient-specific assessments and treatment decisions.

6.2 Future Directions

In this sub-section, we conclude the thesis by shedding light on some possible future research directions:

- We will experiment with adopting the proposed double-dilated convolution module in large datasets with different medical image modalities to further verify its effectiveness in performing different segmentation tasks. Also, the concept of multi-resolution dilated convolution can be extended to develop an N-dilated convolution module which employs a kernel with N sparsity factors on different scales.

- We intend to investigate the phenomenon of high false positive rates associated with CAD systems in order to spare patients from the negative psychological impact and unnecessary biopsies.

- We will consider investigating possible methods that can further improve the survival time estimation accuracy. By utilizing feature selection methods, employing different learning algorithms for different stages, and building hybrid deep neural networks, we will systematically investigate how the model performance may further improve.

- We intend to study the development of an API that allows for seamless integration with different EHR systems. We will also investigate employing federated learning techniques to allow for efficient and private integration with healthcare systems.

Bibliography

[1] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking Atrous Convolution for Semantic Image Segmentation," doi:10.48550/arXiv.1706.05587. [Online]. Available: http://arxiv.org/abs/1706.05587

[2] I. Sim, P. Gorman, R. A. Greenes, R. B. Haynes, B. Kaplan, H. Lehmann, and P. C. Tang, "Clinical Decision Support Systems for the Practice of Evidence-based Medicine," *Journal of the American Medical Informatics Association*, vol. 8, no. 6, pp. 527–534, Nov. 2001, doi:10.1136/jamia.2001.0080527. [Online]. Available: https://doi.org/10.1136/jamia.2001.0080527

[3] R. T. Sutton, D. Pincock, D. C. Baumgart, D. C. Sadowski, R. N. Fedorak, and K. I. Kroeker, "An overview of clinical decision support systems: benefits, risks, and strategies for success," *npj Digital Medicine*, vol. 3, no. 1, pp. 1–10, Feb. 2020, doi:10.1038/s41746-020-0221-y. [Online]. Available: https://www.nature.com/articles/s41746-020-0221-y

[4] K. Doi, "Computer-aided diagnosis in medical imaging: Historical review, current status and future potential," *Computerized Medical Imaging and Graphics*, vol. 31, no. 4, pp. 198–211, Jun. 2007, doi:10.1016/j.compmedimag.2007.02.002. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0895611107000262

[5] H.-P. Chan, L. M. Hadjiiski, and R. K. Samala, "Computer-aided diagnosis in the era of deep learning," *Medical Physics*, vol. 47, no. 5, pp. e218–e227, 2020, doi:10.1002/mp.13764. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/mp.13764

[6] T. M. Gill, "THE CENTRAL ROLE OF PROGNOSIS IN CLINICAL DECISION MAKING," *Jama*, vol. 307, no. 2, pp. 199–200, Jan. 2012, doi:10.1001/jama.2011.1992. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3273781/

[7] A. Madabhushi, S. Agner, A. Basavanhally, S. Doyle, and G. Lee, "Computer-aided prognosis: Predicting patient and disease outcome via quantitative fusion of

multi-scale, multi-modal data," *Computerized Medical Imaging and Graphics*, vol. 35, no. 7, pp. 506–514, Oct. 2011, doi:10.1016/j.compmedimag.2011.01.008. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S089561111100019X

[8] E. Abd, N. Intan, M. Bouchahma, and K. Musa, "Framework for a Computer-Aided Treatment Prediction (CATP) System for Breast Cancer," *Intelligent Automation & Soft Computing*, vol. 36, no. 3, pp. 3007–3028, doi:10.32604/iasc.2023.032580. [Online]. Available: https://www.techscience.com/iasc/v36n3/51875

[9] B. Meskó and M. Görög, "A short guide for medical professionals in the era of artificial intelligence," *npj Digital Medicine*, vol. 3, no. 1, pp. 1–8, Sep. 2020, doi:10.1038/s41746-020-00333-z. [Online]. Available: https://www.nature.com/articles/s41746-020-00333-z

[10] G. Baselli, M. Codari, and F. Sardanelli, "Opening the black box of machine learning in radiology: can the proximity of annotated cases be a way?" *European Radiology Experimental*, vol. 4, p. 30, May 2020, doi:10.1186/s41747-020-00159-0. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7200961/

[11] A. M. Froomkin, I. R. Kerr, and J. Pineau, "When AIs Outperform Doctors: Confronting the Challenges of a Tort-Induced Over-Reliance on Machine Learning," Rochester, NY, Feb. 2019, doi:10.2139/ssrn.3114347. [Online]. Available: https://papers.ssrn.com/abstract=3114347

[12] J. Osheroff, J. Teich, D. Levick, L. Saldana, F. Velasco, D. Sittig, K. Rogers, and R. Jenders, *Improving Outcomes with Clinical Decision Support: An Implementer's Guide, Second Edition*, Feb. 2012, doi:10.4324/9781498757461.

[13] E. S. Berner, Ed., *Clinical decision support systems: theory and practice*, 2nd ed., ser. Health informatics. New York, NY: Springer, 2007.

[14] J. Quinlan, *C4. 5: programs for machine learning*. Elsevier, 2014.

[15] T. K. Ho, "Random decision forests," in *Proceedings of 3rd International Conference on Document Analysis and Recognition*, vol. 1, pp. 278–282 vol.1, doi:10.1109/ICDAR.1995.598994.

[16] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, doi:10.1007/BF00994018. [Online]. Available: https://doi.org/10.1007/BF00994018

[17] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *The bulletin of mathematical biophysics*, vol. 5, no. 4, pp. 115–133, doi:10.1007/BF02478259. [Online]. Available: https://doi.org/10.1007/BF02478259

[18] E. Choi, M. T. Bahadori, J. Sun, J. Kulas, A. Schuetz, and W. Stewart, "RETAIN: An Interpretable Predictive Model for Healthcare using Reverse Time Attention Mechanism," in *Advances in Neural Information Processing Systems*, vol. 29. Curran Associates, Inc. [Online]. Available: https://proceedings.neurips.cc/paper/2016/hash/231141b34c82aa95e48810a9d1b33a79-Abstract.html

[19] P. Liu and S. Fei, "Two-Stage Prediction of Comorbid Cancer Patient Survivability Based on Improved Infinite Feature Selection," *IEEE Access*, vol. 8, pp. 169 559–169 567, 2020, doi:10.1109/ACCESS.2020.3016998.

[20] J. Teng, A. Abdygametova, J. Du, B. Ma, R. Zhou, Y. Shyr, and F. Ye, "Bayesian Inference of Lymph Node Ratio Estimation and Survival Prognosis for Breast Cancer Patients," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 2, pp. 354–364, Feb. 2020, doi:10.1109/JBHI.2019.2943401.

[21] M. N. Haque, T. Tazin, M. M. Khan, S. Faisal, S. M. Ibraheem, H. Algethami, and F. A. Almalki, "Predicting Characteristics Associated with Breast Cancer Survival Using Multiple Machine Learning Approaches," *Computational and Mathematical Methods in Medicine*, vol. 2022, p. e1249692, doi:10.1155/2022/1249692. [Online]. Available: https://www.hindawi.com/journals/cmmm/2022/1249692/

[22] K. Huang, J. Zhang, Y. Yu, Y. Lin, and C. Song, "The impact of chemotherapy and survival prediction by machine learning in early Elderly Triple Negative Breast Cancer (eTNBC): A population based study from the SEER database," *BMC Geriatrics*, vol. 22, no. 1, p. 268, doi:10.1186/s12877-022-02936-5. [Online]. Available: https://doi.org/10.1186/s12877-022-02936-5

[23] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, p. d. u. family=Laak, given=Jeroen A. W. M., p. u. family=Ginneken, given=Bram, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, doi:10.1016/j.media.2017.07.005. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841517301135

[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, doi:10.1145/3065386. [Online]. Available: https://dl.acm.org/doi/10.1145/3065386

[25] I. o. M. U. N. C. P. Forum, *Supply and Demand in the Health Care Workforce*. National Academies Press (US), 2009, publication Title: Ensuring Quality Cancer Care through the Oncology Workforce: Sustaining Care in the 21st Century:

Workshop Summary. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK215247/

[26] R. R. Kontham, A. K. Kondoju, M. M. Fouda, and Z. M. Fadlullah, "An End-To-End Explainable AI System for Analyzing Breast Cancer Prediction Models," in *2022 IEEE International Conference on Internet of Things and Intelligence Systems (IoTaIS)*, Nov. 2022, pp. 402–407, doi:10.1109/IoTaIS56727.2022.9975896.

[27] V. Pitroda, M. M. Fouda, and Z. M. Fadlullah, "An Explainable AI Model for Interpretable Lung Disease Classification," in *2021 IEEE International Conference on Internet of Things and Intelligence Systems (IoTaIS)*, Nov. 2021, pp. 98–103, doi:10.1109/IoTaIS53735.2021.9628573.

[28] J. Amann, A. Blasimme, E. Vayena, D. Frey, V. I. Madai, and the Precise4Q consortium, "Explainability for artificial intelligence in healthcare: a multidisciplinary perspective," *BMC Medical Informatics and Decision Making*, vol. 20, no. 1, p. 310, Nov. 2020, doi:10.1186/s12911-020-01332-6. [Online]. Available: https://doi.org/10.1186/s12911-020-01332-6

[29] O. R. Maarsingh, H. E. van der Horst, and D. a. W. M. van der Windt, "[Clinical decision-making: from diagnosis to prognosis-oriented approach]," *Nederlands Tijdschrift Voor Geneeskunde*, vol. 165, p. D6064, Jun. 2021.

[30] R. Kunhimangalam, S. Ovallath, and P. K. Joseph, "A clinical decision support system with an integrated EMR for diagnosis of peripheral neuropathy," *Journal of Medical Systems*, vol. 38, no. 4, p. 38, Apr. 2014, doi:10.1007/s10916-014-0038-9.

[31] S. Razzaki, A. Baker, Y. Perov, K. Middleton, J. Baxter, D. Mullarkey, D. Sangar, M. Taliercio, M. Butt, A. Majeed, A. DoRosario, M. Mahoney, and S. Johri, "A comparative study of artificial intelligence and human doctors for the purpose of triage and diagnosis," Jun. 2018, doi:10.48550/arXiv.1806.10698. [Online]. Available: http://arxiv.org/abs/1806.10698

[32] A. I. Martinez-Franco, M. Sanchez-Mendiola, J. J. Mazon-Ramirez, I. Hernandez-Torres, C. Rivero-Lopez, T. Spicer, and A. Martinez-Gonzalez, "Diagnostic accuracy in Family Medicine residents using a clinical decision support system (DXplain): a randomized-controlled trial," *Diagnosis (Berlin, Germany)*, vol. 5, no. 2, pp. 71–76, Jun. 2018, doi:10.1515/dx-2017-0045.

[33] M. H. Hesamian, W. Jia, X. He, and P. Kennedy, "Deep Learning Techniques for Medical Image Segmentation: Achievements and Challenges," *Journal of Digital*

Imaging, vol. 32, no. 4, pp. 582–596, Aug. 2019, doi:10.1007/s10278-019-00227-x. [Online]. Available: https://doi.org/10.1007/s10278-019-00227-x

[34] F. Yu and V. Koltun, "Multi-Scale Context Aggregation by Dilated Convolutions," *arXiv:1511.07122 [cs]*, Apr. 2016.

[35] G. Lin, C. Shen, A. van den Hengel, and I. Reid, "Efficient Piecewise Training of Deep Structured Models for Semantic Segmentation," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV: IEEE, Jun. 2016, pp. 3194–3203, doi:10.1109/CVPR.2016.348. [Online]. Available: https://ieeexplore.ieee.org/document/7780717/

[36] M. H. Hesamian, W. Jia, X. He, and P. Kennedy, "Deep Learning Techniques for Medical Image Segmentation: Achievements and Challenges," *Journal of Digital Imaging*, vol. 32, no. 4, pp. 582–596, Aug. 2019, doi:10.1007/s10278-019-00227-x. [Online]. Available: https://doi.org/10.1007/s10278-019-00227-x

[37] E. Tappeiner, M. Welk, and R. Schubert, "Tackling the class imbalance problem of deep learning-based head and neck organ segmentation," *International Journal of Computer Assisted Radiology and Surgery*, vol. 17, no. 11, pp. 2103–2111, Nov. 2022, doi:10.1007/s11548-022-02649-5. [Online]. Available: https://doi.org/10.1007/s11548-022-02649-5

[38] C.-M. Nam, J. Kim, and K. J. Lee, "Lung nodule segmentation with convolutional neural network trained by simple diameter information." [Online]. Available: https://openreview.net/pdf?id=r1ib989jG

[39] A. Singh, S. Sengupta, and V. Lakshminarayanan, "Explainable Deep Learning Models in Medical Image Analysis," *Journal of Imaging*, vol. 6, no. 6, p. 52, doi:10.3390/jimaging6060052. [Online]. Available: https://www.mdpi.com/2313-433X/6/6/52

[40] P. Kim, J. M. Daly, M. A. Berry-Stoelzle, M. E. Schmidt, L. C. Michaels, D. A. Dorr, and B. T. Levy, "Prognostic Indices for Advance Care Planning in Primary Care: A Scoping Review," *Journal of the American Board of Family Medicine : JABFM*, vol. 33, no. 2, pp. 322–338, doi:10.3122/jabfm.2020.02.190173. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7772823/

[41] T. Iwasawa, A. Asakura, F. Sakai, T. Kanauchi, T. Gotoh, T. Ogura, T. Yazawa, J. Nishimura, and T. Inoue, "Assessment of Prognosis of Patients With Idiopathic Pulmonary Fibrosis by Computer-aided Analysis of CT Images," *Journal of Thoracic*

Imaging, vol. 24, no. 3, p. 216, Aug. 2009, doi:10.1097/RTI.0b013e3181a6527d. [Online]. Available: https://journals.lww.com/thoracicimaging/Abstract/2009/08000/Assessment_of_Prognosis_of_Patients_With.10.aspx

[42] O. Sertel, J. Kong, H. Shimada, U. V. Catalyurek, J. H. Saltz, and M. N. Gurcan, "Computer-aided prognosis of neuroblastoma on whole-slide images: Classification of stromal development," *Pattern Recognition*, vol. 42, no. 6, pp. 1093–1103, Jun. 2009, doi:10.1016/j.patcog.2008.08.027. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320308003439

[43] J.-M. Chen, Y. Li, J. Xu, L. Gong, L.-W. Wang, W.-L. Liu, and J. Liu, "Computer-aided prognosis on breast cancer with hematoxylin and eosin histopathology images: A review," *Tumor Biology*, vol. 39, no. 3, p. 1010428317694550, Mar. 2017, doi:10.1177/1010428317694550. [Online]. Available: https://doi.org/10.1177/1010428317694550

[44] R. J. Kate and R. Nadig, "Stage-specific predictive models for breast cancer survivability," *International Journal of Medical Informatics*, vol. 97, pp. 304–311, Jan. 2017, doi:10.1016/j.ijmedinf.2016.11.001.

[45] "Survival Rates for Breast Cancer." [Online]. Available: https://www.cancer.org/cancer/breast-cancer/understanding-a-breast-cancer-diagnosis/breast-cancer-survival-rates.html

[46] R. Kleinlein and D. Riaño, "Persistence of data-driven knowledge to predict breast cancer survival," *International Journal of Medical Informatics*, vol. 129, pp. 303–311, Sep. 2019, doi:10.1016/j.ijmedinf.2019.06.018.

[47] J. L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, and Y. Kluger, "DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network," *BMC Medical Research Methodology*, vol. 18, no. 1, p. 24, Feb. 2018, doi:10.1186/s12874-018-0482-1. [Online]. Available: https://doi.org/10.1186/s12874-018-0482-1

[48] S. H. Ahn, E. Kim, C. Kim, W. Cheon, M. Kim, S. B. Lee, Y. K. Lim, H. Kim, D. Shin, D. Y. Kim, and J. H. Jeong, "Deep learning method for prediction of patient-specific dose distribution in breast cancer," *Radiation Oncology*, vol. 16, no. 1, p. 154, Aug. 2021, doi:10.1186/s13014-021-01864-9. [Online]. Available: https://doi.org/10.1186/s13014-021-01864-9

[49] Y. Sheng, T. Li, S. Yoo, F.-F. Yin, R. Blitzblau, J. K. Horton, Y. Ge, and Q. J. Wu, "Automatic Planning of Whole Breast Radiation Therapy Using

Machine Learning Models," *Frontiers in Oncology*, vol. 9, 2019. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fonc.2019.00750

[50] N. Bakx, H. Bluemink, E. Hagelaar, M. van der Sangen, J. Theuws, and C. Hurkmans, "Development and evaluation of radiotherapy deep learning dose prediction models for breast cancer," *Physics and Imaging in Radiation Oncology*, vol. 17, pp. 65–70, Jan. 2021, doi:10.1016/j.phro.2021.01.006. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2405631621000063

[51] P. D. H. Wall and J. D. Fontenot, "Application and comparison of machine learning models for predicting quality assurance outcomes in radiation therapy treatment planning," *Informatics in Medicine Unlocked*, vol. 18, p. 100292, Jan. 2020, doi:10.1016/j.imu.2020.100292. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352914819303661

[52] (02/21/2023 - 07:00) Mammograms - NCI. National Cancer Institute (NCI). [Online]. Available: https://www.cancer.gov/types/breast/mammograms-fact-sheet

[53] (02/02/2011 - 07:00) Definition of heterogeneously dense breast tissue - NCI Dictionary of Cancer Terms - NCI. National Cancer Institute (NCI). [Online]. Available: https://www.cancer.gov/publications/dictionaries/cancer-terms/def/heterogeneously-dense-breast-tissue

[54] "Surveillance, epidemiology, and end results (seer) program (www.seer.cancer.gov) seer*stat database: Incidence - seer research data, 8 registries, nov 2021 sub (1975-2019) - linked to county attributes - time dependent (1990-2019) income/rurality, 1969-2020 counties, national cancer institute, dccps, surveillance research program, released april 2022, based on the november 2021 submission."

[55] "Canadian Cancer Statistics Advisory Committee in collaboration with the Canadian Cancer Society, Statistics Canada and the Public Health Agency of Canada. Canadian Cancer Statistics 2021. Toronto, ON: Canadian Cancer Society; 2021." http://cancer.ca/Canadian-Cancer-Statistics-2021-EN.

[56] M. Arnold, E. Morgan, H. Rumgay, A. Mafra, D. Singh, M. Laversanne, J. Vignat, J. R. Gralow, F. Cardoso, S. Siesling, and I. Soerjomataram, "Current and future burden of breast cancer: Global statistics for 2020 and 2040," *The Breast*, vol. 66, pp. 15–23, Dec. 2022, doi:10.1016/j.breast.2022.08.010. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0960977622001448

[57] S. Misra, N. L. Solomon, F. L. Moffat, and L. G. Koniaris, "Screening Criteria for Breast Cancer," in *Advances in Surgery*, Sep. 2010. [Online]. Available: https://www.us.elsevierhealth.com/advances-in-surgery-2010-9780323068239.html

[58] J. L. Diffey, "A comparison of digital mammography detectors and emerging technology," *Radiography*, vol. 21, no. 4, pp. 315–323, Nov. 2015, doi:10.1016/j.radi.2015.06.007. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S107881741500084X

[59] I. C. Moreira, I. Amaral, I. Domingues, A. Cardoso, M. J. Cardoso, and J. S. Cardoso, "INbreast: Toward a Full-field Digital Mammographic Database," *Academic Radiology*, vol. 19, no. 2, pp. 236–248, Feb. 2012, doi:10.1016/j.acra.2011.09.014.

[60] A. G. Waks and E. P. Winer, "Breast Cancer Treatment: A Review," *JAMA*, vol. 321, no. 3, pp. 288–300, Jan. 2019, doi:10.1001/jama.2018.19323. [Online]. Available: https://doi.org/10.1001/jama.2018.19323

[61] Access Data and Reports — CIHI. Canadian Institute For Health Information. [Online]. Available: https://www.cihi.ca/en/access-data-and-reports

[62] Epic, Microsoft partner to use generative AI for better EHRs. Healthcare IT News. [Online]. Available: https://www.healthcareitnews.com/news/epic-microsoft-partner-use-generative-ai-better-ehrs

[63] S.-J. Heo, Y. Kim, S. Yun, S.-S. Lim, J. Kim, C.-M. Nam, E.-C. Park, I. Jung, and J.-H. Yoon, "Deep Learning Algorithms with Demographic Information Help to Detect Tuberculosis in Chest Radiographs in Annual Workers' Health Examination Data," *International Journal of Environmental Research and Public Health*, vol. 16, no. 2, p. 250, doi:10.3390/ijerph16020250. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6352082/

[64] J. Konečný, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon. Federated Learning: Strategies for Improving Communication Efficiency. Doi:10.48550/arXiv.1610.05492. [Online]. Available: http://arxiv.org/abs/1610.05492

[65] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural computation*, vol. 1, no. 4, pp. 541–551, 1989.

[66] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[67] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[68] M. Holschneider, R. Kronland-Martinet, J. Morlet, and P. Tchamitchian, "A real-time algorithm for signal analysis with the help of the wavelet transform," in *Wavelets*. Springer, 1990, pp. 286–297.

[69] P. Wang, P. Chen, Y. Yuan, D. Liu, Z. Huang, X. Hou, and G. Cottrell, "Understanding convolution for semantic segmentation," in *2018 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2018, pp. 1451–1460.

[70] Y. Ho and S. Wookey, "The Real-World-Weight Cross-Entropy Loss Function: Modeling the Costs of Mislabeling," *IEEE Access*, vol. 8, pp. 4806–4813, 2020, doi:10.1109/ACCESS.2019.2962617.

[71] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation," in *2016 Fourth International Conference on 3D Vision (3DV)*, Oct. 2016, pp. 565–571, doi:10.1109/3DV.2016.79.

[72] S. S. M. Salehi, D. Erdogmus, and A. Gholipour, "Tversky loss function for image segmentation using 3d fully convolutional deep networks," *CoRR*, vol. abs/1706.05721, 2017. [Online]. Available: http://arxiv.org/abs/1706.05721

[73] P. Dardouillet, A. Benoit, E. Amri, P. Bolon, D. Dubucq, and A. Crédoz, "Explainability of Image Semantic Segmentation Through SHAP Values," in *ICPR-XAIE -26TH International Conference on Pattern Recognition 2-nd Workshop on Explainable and Ethical AI*, Montreal, Canada, Aug. 2022. [Online]. Available: https://hal.science/hal-03719597

[74] M. Abukmeil, A. Genovese, V. Piuri, F. Rundo, and F. Scotti, "Towards Explainable Semantic Segmentation for Autonomous Driving Systems by Multi-Scale Variational Attention," in *2021 IEEE International Conference on Autonomous Systems (ICAS)*, Aug. 2021, pp. 1–5, doi:10.1109/ICAS49788.2021.9551172.

[75] J. Mendes, J. Domingues, H. Aidos, N. Garcia, and N. Matela, "AI in Breast Cancer Imaging: A Survey of Different Applications," *Journal of Imaging*, vol. 8, no. 9, p. 228, Sep. 2022, doi:10.3390/jimaging8090228.

[76] X. Yu, Q. Zhou, S. Wang, and Y.-D. Zhang, "A systematic survey of deep learning in breast cancer," *International Journal of Intelligent Systems*, vol. 37, no. 1, pp. 152–216, 2022, doi:10.1002/int.22622.

[77] J. Long, E. Shelhamer, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[78] X. He, R. S. Zemel, and M. A. Carreira-Perpinán, "Multiscale conditional random fields for image labeling," in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, vol. 2. IEEE, 2004, pp. II–II.

[79] J. Yao, S. Fidler, and R. Urtasun, "Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 702–709.

[80] L. Pei, L. Vidyaratne, M. M. Rahman, and K. M. Iftekharuddin, "Context aware deep learning for brain tumor segmentation, subtype classification, and survival prediction using radiology images," *Scientific Reports*, vol. 10, no. 1, p. 19726, Nov. 2020, doi:10.1038/s41598-020-74419-9.

[81] P. F. Christ, F. Ettlinger, F. Grün, M. E. A. Elshaer, J. Lipková, S. Schlecht, F. Ahmaddy, S. Tatavarty, M. Bickel, P. Bilic, M. Rempfler, F. Hofmann, M. D'Anastasi, S. Ahmadi, G. Kaissis, J. Holch, W. H. Sommer, R. Braren, V. Heinemann, and B. H. Menze, "Automatic liver and tumor segmentation of CT and MRI volumes using cascaded fully convolutional neural networks," *CoRR*, vol. abs/1702.05970, 2017. [Online]. Available: http://arxiv.org/abs/1702.05970

[82] M. Ben naceur, M. Akil, R. Saouli, and R. Kachouri, "Fully automatic brain tumor segmentation with deep learning-based selective attention using overlapping patches and multi-class weighted cross-entropy," *Medical Image Analysis*, vol. 63, p. 101692, Jul. 2020, doi:10.1016/j.media.2020.101692.

[83] S. Lu, F. Gao, C. Piao, and Y. Ma, "Dynamic Weighted Cross Entropy for Semantic Segmentation with Extremely Imbalanced Data," in *2019 International Conference on Artificial Intelligence and Advanced Manufacturing (AIAM)*, Oct. 2019, pp. 230–233, doi:10.1109/AIAM48774.2019.00053.

[84] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. J. Cardoso, "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," *CoRR*, vol. abs/1707.03237, 2017. [Online]. Available: http://arxiv.org/abs/1707.03237

[85] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should I trust you?": Explaining the predictions of any classifier," *CoRR*, vol. abs/1602.04938, 2016. [Online]. Available: http://arxiv.org/abs/1602.04938

[86] G. Montavon, S. Bach, A. Binder, W. Samek, and K. Müller, "Explaining nonlinear classification decisions with deep taylor decomposition," *CoRR*, vol. abs/1512.02479, 2015. [Online]. Available: http://arxiv.org/abs/1512.02479

[87] S. Lapuschkin, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation," *PLoS ONE*, vol. 10, pp. 130–140, doi:10.1371/journal.pone.0130140.

[88] V. Pitroda, M. M. Fouda, and Z. M. Fadlullah, "An Explainable AI Model for Interpretable Lung Disease Classification," in *2021 IEEE International Conference on Internet of Things and Intelligence Systems (IoTaIS)*, pp. 98–103, doi:10.1109/IoTaIS53735.2021.9628573.

[89] V. Couteaux, O. Nempont, G. Pizaine, and I. Bloch, "Towards Interpretability of Segmentation Networks by Analyzing DeepDreams," in *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support*. Springer International Publishing, vol. 11797, pp. 56–63, doi:10.1007/978-3-030-33850-3˙7.

[90] P. Lakhani, "The Importance of Image Resolution in Building Deep Learning Models for Medical Imaging," *Radiology: Artificial Intelligence*, vol. 2, no. 1, p. e190177, Jan. 2020, doi:10.1148/ryai.2019190177.

[91] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1520–1528.

[92] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3640–3649.

[93] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs," *CoRR*, vol. abs/1412.7062, Jun. 2016. [Online]. Available: http://arxiv.org/abs/1412.7062

[94] Z. Wu, C. Shen, and A. van den Hengel, "Bridging category-level and instance-level semantic image segmentation," *CoRR*, vol. abs/1605.06885, 2016. [Online]. Available: http://arxiv.org/abs/1605.06885

[95] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *CoRR*, vol. abs/1606.00915, 2016. [Online]. Available: http://arxiv.org/abs/1606.00915

[96] L. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," *CoRR*, vol. abs/1802.02611, 2018. [Online]. Available: http://arxiv.org/abs/1802.02611

[97] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal Visual Object Classes Challenge: A Retrospective," *International Journal of Computer Vision*, vol. 111, no. 1, pp. 98–136, Jan. 2015, doi:10.1007/s11263-014-0733-5.

[98] R. Hussein, S. Lee, R. Ward, and M. J. McKeown, "Semi-dilated convolutional neural networks for epileptic seizure prediction," *Neural Networks*, vol. 139, no. Complete, pp. 212–222, 2021, doi:10.1016/j.neunet.2021.03.008.

[99] L. W. Bassett, K. Conner, and I. Ms, "The Abnormal Mammogram," *Holland-Frei Cancer Medicine. 6th edition*, 2003. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK12642/

[100] S. Jadon, "A survey of loss functions for semantic segmentation," in *2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, Oct. 2020, pp. 1–7, doi:10.1109/CIBCB48159.2020.9277638.

[101] I. J. Good, "Rational Decisions," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 14, no. 1, pp. 107–114, 1952.

[102] M. Yi-de, L. Qing, and Q. Zhi-bai, "Automated image segmentation using improved PCNN model based on cross-entropy," in *Proceedings of 2004 International Symposium on Intelligent Multimedia, Video and Speech Processing, 2004.*, Oct. 2004, pp. 743–746, doi:10.1109/ISIMP.2004.1434171.

[103] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *CoRR*, vol. abs/1505.04597, 2015. [Online]. Available: http://arxiv.org/abs/1505.04597

[104] M. Masood, T. Nazir, M. Nawaz, A. Mehmood, J. Rashid, H.-Y. Kwon, T. Mahmood, and A. Hussain, "A Novel Deep Learning Method for Recognition and Classification of Brain Tumors from MRI Images," *Diagnostics*, vol. 11, no. 5, p. 744, May 2021, doi:10.3390/diagnostics11050744.

[105] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," *International Journal of Computer Vision*, vol. 128, no. 2, pp. 336–359, doi:10.1007/s11263-019-01228-7. [Online]. Available: http://arxiv.org/abs/1610.02391

[106] M. D. Zeiler and R. Fergus, "Visualizing and Understanding Convolutional Networks," in *Computer Vision – ECCV 2014*, ser. Lecture Notes in Computer Science, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Springer International Publishing, pp. 818–833, doi:10.1007/978-3-319-10590-1˙53.

[107] J. Kauffmann, K.-R. Müller, and G. Montavon, "Towards Explaining Anomalies: A Deep Taylor Decomposition of One-Class Models," *Pattern Recognition*, vol. 101, pp. 107–198, doi:10.1016/j.patcog.2020.107198.

[108] W. Samek, G. Montavon, S. Lapuschkin, C. J. Anders, and K.-R. Müller, "Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications," *Proceedings of the IEEE*, vol. 109, no. 3, pp. 247–278, doi:10.1109/JPROC.2021.3060483. [Online]. Available: http://arxiv.org/abs/2003.07631

[109] Z. Wang, E. Wang, and Y. Zhu, "Image segmentation evaluation: A survey of methods," *Artificial Intelligence Review*, vol. 53, no. 8, pp. 5637–5674, Dec. 2020, doi:10.1007/s10462-020-09830-9.

[110] L. R. Dice, "Measures of the Amount of Ecologic Association Between Species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945, doi:10.2307/1932409.

[111] P. Jaccard, "The Distribution of the Flora in the Alpine Zone.1," *New Phytologist*, vol. 11, no. 2, pp. 37–50, 1912, doi:10.1111/j.1469-8137.1912.tb05611.x.

[112] A. A. Taha and A. Hanbury, "Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool," *BMC Medical Imaging*, vol. 15, no. 1, p. 29, Aug. 2015, doi:10.1186/s12880-015-0068-x.

[113] N. Dhungel, G. Carneiro, and A. P. Bradley, "Automated Mass Detection in Mammograms Using Cascaded Deep Learning and Random Forests," in *2015 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, Nov. 2015, pp. 1–8, doi:10.1109/DICTA.2015.7371234.

[114] R. M. Nishikawa, M. Kallergi, and C. G. Orton, "Computer-aided detection, in its present form, is not an effective aid for screening mammography," *Medical Physics*, vol. 33, no. 4, pp. 811–814, 2006, doi:10.1118/1.2168063.

[115] K. Loizidou, R. Elia, and C. Pitris, "Computer-aided breast cancer detection and classification in mammography: A comprehensive review," *Computers in Biology and Medicine*, vol. 153, p. 106554, Feb. 2023, doi:10.1016/j.compbiomed.2023.106554.

[116] S. Ciatto, N. Houssami, D. Bernardi, F. Caumo, M. Pellegrini, S. Brunelli, P. Tuttobene, P. Bricolo, C. Fantò, M. Valentini, S. Montemezzi, and P. Macaskill, "Integration of 3D digital mammography with tomosynthesis for population breast-cancer screening (STORM): a prospective comparison study," *The Lancet Oncology*, vol. 14, no. 7, pp. 583–589, Jun. 2013, doi:10.1016/S1470-2045(13)70134-7. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1470204513701347

[117] L. Rieger, P. Chormai, G. Montavon, L. K. Hansen, and K.-R. Müller, "Structuring Neural Networks for More Explainable Predictions," in *Explainable and Interpretable Models in Computer Vision and Machine Learning*, ser. The Springer Series on Challenges in Machine Learning, H. J. Escalante, S. Escalera, I. Guyon, X. Baró, Y. Güçlütürk, U. Güçlü, and p. u. family=Gerven, given=Marcel, Eds. Springer International Publishing, pp. 115–131, doi:10.1007/978-3-319-98131-4˙5. [Online]. Available: https://doi.org/10.1007/978-3-319-98131-4_5

[118] **© 2022 IEEE. Reprinted, with permission, from A. Farrag, Z. M. Fadlullah, and M. M. Fouda, "A Two-Step Machine Learning Model for Stage-Specific Disease Survivability Prediction," in 2022 IEEE International Conference on Internet of Things and Intelligence Systems (IoTaIS), Nov. 2022.**

[119] D. Delen, G. Walker, and A. Kadam, "Predicting breast cancer survivability: A comparison of three data mining methods," *Artificial intelligence in medicine*, vol. 34, pp. 113–27, Jul. 2005, doi:10.1016/j.artmed.2004.07.002.

[120] Y. Wang, D. Wang, X. Ye, Y. Wang, Y. Yin, and Y. Jin, "A tree ensemble-based two-stage model for advanced-stage colorectal cancer survival prediction," *Information Sciences*, vol. 474, pp. 106–124, Feb. 2019, doi:10.1016/j.ins.2018.09.046.

[121] "Summary Stage 2018 - SEER," National Cancer Institute (NCI). [Online]. Available: https://seer.cancer.gov/tools/ssm/

[122] S. Haykin, *Neural networks: a comprehensive foundation*. Prentice Hall PTR, 1994.

[123] A. Bellaachia and E. Guven, "Predicting breast cancer survivability using data mining techniques," *Department of Computer Science. George Washington University. Washington DC*, jan 2006.

[124] D. Solti and H. Zhai, "Predicting Breast Cancer Patient Survival Using Machine Learning," in *Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics*, ser. BCB'13. New

York, NY, USA: Association for Computing Machinery, Sep. 2013, pp. 704–705, doi:10.1145/2506583.2512376.

[125] S. Hussain, N. Z. Quazilbash, S. Bai, and S. Khoja, "Reduction of Variables for Predicting Breast Cancer Survivability Using Principal Component Analysis," in *2015 IEEE 28th International Symposium on Computer-Based Medical Systems*, Jun. 2015, pp. 131–134, doi:10.1109/CBMS.2015.62.

[126] I. C. Tee and A. H. Gazala, "A novel breast cancer prediction system," in *2011 International Symposium on Innovations in Intelligent Systems and Applications*, 2011, pp. 621–625.

[127] "Surveillance research program, national cancer institute seer*stat software (seer.cancer.gov/seerstat) version 8.4.0.1."

[128] A. S. Sarvestani, A. A. Safavi, N. Parandeh, and M. Salehi, "Predicting breast cancer survivability using data mining techniques," in *2010 2nd International Conference on Software Technology and Engineering*. San Juan, PR, USA: IEEE, Oct. 2010, p. 5608818, doi:10.1109/ICSTE.2010.5608818.

[129] Z. Sedighi-Maman and A. Mondello, "A two-stage modeling approach for breast cancer survivability prediction," *International Journal of Medical Informatics*, vol. 149, p. 104438, May 2021, doi:10.1016/j.ijmedinf.2021.104438.

[130] A. Z. Dag, Z. Akcam, E. Kibis, S. Simsek, and D. Delen, "A probabilistic data analytics methodology based on Bayesian Belief network for predicting and understanding breast cancer survival," *Knowledge-Based Systems*, vol. 242, p. 108407, Apr. 2022, doi:10.1016/j.knosys.2022.108407.

[131] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[132] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, Jun. 2002, doi:10.1613/jair.953.

[133] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-smote: A new over-sampling method in imbalanced data sets learning," in *Advances in Intelligent Computing*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 878–887, doi: 10.1007/11538059˙91.

[134] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, Jun. 2008, pp. 1322–1328, doi:10.1109/IJCNN.2008.4633969.

[135] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, and F. Herrera, *Cost-Sensitive Learning*. Cham: Springer International Publishing, 2018, pp. 63–78.

[136] G. Lemaître, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning," *Journal of Machine Learning Research*, vol. 18, no. 17, pp. 1–5, 2017.

[137] S. Shayesteh, M. Nazari, A. Salahshour, S. Sandoughdaran, G. Hajianfar, M. Khateri, A. Yaghobi Joybari, F. Jozian, S. H. Fatehi Feyzabad, H. Arabi, I. Shiri, and H. Zaidi, "Treatment response prediction using MRI-based pre-, post-, and delta-radiomic features and machine learning algorithms in colorectal cancer," *Medical Physics*, vol. 48, no. 7, pp. 3691–3701, 2021, doi:10.1002/mp.14896. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/mp.14896

[138] B.-C. Ahn, J.-W. So, C.-B. Synn, T. H. Kim, J. H. Kim, Y. Byeon, Y. S. Kim, S. G. Heo, S.-D. Yang, M. R. Yun, S. Lim, S.-J. Choi, W. Lee, D. K. Kim, E. J. Lee, S. Lee, D.-J. Lee, C. G. Kim, S. M. Lim, M. H. Hong, B. C. Cho, K.-H. Pyo, and H. R. Kim, "Clinical decision support algorithm based on machine learning to assess the clinical response to anti–programmed death-1 therapy in patients with non–small-cell lung cancer," *European Journal of Cancer*, vol. 153, pp. 179–189, Aug. 2021, doi:10.1016/j.ejca.2021.05.019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0959804921003282

[139] A. Abajian, N. Murali, L. J. Savic, F. M. Laage-Gaupp, N. Nezami, J. S. Duncan, T. Schlachter, M. Lin, J.-F. Geschwind, and J. Chapiro, "Predicting Treatment Response to Intra-arterial Therapies for Hepatocellular Carcinoma with the Use of Supervised Machine Learning—An Artificial Intelligence Concept," *Journal of Vascular and Interventional Radiology*, vol. 29, no. 6, pp. 850–857.e1, Jun. 2018, doi:10.1016/j.jvir.2018.01.769. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1051044318307930