Hypothesizing Precise Cancer Treatments Based on Patient Survival Using Machine Learning & Deep Learning

by

Hasibul Sakib
19101283
Aditto Baidya Alok
19101509
Fardin Huq
22241141
Shamsil Arafin Ullah
19101164
Riya Ghosh
19101327

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
September 2023

© 2023. Brac University
All rights reserved.

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Hasibul Sakib
19101283

Aditto Baidya Alok
19101509

Fardin Huq
22241141

Shamsil Arafin Ullah
19101164

Riya Ghosh
19101327

Approval

The thesis/project titled "Hypothesizing Precise Cancer Treatments Based on Patient Survival Using Machine Learning & Deep Learning" was submitted by

1. Hasibul Sakib (19101283)
2. Aditto Baidya Alok (19101509)
3. Fardin Huq (22241141)
4. Shamsil Arafin Ullah (19101164)
5. Riya Ghosh (19101327)

Of Summer, 2023 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on September 21, 2023.

Examining Committee:

Supervisor:
(Member)

Arif Shakil
Lecturer
Department of Computer Science and Engineering
BRAC University

Program Coordinator:
(Member)

Md. Golam Rabiul Alam,Phd
Professor
Department of Computer Science and Engineering
BRAC University

Head of Department:
(Chair)

Sadia Hamid Kazi,Phd
Chairperson and Associate Professor
Department of Computer Science and Engineering
BRAC University

Ethics Statement

This research project upholds the highest ethical standards in conducting the study on cancer treatment and prediction. It ensures informed consent, privacy protection, data security, impartiality, responsible use of findings, research integrity, collaboration with experts, and compliance with ethical guidelines. The project aims to contribute to knowledge while prioritizing participant well-being and rights.

Abstract

Cancer, an enduring medical enigma with historical recognition dating back to ancient civilizations, remains without a definitive cure. This research undertakes a comprehensive investigation encompassing nine prevalent global cancer types, including those with significant implications for the population of Bangladesh. Employing cutting-edge machine learning (ML) and deep learning (DL) models, as well as traditional machine learning techniques, our study derives its strength from an extensive dataset sourced from the Surveillance, Epidemiology, and End Results (SEER) program. Our research endeavors to unravel the intricate tapestry of cancer by distilling pivotal insights from substantial datasets. At its core, our mission is to redefine the landscape of cancer treatment through the creation of predictive models, thus heralding an era of personalized and highly efficacious cancer therapies. Based on a hypothesis, our objective seeks to improve cancer treatment by developing predictive models. Through a comparative analysis involving traditional machine learning models, deep learning algorithms, and boosting models, we have discovered that the boosting models stand out in terms of accuracy, indicating their potential to enhance predictive precision for therapeutic response. We hypothesize that the surgical removal of localized tumors can effectively arrest cancer progression, thereby increasing patient survival. This encapsulates the main focus of our study, which is a committed attempt to identify unique answers to a persistent medical dilemma by integrating the knowledge of the past with the potential of the future.

Keywords: SEER Data, Machine Learning, Deep Learning, Cancer, Prediction, Survivability

Acknowledgement

Firstly, all praise to the Great Allah for whom our thesis have been completed without any major interruption.
Secondly, to our advisor Mr. Arif Shakil sir for his kind support and advice in our work. He helped us whenever we needed help.
And finally to our parents without their throughout support it may not be possible.

Table of Contents

List of Figures

List of Tables

Nomenclature

The next list describes several symbols & abbreviation that will be later used within the body of the document

AI Artificial Intelligence

AUC Area Under Curve

DL Deep Learning

HGB Histogram Based Gradient Boosting

IARC International Agency for Research on Cancer

ML Machine Learning

MLP Multi layer Perceptron

NCD Non-communicable Diseases

NIH National Institute of Health

ROC Receiver Operating Characteristic

SEER Surveillance, Epidemiology, and End Results

SVM Support Vector Machine

TRF Trees Random Forest

WHO World Health Organisation

Chapter 1

Introduction

Cancer is a disease characterized by the uncontrolled growth and spread of abnormal cells, [22] affecting various organs and tissues in the human body. It is expected to become the leading cause of death worldwide in the 21st century, as non-communicable diseases (NCDs) dominate global fatalities [45].

1.1 Global Impact of Cancer

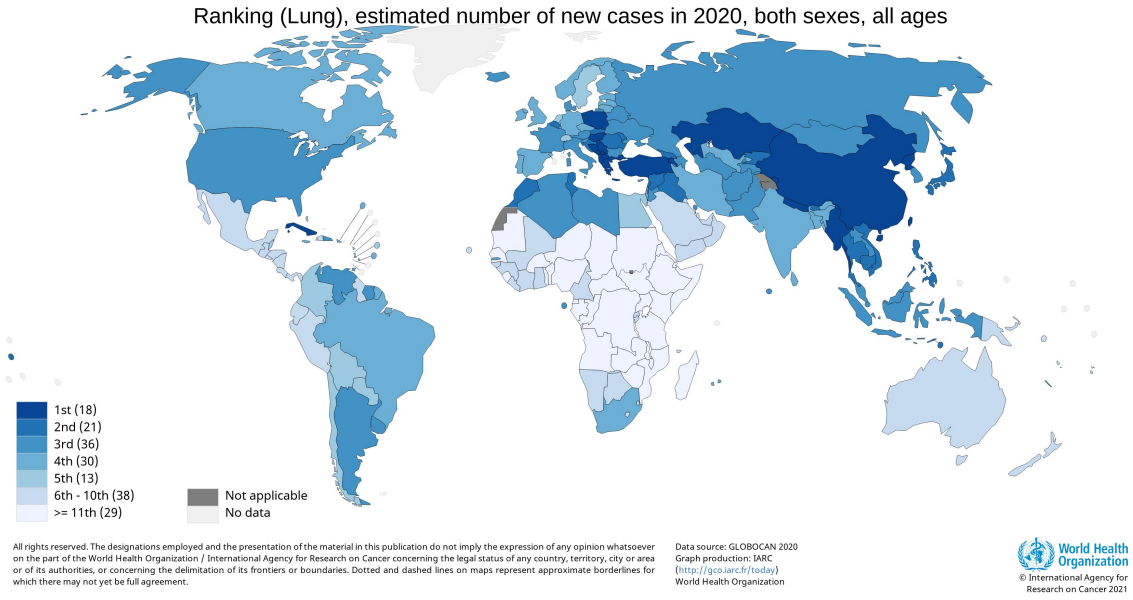

Figure 1.1: Global Map Presenting the estimated age-standardized incidence rates (World) in 2020, all cancers, both sexes, all ages.
Source- World Health Organization [17]

According to the World Health Organization (WHO), cancer was the second leading cause of death globally in 2018, responsible for approximately 9.6 million deaths, equating to one in every six deaths [43]. Men have historically been at a higher risk of dying from cancer [26]. The International Agency for Research on Cancer (IARC) estimates that one in five individuals worldwide will experience cancer in their lifetime, with one in eight men and one in eleven women succumbing to the disease. In 2020, there were projected to be 10.3 million cancer-related deaths and 19.3 million new cases worldwide [16]. Additionally, it is anticipated that 400,000 children and adolescents aged 0-19 will be diagnosed with cancer [42]. By 2030, the number of new cancer cases is predicted to rise from 12.7 million to 21.4 million[46].

1.2 Cancer in Bangladesh

In Bangladesh, cancer ranks as the sixth leading cause of death, and the country is expected to experience a significant increase in cancer fatalities, from 7.5% in 2005 to 13% in 2030, according to the Bangladesh Bureau of Statistics and the IARC [46]. The number of cancer patients in Bangladesh is estimated to be between 1.3 and 1.5 million, with approximately 200,000 new cases diagnosed each year [21]. In 2020, there were 156,775 new cancer cases reported, with an estimated 11% chance of developing cancer before the age of 75 and an 8.4% risk of dying from cancer before that age [44]. In 2020, there were 108,990 cancer-related deaths, a slight increase from 108,137 in 2018 [44].

1.3 Prevention and Early Detection of Cancer

Prevention and early detection play pivotal roles in the battle against cancer. Prevention strategies focus on reducing exposure to risk factors that contribute to cancer development. This includes adopting a healthy lifestyle, such as avoiding tobacco use, maintaining a balanced diet, engaging in regular physical activity, and limiting alcohol consumption [8]. Public health initiatives also emphasize vaccination against cancer-causing viruses like the human papillomavirus (HPV) and hepatitis B [11]. Additionally, promoting awareness and education about cancer prevention helps empower individuals to make informed choices and reduce their cancer risk [27].

Early detection is equally critical as it enables the identification of cancer at its initial stages when treatment outcomes are generally more favorable. Regular screenings, such as mammograms for breast cancer and colonoscopies for colorectal cancer, aid in detecting abnormalities even before symptoms manifest. Heightened awareness of cancer warning signs and seeking prompt medical attention for any concerning symptoms contribute to early diagnosis and timely interventions [33].

By investing in prevention measures and early detection efforts, we can significantly reduce the burden of cancer, enhance survival rates, and improve overall public health outcomes. The collective commitment towards prevention and early detection serves as a powerful approach in the fight against cancer, fostering a healthier and more resilient global population [27].

1.4 Treatment and Recovery

Treatment and recovery are integral components of the cancer journey, providing hope and support to individuals diagnosed with the disease. Cancer treatment aims to target and eliminate cancer cells, halt their growth, or alleviate symptoms to improve the quality of life [15]. Depending on the cancer type, stage, and individual factors, treatment options may include surgery, chemotherapy, radiation therapy, immunotherapy, targeted therapy, or a combination of these modalities. The treatment process often demands resilience, as patients may experience side effects and emotional challenges [23].
Recovery, on the other hand, embodies the process of healing and rebuilding physical and emotional strength after completing cancer treatment. This phase involves regular follow-up visits with healthcare professionals to monitor for any signs of recurrence and to manage any potential long-term effects of treatment [13]. During recovery, individuals often seek support from their loved ones, cancer support groups, and mental health professionals to cope with emotional and psychological impacts. The journey from diagnosis to treatment and recovery represents an incredible testament to human strength and determination. While it may be challenging, many cancer survivors find new perspectives on life, discover renewed gratitude, and embrace a deeper appreciation for the present moment. Empowering cancer patients with compassionate care, ongoing support, and comprehensive survivorship programs aid in fostering resilience and enhancing their well-being throughout their journey toward recovery and beyond.

1.5 Role of AI and ML in Healthcare

The healthcare sector has been at the forefront of adopting new technologies, and machine learning is increasingly playing a pivotal role in medical advancements and the development of novel medical techniques [35]. Numerous applications of machine learning, particularly in outcome prediction models, have been reported in recent clinical literature, encompassing various outcomes such as death, cardiac arrest, and acute.

1.6 Problem Statement

The number of cancer cases is increasing at an alarming rate. The National Institute of Health(NIH) estimated 1,918,030 new cancer cases and an estimated death of 609,360 in 2022. Where 15 percent of the total number is estimated to be breast cancer in terms of new cases[41].

According to NIH, the percentage of patients who are predicted to survive the effects of their cancer is estimated as relative survival. It eliminates the chance of being passed away from other circumstances. A patient's specific outcome cannot be predicted using survival statistics because they are based on large numbers of people. Since no two patients are the same, treatment and patient responses might differ substantially[41]. Relative survival is calculated as the difference between the proportion of expected survivors in a cohort of people without cancer and the pro-

portion of observed survivors (all causes of death) in a cohort of cancer patients. The approach is predicated on the idea that several competing causes of death exist independently[37].

This research aims to predict the effectiveness of a treatment considering the various kinds of cancers. Additionally, this research will be able to predict the success rate of a treatment considering the matrices of cancer.

1.7 Objective

The purpose of this study is to determine the effectiveness of various cancer treatments and the success rates of those treatments for various cancer kinds. Additionally, we'll look into survival statistics for various cancer kinds as part of our study. Although there are several varieties of cancer and therapies are available, none of them are currently known to be curable. However, some treatments may be able to do so. So, using various techniques, we will gather data and build a model, which we will then analyze and forecast. Our primary goal is to investigate effective cancer treatments and survival rates, as well as to communicate the information we learn from our research so that doctors may treat cancer patients properly and ensure they receive the correct care for their condition.

- To examine the dataset and pick out the optimal variables.
- To fully comprehend the cutting-edge applications of Machine Learning. To deeply understand the Advanced use of Deep Learning.
- To construct a model that can be used to monitor and analyze large amounts of data.
- To create an efficient model and utilize the benefits.
- Providing suggestions for enhancing cancer therapies based on success rate in a shorter amount of time.

1.7.1 Machine Learning

We will use many algorithms to find accuracy and success rates, Machine learning is one of them. Besides, learning ML and different algorithms of ML is also an objective for our team.
In computer science and artificial intelligence, machine learning is a subfield. It focuses on creating models utilizing data and algorithms to help people comprehend any situation better while also increasing precision. We need to know about the ML algorithm properly so that we use appropriate ones to extract results[47].

1.7.2 Deep Learning

Deep learning uses artificial neural networks to do intricate computations on massive amounts of data. It is a type of artificial intelligence that is based on how the human brain functions and is put together [32]. Deep learning approaches use samples to teach computers new things.
Deep learning models employ several algorithms. Although no network is thought to be flawless, specific algorithms are more effective at carrying out particular tasks [38]. It helps to have a firm grasp of all fundamental algorithms to select the appropriate ones[34].

1.8 Motivation

Cancer, an unrelenting adversary, continues to exact a devastating toll on lives worldwide. Despite advancements in medical science, the quest for effective cancer treatments remains an ongoing challenge. Amidst this battle, the urgent need emerges for tailored and precise approaches to cancer treatment, addressing the critical question: In which situations can we optimize survival rates for cancer patients? This study finds its motivation in the alarming reality that cancer's grasp on human lives persists, with millions succumbing to its relentless advance each year. The primary goal is clear - to decipher the nuanced factors that influence cancer survival and to construct predictive models that can offer a beacon of hope.
The essence of our motivation lies in the belief that we can redefine the odds. By exploring the intricacies of cancer data, we aim to unveil the circumstances where survival rates are optimized. Our focus is on constructing models, guided by state-of-the-art machine learning techniques, that can provide insights into these critical moments.
The ultimate aspiration is to pave the way for personalized cancer treatment strategies. We endeavor to harness the power of data to tailor interventions to each patient's unique situation, offering them the best possible chance at survival. Through meticulous analysis, we aim to provide a roadmap for healthcare professionals, helping them make informed decisions that can transform the lives of cancer patients.
In the following pages, we embark on a mission that combines the urgency of the present with the hope of a brighter future. By constructing models rooted in data-driven insights, we aim to bridge the gap between scientific knowledge and tangible benefits for cancer patients. With unwavering determination, we aspire to enhance the precision and efficacy of cancer treatments, offering a glimmer of hope to those facing this formidable adversary.

Chapter 2

Literature Review

A literature review is a document that summarizes and exhibits an understanding of the scholarly papers on a certain topic in context. If we consider the current situation a lot of countries are finding cures for cancer, and researchers are finding patterns to describe the health condition and how to heal the human body. Data analysis will present an extraordinary chance to comprehend cancer at every level and support decisions about how to treat patients using the knowledge gleaned from these enormous databases. Some are forecasting It's possible that computer scientists analyzing statistics rather than biologists staring through microscopes may discover the next generation of cancer medicines. In our research, we focus on cancer and its therapies, making predictions about effective therapy as well as survival rates. This section contains keywords, earlier research on the subject from other scholars, existing circumstances, and future projects[9].

Ganggayah et al.[28] presents a machine learning model for predicting factors influencing the survival of breast cancer patients. The model's accuracy in predicting patient survival is a notable advantage, offering potential benefits for personalized treatment planning. By identifying crucial survival factors, this model provides clinicians with valuable insights into patient-specific care strategies. However, the model's applicability to other cancer types and its generalizability need further exploration. Cancer is a heterogeneous disease, and different cancer types may require unique predictive models. Researchers must assess the model's transferability to other cancer domains to maximize its impact on clinical practice.

Rafique et al.'s[40] paper discusses the use of machine learning in predicting cancer therapy. It highlights the advantages of improved accuracy and efficiency in therapy prediction, potentially revolutionizing cancer treatment planning. By tailoring therapies to individual patient needs, machine learning can enhance treatment effectiveness while minimizing adverse effects. Nonetheless, this paper emphasizes several challenges. Data scarcity remains a significant impediment, as machine learning models thrive on extensive datasets. Ethical concerns, such as algorithm transparency and bias mitigation, are critical considerations in treatment decision-making. Researchers must collaborate closely with healthcare professionals to address these challenges and ensure responsible AI implementation in clinical practice.

Huang et al.'s[36] paper explores the opportunities and challenges of using artificial intelligence (AI) for cancer diagnosis and prognosis. It highlights AI's potential to improve accuracy and efficiency in cancer diagnosis, treatment planning, and prognosis. AI's capacity to process vast amounts of data quickly can assist clinicians in making more informed decisions, leading to better patient outcomes. However, the paper also underscores several challenges. Data availability remains a significant hurdle, as AI algorithms require extensive, high-quality datasets for training and validation. Additionally, ethical considerations, such as data privacy and algorithm transparency, demand careful attention. Researchers and practitioners must collaborate to ensure responsible AI adoption in clinical settings.

Alzubaidi et al.'s[41] comprehensive review of DL in cancer treatment provides valuable insights into the potential of DL in risk assessment, predictive models, and image analysis. One significant advantage of this paper is its broad coverage of topics, offering a holistic view of DL's applications in precision cancer treatment. This wide-ranging approach allows researchers to understand the diverse landscape of DL applications in the field of cancer treatment. However, the review also underlines several challenges. The ethical considerations in using AI for clinical decision-making are paramount. Ensuring patient data privacy and establishing transparent decision-making processes are critical steps toward responsible DL implementation. Addressing these challenges is essential for realizing the full potential of DL in cancer treatment.

C. Stein's[8] paper focuses on modifiable risk factors for cancer, shedding light on factors that individuals can change to reduce their cancer risk. While it doesn't delve into machine learning or deep learning, it provides crucial information about lifestyle-related risk factors such as smoking, obesity, and physical inactivity that are vital in cancer prevention strategies. The strength of this paper lies in its emphasis on prevention and lifestyle choices, which play a significant role in reducing cancer incidence. Researchers in cancer epidemiology and public health can benefit from this paper to understand the importance of addressing modifiable risk factors. However, it doesn't directly address machine learning or data analysis in cancer research.

H. Ming's[10]paper introduces an enhanced decision tree classification algorithm based on the ID3 algorithm. While it doesn't focus on cancer applications, it presents a valuable contribution by improving the accuracy and efficiency of the ID3 algorithm. Decision tree algorithms are commonly used in healthcare, including cancer research, for tasks like patient prognosis. The strength of this paper is its practicality, as decision trees are widely used in medical data analysis. Researchers interested in machine learning for cancer applications can benefit from this algorithm enhancement. However, this paper lacks specific examples or applications in the context of cancer data.

This paper by N. Viney et al.[12] discusses the use of ensemble modeling to assess the impact of land use change on hydrology. While not cancer-specific, it demonstrates the value of ensemble techniques in predicting environmental changes. The advantage of this paper is its application of ensemble modeling, a technique relevant to predictive modeling in cancer research, such as assessing the impact of

environmental factors on cancer incidence. Researchers interested in environmental epidemiology may find this paper insightful. However, it doesn't directly address cancer datasets or outcomes.

The American Cancer Society's[2] paper provides an overview of statistics used to guide prognosis and evaluate treatment in cancer patients. While not a traditional research paper, it serves as an educational resource on the statistical aspects of cancer care. The advantage of this paper is its focus on statistical methods and their importance in clinical decision-making. Researchers, clinicians, and students in oncology can benefit from a deeper understanding of the statistical tools used in cancer research and treatment evaluation. However, it doesn't present original research or data analysis.

This paper by A. Alama[15] and colleagues discusses the potential of targeting cancer-initiating cells (CICs) to overcome drug resistance in cancer patients. Although not machine learning-oriented, it addresses a critical aspect of cancer therapy and the challenges associated with targeting CICs. The strength of this paper lies in its exploration of innovative approaches to combat drug resistance in cancer, a topic of great importance in oncology research. It encourages researchers to consider novel therapeutic strategies. However, it doesn't delve into data analysis techniques or machine learning applications.

This paper by M. S. Hossain et al.[21] presents a retrospective analysis of over 5,000 cases of diagnosed hematological malignancies in Bangladesh. While not explicitly focusing on machine learning, it provides valuable epidemiological insights into the prevalence and demographics of hematological malignancies. The strength of this paper is its contribution to understanding the landscape of hematological malignancies in a specific region, which can inform healthcare policies and resource allocation. Researchers interested in the epidemiology of hematological cancers may find this paper informative. However, it doesn't delve into advanced data analysis methods.

In this paper, G. Pennock[23] explores the dynamic landscape of immune checkpoint inhibitors (ICIs) in cancer treatment. ICIs, a class of immunotherapeutic drugs, hold immense promise in unleashing the body's immune system against cancer cells. The paper underscores their success in treating various cancers but also highlights the need for a deeper understanding of their optimal utilization and management of associated side effects. This paper's strength lies in its relevance to the rapidly evolving field of cancer immunotherapy. Researchers and clinicians involved in cancer treatment will find valuable insights into the latest developments and challenges related to ICIs. However, it doesn't delve into machine learning or data-driven approaches.

K.H. Yu et al.[26] explore the transformative potential of artificial intelligence (AI) in healthcare. AI's applications in diagnostic tools, personalized care, and healthcare efficiency enhancement are discussed. The paper also addresses the challenges that must be overcome to fully realize AI's impact on healthcare. The strength of this paper is its relevance to the integration of AI in oncology. Cancer diagnosis and treatment benefit greatly from AI-driven advancements, and this paper provides an

overview of the opportunities and obstacles in this context. Researchers and healthcare professionals in cancer care should find this paper informative. However, it's more focused on healthcare in general than cancer-specific applications.

E. Alawadhi[27] presents a population-based study on cancer survival in Kuwait. The research identifies lower survival rates compared to developed countries, with potential causes including late-stage diagnosis, limited access to high-quality care, and cultural factors. The strength of this paper lies in its epidemiological insights into cancer survival specific to Kuwait, which can inform regional healthcare strategies. Researchers and policymakers interested in cancer epidemiology in the Middle East may find this study valuable. However, it doesn't primarily focus on data analysis or machine learning.

H. Kamel[30] and colleagues employ the Gaussian naive Bayes algorithm for cancer patient classification. This machine learning algorithm assumes independence and normal distribution of features, and the paper demonstrates its high accuracy in cancer patient classification. The strength of this paper is its direct application to cancer classification, showcasing the effectiveness of the Gaussian naive Bayes algorithm. Researchers and practitioners in the field of cancer diagnostics may find this study informative, as it provides insights into an algorithm's potential for accurate classification. However, it primarily focuses on algorithmic applications rather than exploring broader machine-learning topics.

T. Prasanth[32] proposes a deep learning modified neural network for efficient big data retrieval. The network is designed to learn the underlying structure of big data, leading to improved retrieval accuracy and efficiency. The strength of this paper is its relevance to big data challenges, which are increasingly prevalent in cancer research. Big data retrieval is crucial in managing and analyzing large-scale cancer datasets. Researchers dealing with extensive oncology data can benefit from this paper's insights into efficient retrieval techniques. However, it doesn't primarily focus on cancer-specific applications.

In this paper, J. A. M. Sidey-Gibbons[35] provides a practical introduction to machine learning in medicine. The paper discusses different types of machine learning algorithms and their applications in solving medical problems. The strength of this paper is its applicability to healthcare, which includes cancer diagnosis and treatment. Machine learning plays a growing role in various aspects of oncology, from image analysis to predicting treatment outcomes. Researchers and healthcare professionals in the field of cancer care can find practical insights in this introduction. However, it doesn't focus specifically on cancer but provides a broader overview.

This technical report authored by P. J. M. Ali et al.[20] provides an in-depth overview of data normalization and standardization techniques, fundamental in preparing data for analysis. While not cancer-specific, it offers essential knowledge for researchers dealing with diverse datasets, including cancer data. The strength of this paper is its technical depth, guiding researchers on the nuances of data preprocessing. Researchers in the field of oncology, where data quality is paramount, can benefit from understanding data normalization and standardization. However, it

lacks direct cancer-related examples.

A. Natekin and A. Knoll's paper[18] serves as a tutorial on gradient boosting machines, a versatile machine learning algorithm applicable to classification and regression tasks. This paper is valuable for researchers and practitioners interested in harnessing the power of ensemble learning in cancer research. The strength of this paper lies in its educational aspect, providing a step-by-step guide on training and using gradient boosting machines. Given the increasing adoption of machine learning techniques in oncology, this tutorial can empower researchers to leverage gradient boosting for predictive modeling. However, it lacks direct cancer-related case studies.

R. E. Schapire's paper[19] offers an explanation of Adaboost, another ensemble learning algorithm applicable to classification and regression tasks. While not cancer-specific, it provides insights into the workings of Adaboost, which can be relevant in the context of cancer data analysis. The advantage of this paper is its focus on algorithmic understanding, aiding researchers in grasping the fundamentals of Adaboost. Knowledge of Adaboost can be beneficial for those seeking to explore ensemble learning methods in cancer research. However, it doesn't present cancer-related applications.

G. Biau and E. Scornet[24] offer a comprehensive tour of random forests, a widely used machine learning algorithm renowned for its versatility in classification and regression tasks. Random forests, composed of decision trees, excel in accuracy and resilience. This paper is particularly valuable for its clarity in explaining the workings of random forests. The paper's strength is its educational value, making it an essential resource for cancer researchers looking to harness the power of random forests for predictive modeling. While not cancer-specific, random forests are employed extensively in cancer research, making this paper highly relevant. However, it lacks direct cancer-related case studies.

This paper by K. Potdar et al.[25] conducts a comparative study of encoding techniques for categorical variables in neural network classifiers. While not cancer-specific, it delves into an essential aspect of data preprocessing crucial in handling diverse datasets, including those in oncology. The paper's strength lies in its practicality, offering insights into choosing the most suitable encoding technique based on data characteristics and neural network architecture. Researchers dealing with cancer data, which often contain categorical variables, can benefit from this guidance. However, it lacks direct cancer-related applications.

A. Guryanov[29] introduces a histogram-based algorithm for constructing gradient-boosting ensembles of piecewise linear decision trees. Gradient boosting is a powerful machine-learning technique applicable to classification and regression tasks, known for its accuracy. This paper showcases an algorithm that enhances the efficiency and accuracy of traditional gradient-boosting ensembles. The strength of this paper is its contribution to machine learning methodology, which has potential applications in cancer data analysis. It addresses an algorithmic aspect that can improve the performance of gradient boosting, a technique used in various areas of oncology

research. Researchers interested in optimizing ensemble methods may find this paper valuable. However, it doesn't focus on cancer-specific applications.

Chapter 3

Methodology

3.1 Dataset

Real-time data analysis has become increasingly important in advancing cancer research and patient care. Cancer remains a major worldwide health concern. In order to guarantee the authenticity and dependability of our real-time cancer data analysis, we set out on a careful data management journey in this thesis. To extract valuable insights, draw wise conclusions, and improve cancer research, effective data management is crucial. We lay forth a thorough methodology that includes data pre-processing, feature selection, correlation analysis, cross-validation, and model evaluation. This methodical approach maximizes the data's usefulness while maintaining its integrity throughout the study.

In order to assure data homogeneity and acceptability for future analysis, data pre-processing is a crucial first step in any data analysis effort. Data pre-processing in our study includes standardizing and cleaning the data to get rid of discrepancies and improve the precision of later analyses. Data normalization, in which the data were converted into a standardized scale, was essential to this procedure. By doing this, we eliminated the impact of several measuring units and established a uniform baseline for the data, allowing for precise comparisons and lowering the possibility of skewed conclusions. The ensuing phases of our research were built on a firm foundation created by this rigorous attention to data pre-processing, which also guaranteed the accuracy and dependability of the data throughout the analysis.

To ensure the success and validity of our investigation, it is crucial that the appropriate features are chosen for analysis. We conducted a thorough data evaluation procedure, working closely with a clinical specialist, to pinpoint the dataset's most pertinent and significant elements. We were able to understand the medical meaning and context of particular traits related to cancer research thanks in large part to the knowledge and perceptions of the medical professional. By choosing these relevant criteria, we can make sure that our analysis is concentrated on the most important facets of cancer results and patient treatment. We increase the clinical relevance and validity of our study by incorporating the domain expert's knowledge, producing valuable and useful insights.

In order to get a thorough grasp of cancer dynamics, we looked into the interactions

between various features in our data exploration. We used the correlation coefficient, a potent statistical indicator, to do this. The correlation coefficient measures how closely two variables are related, allowing us to spot any dependencies and connections in the data. This analysis reveals significant patterns that can guide our further research by shedding light on the links between various parameters. The discovery of possible risk factors, predictive variables, and underlying mechanisms affecting cancer outcomes is made easier by an understanding of these linkages.

We used cross-validation techniques to ensure the accuracy of our predictive model and prevent overfitting. In machine learning and data analysis, the practice of cross-validation includes segmenting the dataset into various subsets. The additional subsets are used for repeatedly training while one subset is used to evaluate the performance of the model. This method enables us to thoroughly assess the performance of our model under numerous data circumstances, producing results that are more trustworthy and generalizable. We strengthen the validity and ability to generalize our findings by including cross-validation, guaranteeing that the predictive model's effectiveness translates to practical applications.

The creation of a predictive model that provides useful information about the course of cancer is the ultimate goal of our data management efforts. We make use of machine learning methods by training and assessing models. The model is trained by feeding it the chosen features and data points, which helps it recognize and adjust to the patterns in the dataset. The model is then rigorously tested using a cross-validated subset of data to determine how well it performs. Through this iterative approach, we may fine-tune the model and increase its predicted accuracy and predictability. Our predictive model has been optimized to increase its precision and applicability, resulting in significant advancements in both cancer studies and the treatment of patients.

In summary, our thorough approach to data management is essential for gleaning insightful information from real-time cancer information and drawing defensible conclusions. The core of our research is built on rigorous data pre-processing, guided feature selection, correlation analysis, strong cross-validation, and optimal model evaluation. The effect of our findings is enhanced by the combined effort with a medical specialist, which confirms the clinical significance and validity of our analysis. In the end, our approach to data management advances patient care and cancer research, creating a more promising future for the battle against cancer. We can open up new vistas for comprehending cancer dynamics, resulting in improved treatment plans and better patient outcomes, by continuously improving and refining our data management processes. With the help of this study, we hope to have a long-lasting impact on the global fight against cancer and improve the lives of millions of people who are impacted by this difficult and complex illness. Figure 3.1 in page 14 shows the workflow of the study.

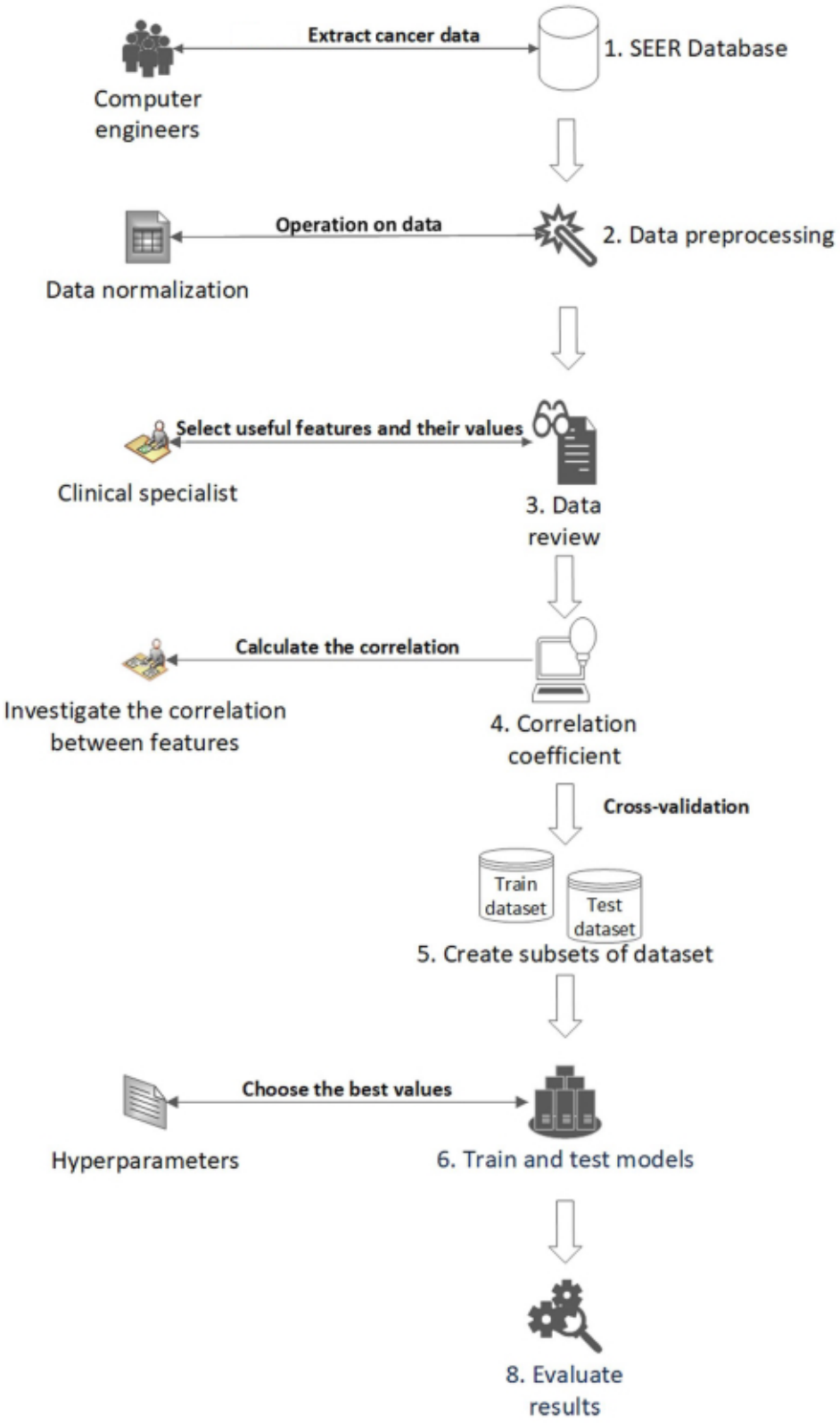

Figure 3.1: Workflow

3.2 Data Analysis

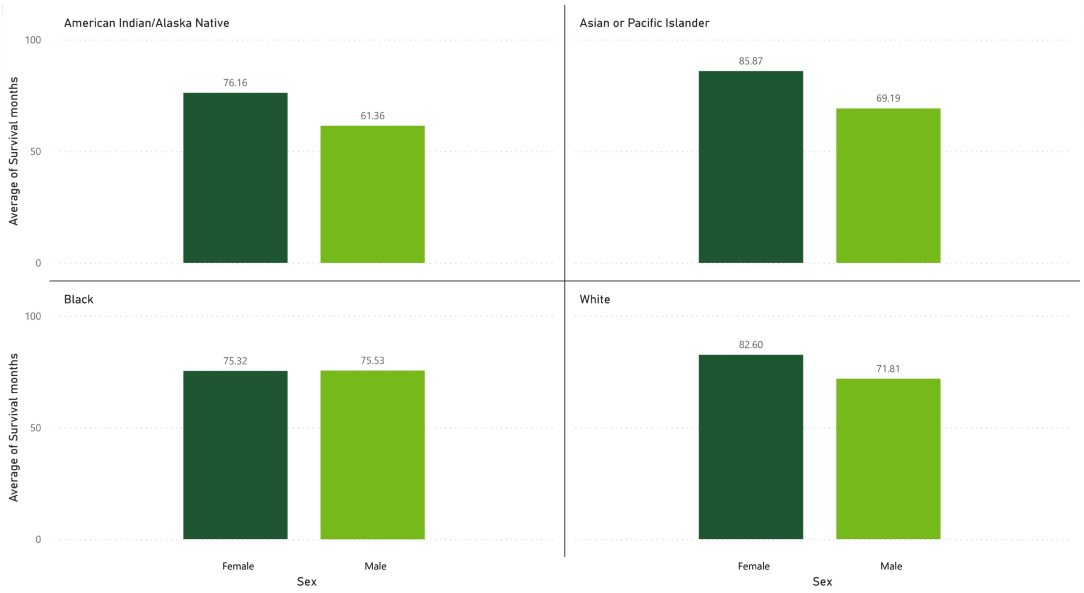

Figure 3.2: Survival in regard to race & sex

In figure 3.2 the column which represents females has the highest survival average among the known four races according to this graph, which illustrates the average survival month by race. Male individuals have a low average survival rate among all races.

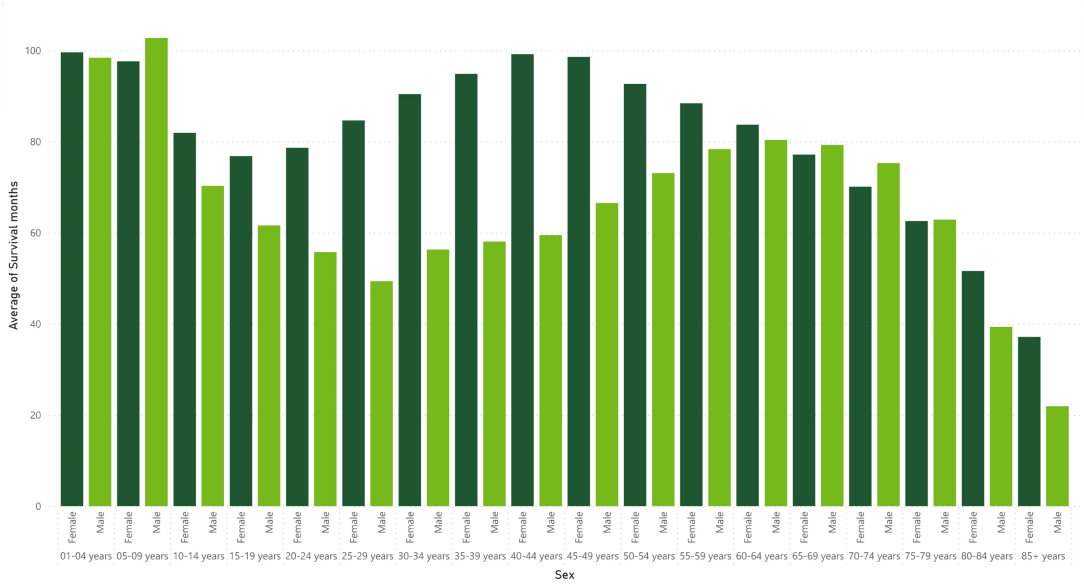

Figure 3.3: Average Survival in regards to Age & Sex

From figure 3.3, the graph shows the age ranges for both male and female sex as well as the average survival month. As we can see, as people get older, their chances of surviving or having the greatest survival rate decline. Women generally have a longer lifespan at all ages, with one notable exception of newborns, whose chronological age falls before the age group of 60.

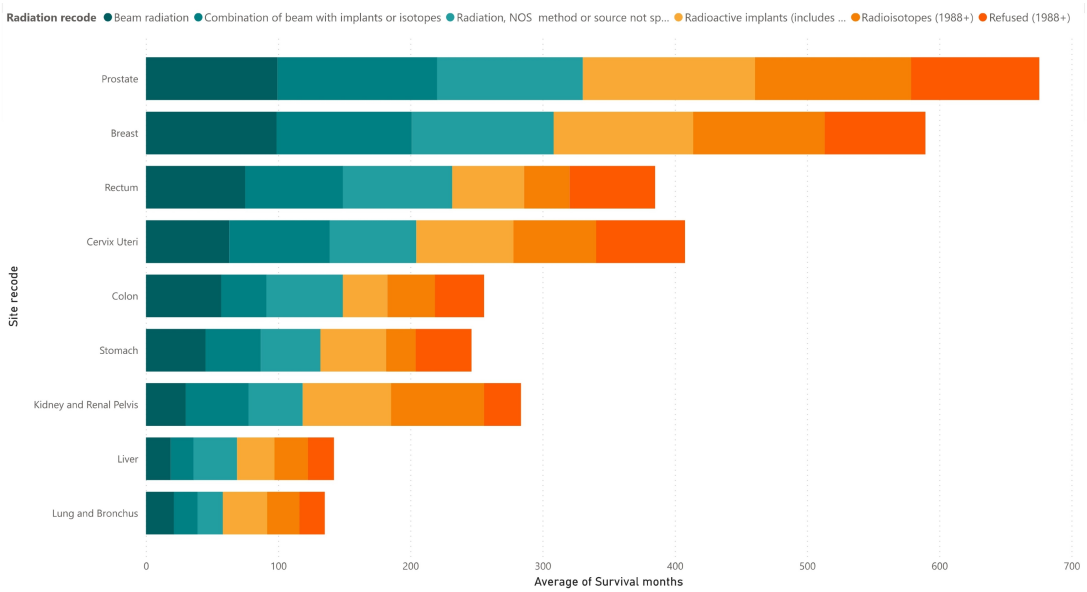

Figure 3.4: Average Survival in regards to site record & radiation record

Figure 3.4 shows how different radiation and other treatments affect the survivability in different types of cancer. This graph shows which types of cancer are treated with radiation treatment. We can tell that every sort of cancer uses beam radiation the most frequently from this. Additionally, we may see certain important elements, such as the predominant usage of a certain type of radiation treatment. This graph is crucial for a better understanding of radiation therapy for various types of cancer.

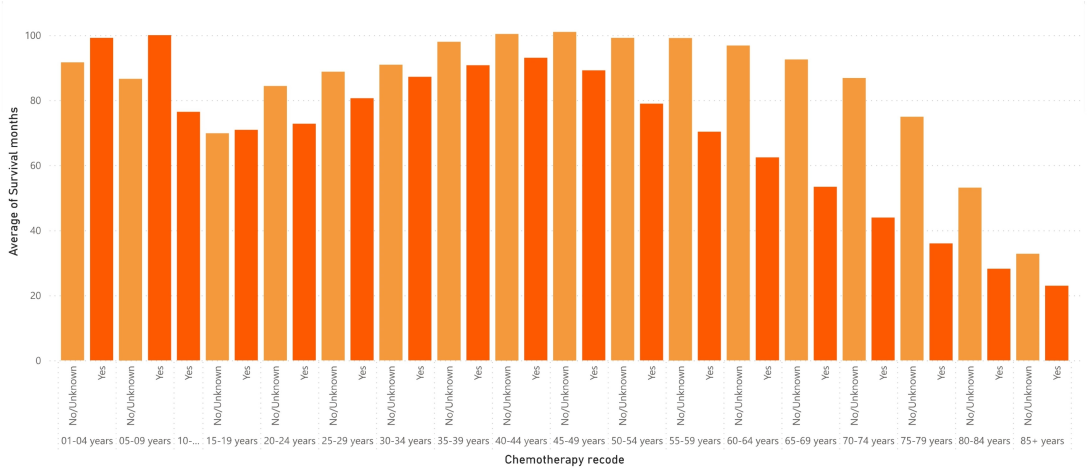

Figure 3.5: Average Survival in regards to Chemotherapy Record & Age

In figure 3.5 we were looking for the chemotherapeutic outcomes of the statistical analysis of the data. So, using chemotherapy data as the legends, we built up a bar chart with the average survival on the Y-axis, and age groups on the X-axis. The data show that the survival rate is below average for the age group 60 and upwards.

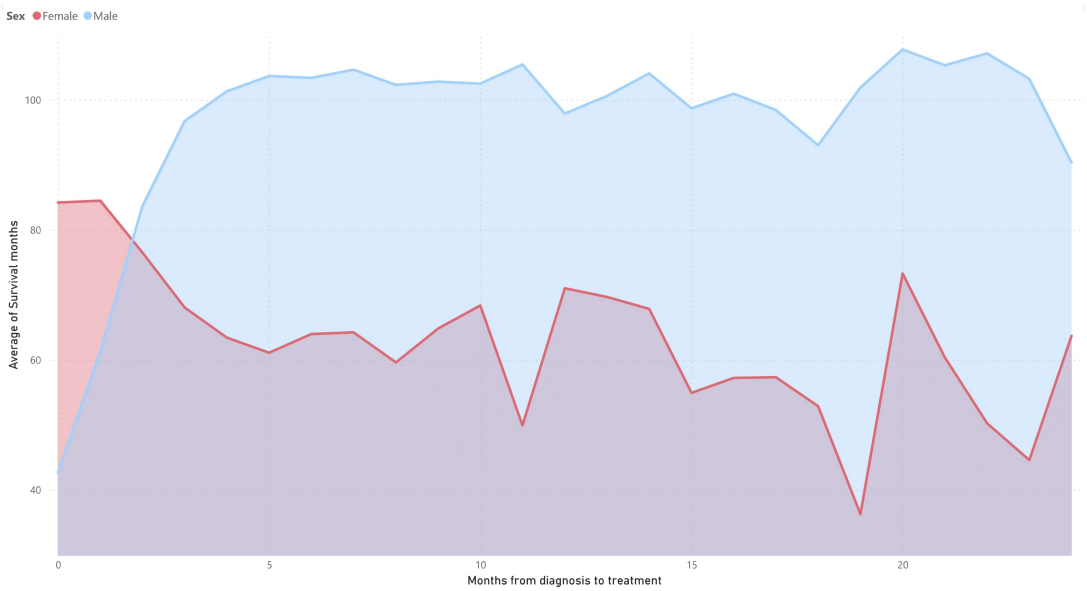

Figure 3.6: Average Survival in regards to Month of Treatment

From the figure 3.6, the statistics between the diagnosis data and the average surviving months are shown in this area chart. The average survival month reaches its maximum when the diagnosis month reaches 60. From that point on, the main conclusions are that a greater survival rate results from a higher diagnosis month, and a lower survival rate results from a lower diagnosis month. However, a number greater than 60 months indicates an average value of survival month.

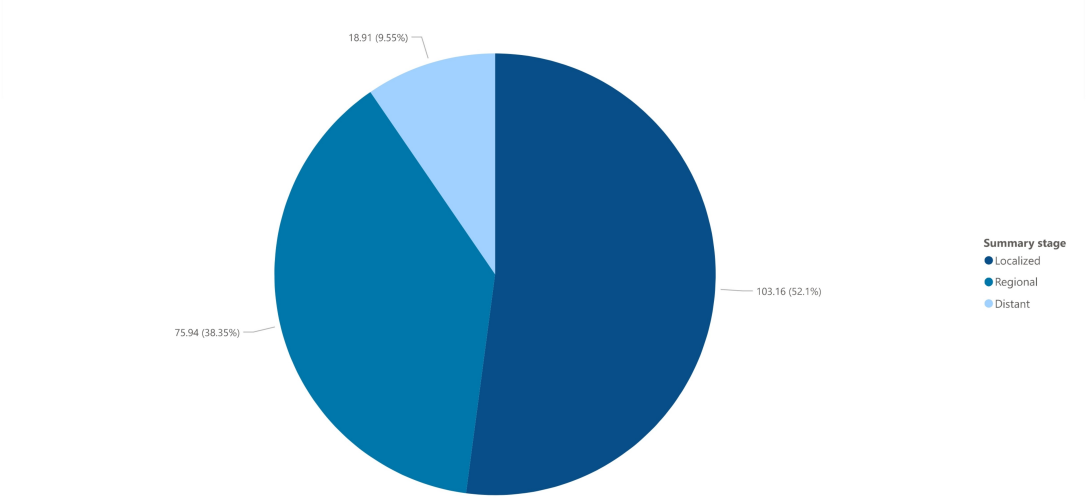

Figure 3.7: Average Survival in regards to Summary Stage

According to this graph 3.7, localized cancer, which refers to cancer cells that are only present in a specific area of the body, has the greatest survival rate (52.1%). The highest survival rate then varies according to the location of the cancer cell (38.35%). The survival percentage varies depending on whether cancer has spread from the main tumor to adjacent lymph nodes, organs, or tissues. The lowest survival rate is (9.55%), and it depends on how far the cancer has progressed from the primary tumor to distant organs.

Our next stages include concluding the consultation with a medical expert,

understanding the vocabulary of the dataset, and dealing with the null values in the data. As a result, we will be able to do thorough data preparation, which will include managing missing values, normalization, and feature engineering. We will next use sophisticated statistical methods to analyze the data, uncover relevant patterns, and draw conclusions.

3.3 Data Preprocessing

In this section, we describe the measures taken to prepare the dataset [3] for accurate cancer treatment survival prediction through data preprocessing. The initial batch of data included information on fifteen types of cancer. To concentrate on the most prevalent cancers, we selected the five most prevalent forms of cancer worldwide and in Bangladesh [17]. By combining these two sets, we obtained a dataset containing seven cancer types. Nevertheless, two additional cancer types exhibited a substantial quantity of data, prompting us to include a total of nine cancer types in our study.

Site Record	**Number of Data**
Breast	332939
Lung and Bronchus	198690
Prostate	163191
Rectum	39136
Cervix Uteri	19562
Stomach	10349
Kidney and Renal Pelvis	7487
Liver	4904
Colon	3773

Table 3.1: Number of incidences on site record

Before developing a model, we meticulously inspected the dataset for absent values. Given the sensitive nature of the data and the fact that the values corresponded to cancer cases, we decided to exclude any entries with missing data. This method assured the integrity and dependability of the dataset because we did not impute missing values, which could introduce bias into the predictive models[31].

To standardize the characteristics and facilitate the convergence of optimization algorithms, we scaled the data, specifically by normalizing the data. This procedure converted the numerical features to a standard range, thereby mitigating the impact of any feature that dominated the learning process of the model [20].

Next, for the purpose of predicting cancer treatment survival, we defined survival as the condition in which a patient lives for at least 60 months as physicians often use 5 years as a time period of observing relative survival rate [14]. Patients who were still alive at the conclusion of the observation period were therefore presumed to have survived the cancer. This binary representation facilitated the

classification assignment, which sought to predict whether a patient would meet the predetermined survival threshold.

To optimize the predictive performance, we assessed the importance of each feature in predicting cancer treatment survival in the feature selection process. Through meticulous analysis, we identified and eliminated superfluous features that could have impeded model generalization and introduced noise into the predictions.

We performed feature encoding to convert the categorical features present in the dataset into a numerical representation. This transformation permitted the incorporation of categorical data into machine learning models, thereby enhancing their capacity to recognize patterns and relationships [25].

To accurately evaluate the performance of our models, we divide the dataset into three subsets: the training set, and the test set. The training set accounted for 70% of the data, while the test set comprised 30%. This partitioning ensured that the models were trained on a substantial quantity of data while also being evaluated on unseen data to determine their generalization capabilities [39].

In conclusion, the data preprocessing methods outlined in this section were intended to generate a robust and trustworthy dataset, optimized for accurate cancer treatment survival prediction. By managing missing values strategically, normalizing data, performing feature selection, and encoding categorical features, we curated a comprehensive dataset that will serve as the basis for our subsequent model development and analysis.

3.4 Model Architecture

We illustrate the convoluted model architecture used in our technique for generating therapy hypotheses for cancer in this section. We used Decision Tree, Random Forest, Logistic Regression, Gradient Boosting, AdaBoost, HistGradientBoosting, and Naive Bayes models to produce a varied collection of classifiers by utilizing the capabilities of ensemble learning techniques. By combining the different capabilities of these models, we were able to increase prediction precision. Additionally, in order to recognize intricate links and patterns in the data, we used deep neural networks based on a multilayer perceptron. The design has several hidden layers with different activation functions that may be used to learn complex feature representations. A reliable and comprehensive method for predicting cancer treatment outcomes from patient survival data is provided by the use of both ensemble learning and deep neural networks, promoting improved precision and personalized cancer care.

3.4.1 Logistic Regression

For binary or multiple-class classification tasks, logistic regression is a popular statistical model. Using a logistic or sigmoid function, it models the link between input variables and the likelihood of a specific outcome. The log odds of the anticipated outcome are affected by each factor according to the model's estimated coefficients for each feature.

In logistic regression, the input variables and their corresponding coefficients are combined linearly to represent the log odds of the anticipated outcome. The logistic function, which transfers the linear combination to an integer in the range of 0 to 1 is then used for altering this linear function. This value has been modified to show the estimated likelihood of a successful outcome.
The coefficients obtained from the Logistic Regression model provide insights into the importance and direction of each feature's influence on the outcome. A positive coefficient indicates that an increase in the corresponding feature's value increases the log odds (and thus the probability) of the positive outcome. Conversely, a negative coefficient suggests a decrease in the log odds (and probability) with an increase in the feature's value[1].

Model Equations and Formulations

In logistic regression, the relationship between the input factors (x) and the likelihood of a specific outcome (y) is represented by the function of logistic regression. The calculation looks like this:

$$\hat{p}\left(X_i\right) = \operatorname{expit}\left(X_i w + w_0\right) = \frac{1}{1 + \exp\left(-X_i w - w_0\right)} \tag{3.1}$$

In this equation, $P\left(y_i = 1 \mid X_i\right)$represents the likelihood of a positive outcome given the input variables. The symbol "e" corresponds to the base of the natural logarithm, and the coefficients (weights) associated with the input features are denoted by $w_0, w_1, w_2, ..., w_n$.
The logistic function converts a probability value from a linear combination of the input factors and their associated coefficients. The exponentiation of the negative z value in the denominator ensures that the probability lies between 0 and 1, capturing the idea of a binary outcome.

Rationale for Model Selection

Logistic Regression models were chosen for several reasons. Firstly, they offer simplicity and interpretability. The coefficients obtained from the model provide a clear understanding of the relationship between patient characteristics and treatment outcomes. They indicate the direction and magnitude of the influence that each feature has on the log odds of the predicted outcome. This interpretability allows medical professionals to assess the significance and contribution of individual features in the context of cancer treatment.
Secondly, Logistic Regression provides the ability to estimate the probability of specific outcomes. This estimation is valuable in cancer prediction, as it enables precise predictions based on various factors. Estimating probabilities allows medical professionals to assess the likelihood of treatment success or failure based on patient characteristics, helping them make informed decisions and tailor treatment strategies accordingly.
Lastly, Logistic Regression is suitable for handling both categorical and continuous input features. It can accommodate a wide range of data types, making it versatile in capturing the relationship between features and the probability of a particular

outcome. This flexibility is particularly advantageous when analyzing the impact of individual features on the likelihood of cancer occurrence or progression, as it allows for the inclusion of diverse patient characteristics.
In summary, Logistic Regression is a valuable tool for hypothesizing precise cancer treatments based on patient characteristics. Its simplicity, interpretability, and ability to estimate probabilities make it an ideal choice. By understanding the relationship between input variables and treatment outcomes, medical professionals can make informed decisions regarding cancer treatment strategies. The coefficients obtained from the model provide insights into the importance and direction of each feature's influence, enhancing the understanding of the underlying mechanisms driving treatment outcomes.

3.4.2 Decision Tree Classifier

The values of various attributes are used to create judgments in decision trees, which are hierarchical structures. They allocated the data recursively at each node, taking into account the features that increase useful information or reduce impurity. A decision tree model that permits predictions based on learned rules is produced as a consequence of this procedure, which is carried out until a stopping requirement is satisfied.
The selection of split criteria is a crucial component of decision trees. Decision tree models are frequently built using algorithms like ID3 (Iterative Dichotomiser 3), C4.5, or CART (Classification and Regression Trees). In order to maximize the information acquired from each split, the ID3 and C4.5 algorithms employ entropy and information gain measurements to choose the optimum feature to use [10]. The Gini impurity, on the other hand, is used by CART as a criterion to reduce the possibility of misclassification.
The root node of the decision tree learning algorithm contains the full dataset, which is then recursively divided into subsets according to various criteria. The feature's capacity to divide the classes or lessen impurity is assessed using the selected split criteria to decide the splitting procedure. Until a termination criterion is satisfied—which might be a predetermined maximum depth, a minimum number of samples in each leaf node, or other stopping criteria—this recursive splitting continues[4].

Model Equations and Formulations

Decision trees do not have a specific mathematical equation that represents the entire model. However, at each internal node of the tree, a decision is made based on a condition or threshold related to a specific feature. This decision-making process recurs until it reaches the leaf nodes, which contain the final predictions. The decision tree's predictions are based on a structure of if-else conditions, forming a rule-based framework.
The decision rule at each node is typically represented as: "If Feature A ¡= Threshold B, go to the left child node; otherwise, go to the right child node."
This process continues until reaching a leaf node, where the final prediction or class label is assigned based on the majority class of the samples in that leaf node. The structure of the decision tree can be visualized as a flowchart, with each node representing a decision based on a feature, and each branch representing a possible

outcome.

Rationale for Model Selection

For our project, Decision trees were selected for several reasons. Firstly, they offer high interpretability, providing a clear understanding of the decision-making process. The decision tree structure can be easily visualized, allowing medical professionals to comprehend the rules and conditions that determine treatment predictions.
Secondly, decision trees excel at feature selection. During tree construction, features that provide the most information gain or decrease in impurity are prioritized for splitting. This feature selection capability allows decision trees to identify and prioritize the most important variables contributing to treatment predictions. Medical professionals can gain insights into which features have the most impact on patient survival.
Lastly, decision trees provide a rule-based modeling approach. The decision tree's structure comprises a series of if-else conditions that determine the path of prediction. This rule-based nature allows for a clear understanding of the decision process and facilitates the creation of rule-based models for cancer prediction. It becomes easier to interpret and communicate the decision rules derived from the decision tree, aiding in clinical decision-making.
In summary, decision trees provide interpretability, feature selection capabilities, and a rule-based modeling approach, making them a valuable tool for hypothesizing precise cancer treatments based on patient survival. By visualizing the decision-making process, identifying important features, and offering rule-based insights, decision trees aid in understanding the factors influencing treatment outcomes and guide medical professionals in making informed decisions.

3.4.3 Random Forest

An ensemble learning approach called Random Forest enhances the effectiveness and reliability of individual decision trees. It functions by building a group of decision trees and combining their predictions via voting or averaging. To lessen overfitting and improve the model's generalization skills, every decision tree in the Random Forest algorithm has been trained on a randomly selected sample of the training data, and for every node, a randomly chosen subset of features is taken into consideration.
Random Forest overcomes the constraints of single decision trees by combining the strength and collective wisdom of multiple decision trees. Random Forest can identify complicated links in the data that may be difficult for a single decision tree to detect by aggregating the predictions generated by multiple trees. The ensemble gains variety as a result of the randomness incorporated into the training process through feature subset selection and data sampling, which minimizes the probability of overfitting and enhances the algorithm's ability to adapt well to fresh data[24].

Model Equations and Formulations

The predicted outcome of a Random Forest model is obtained by combining the predicted outcomes of different decision trees. Let's denote the set of decision trees as T and the prediction of each tree as $h_i(x)$, where i represents the position of the

tree. The final prediction $\hat{y}$ for an input x is calculated as follows:

$$\hat{y} = \frac{1}{|T|} \sum (h_i(x)) \tag{3.2}$$

where $\sum$represents the sum of all decision trees in the Random Forest and $|T|$ is the overall number of trees in the ensemble.
This equation demonstrates that the overall prediction is equal to the mean of all the predictions made by every decision tree. The aggregation process combines the collective knowledge of the trees, resulting in a more robust and accurate prediction for cancer treatment outcomes.

Rationale for Model Selection

Random Forest models were chosen for several reasons. Firstly, they improve the performance and robustness of decision trees by reducing overfitting. By creating a set of decision trees and combining their predictions, Random Forest reduces the impact of individual noisy or biased trees, resulting in more accurate predictions. The randomness introduced during the training process helps to capture different aspects of the data, leading to a more generalized and robust model.
Secondly, Random Forest models can handle high-dimensional datasets effectively. With the ability to consider subsets of features at each node, Random Forest can efficiently handle a large number of features, which is often the case in cancer treatment datasets. Through this feature subset selection process, the model is able to concentrate on relevant attributes and reduce the influence of irrelevant or noisy features, improving the overall prediction performance.
Lastly, Random Forest provides valuable insights into feature importance. By evaluating the contribution of each feature across the ensemble of decision trees, Random Forest can rank the importance of features in predicting treatment outcomes. This information aids in understanding the underlying factors and variables that influence precise cancer treatment predictions. The feature importance rankings can guide medical professionals in identifying the most critical features for accurate predictions and in gaining insights into the underlying mechanisms that drive treatment outcomes.
In summary, Random Forest models are valuable in hypothesizing precise cancer treatments based on patient survival. By creating an ensemble of decision trees, Random Forest reduces overfitting, handles high-dimensional data, and provides insights into feature importance. The aggregation of predictions from multiple trees improves the model's robustness and accuracy, making Random Forest an effective tool in the field of cancer prediction and treatment.

3.4.4 Gaussian Naive Bayes

Naive Bayes is a classification algorithm that calculates the probability of a data point belonging to a particular class. It relies on Bayes' theorem, incorporating the prior probability of the class and the likelihood of the data point given the class. The model's simplicity, speed, and interpretability make it an attractive choice for a wide range of classification tasks. Naive Bayes calculates the probabilities based on the assumption that the features are conditionally independent given the class label, even though this assumption may not hold in real-world data. Despite this

limitation, Naive Bayes remains a powerful and versatile model that can provide valuable insights and accurate predictions for various applications [30].

Model Equations and Formulations:

The key equations for Gaussian Naive Bayes are as follows:

$$P\left(x_i \mid y\right) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{\left(x_i - \mu_y\right)^2}{2\sigma_y^2}\right) \tag{3.3}$$

The parameters σ_y and μ_y are estimated using maximum likelihood. The first part of the equation is the normalizing constant. It ensures that the probability of a data point being in a class sums to 1. The second part of the equation is the exponential term. It measures how likely it is for a data point to have a value x_i given that it belongs to class y.

Rationale for Model Selection:

We chose Naive Bayes for our classification task due to several compelling reasons. Firstly, it is a simple and easy-to-understand model, making it accessible to both researchers and practitioners. Additionally, Naive Bayes is computationally efficient and well-suited for handling large datasets, enabling faster processing times compared to more complex models. Its probabilistic nature provides a measure of uncertainty in predictions, which can be valuable for decision-making in real-world applications. While the independence assumption may not always hold, Naive Bayes often performs remarkably well in practice, delivering accurate results for various tasks. However, it is essential to consider the choice of features and assess the data's characteristics to maximize its effectiveness. Overall, Naive Bayes is a robust and versatile model that strikes a balance between simplicity and performance, making it a reliable choice for our classification problem.

3.4.5 SVM

Support Vector Machines (SVMs) are supervised learning algorithms used for classification and regression tasks. They work by finding the hyperplane that best separates two classes in a dataset. To achieve this, SVMs first map the data to a higher-dimensional space and then identify the hyperplane that maximizes the margin between the classes. The margin is the distance between the hyperplane and the closest data points of each class [7].

Model Equations and Formulations:

SVC solves the following primal problem:

$$\min_{w,b,\zeta} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \zeta_i \tag{3.4}$$

subject to $y_i \left(w^T \phi\left(x_i\right) + b\right) \geq 1 - \zeta_i$

$$\zeta_i \geq 0, i = 1, \ldots, n \tag{3.5}$$

Rationale for Model Selection:

We selected Support Vector Machines (SVMs) for our classification task due to their remarkable power and versatility. SVMs are well-suited for a wide range of classification tasks and have proven to be effective in various real-world applications. Additionally, SVMs are relatively simple models, making them easy to understand and interpret. Their ability to handle nonlinear data enables them to capture complex relationships between features, enhancing their predictive capabilities. SVMs excel in tasks where accuracy and simplicity are crucial, and they can provide reliable results even with limited training data. While SVMs may require careful tuning of parameters, the trade-off is often justified by their strong performance and robustness. Overall, SVMs are a powerful and flexible choice for our classification problem, offering accurate predictions and interpretability.

3.4.6 Gradient Boosting

Gradient Boosting is an ensemble technique that sequentially combines many weak prediction models, including decision trees. It functions by repeatedly training new models to fix the errors created by the previous models. Each new model emphasizes the cases that the prior models misclassified and gives these challenging-to-classify samples more weight. Gradient Boosting can steadily increase the accuracy of its predictions by focusing more on difficult examples thanks to this iterative approach. A new weak model, usually a decision tree, is introduced to the ensemble in each iteration. This model is taught to recognize the correlations and patterns in the data that the other models struggled to understand. A gradient descent optimization approach is used to train the model by minimizing a loss function that measures the mistakes committed by the current ensemble. Gradient Boosting successfully combines the individual strengths of these weak models to create a final forecast that is reliable and powerful by iteratively adding these weak models [18].

Model Equations and Formulations

The gradient is similar for M trees:

$$f_{min}(x) = f_{m-1}(x) + \left(\underset{i_{mcHi}}{\operatorname{argmin}} \left[\sum_{i=1}^{N} L\left(t_2, f_{m-1}\left(x_i\right) + h_m\left(x_i\right)\right) \right] \right)(x) \tag{3.6}$$

where $\sum$ is the cumulative number of weak models x_i. The factor x_i decides how much weight is given to each model's prediction, and the learning rate governs how much each weak model contributes to the overall forecast. By balancing the contribution of new models with that of the current ensemble, the learning rate enables the model to avoid over- or underfitting. Each model's forecast is given a weight based on how important it is to the ensemble as a whole.

Rationale for Model Selection

For several kinds of reasons, Gradient Boosting models were chosen. They are first renowned for their excellent prognostication. Gradient Boosting can capture complicated links and interactions within the data by sequentially merging numerous weak models. The model's capacity to handle complex situations and make precise

predictions is enhanced by the iterative training process, which enables the model to concentrate on hard examples that were incorrectly identified by the prior models.
Secondly, Gradient Boosting is particularly effective when combined with decision trees or other weak learners. Decision trees are capable of capturing complex patterns in the data, and Gradient Boosting enhances their performance by iteratively correcting their mistakes. This combination allows the model to capture fine-grained patterns, interactions, and nonlinear relationships, making it well-suited for precise cancer treatment predictions.
Lastly, Gradient Boosting offers flexibility and efficiency through various implementations such as Scikit-learn, XGBoost, or LightGBM. These libraries provide optimized algorithms, efficient training procedures, and the ability to tune model parameters. They also support parallel computing, which enables faster model training on large datasets. The availability of these libraries makes Gradient Boosting a practical and widely used choice in cancer prediction tasks.
In summary, Gradient Boosting is a powerful ensemble method for accurate cancer prediction. By combining multiple weak models and iteratively correcting mistakes, it captures complex relationships and improves predictive performance. The iterative training process and the weighting of model predictions enhance the model's ability to handle challenging instances. With various implementations available, Gradient Boosting offers flexibility, efficiency, and optimization for cancer prediction tasks.

3.4.7 AdaBoost

AdaBoosting, also known as adaptive boosting, is an ensemble technique that iteratively combines many weak models to increase overall prediction accuracy. It addresses difficult-to-classify samples by assigning higher weights to instances that were misclassified in previous iterations, placing more emphasis on challenging cases. This adaptive approach allows AdaBoosting to focus on areas where the model struggles, gradually refining its predictions and achieving better performance.
AdaBoosting assigns weights to each instance based on their classification performance. In each iteration, a weak model is trained on the updated dataset, where the weights of the misclassified instances are increased. The weak model is designed to focus on the challenging samples and improve its ability to classify them correctly [6]. Subsequent weak models are then trained with updated weights, and their predictions are combined using a weighted voting scheme.
The weights of the instances are changed during training to give misclassified samples more weight. This enables the weak models to focus on challenging scenarios and provide precise forecasts. Each weak model that is trained seeks to fix the errors caused by the prior models, progressively enhancing the performance of the ensemble as a whole [19].

Model Equations and Formulations

The final prediction in AdaBoosting is obtained by summing the predictions from each weak model, weighted by the corresponding model's importance factor. For example, in binary classification, the prediction H for an input x is given by:

$$H(x) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right) \tag{3.7}$$

After the weak classifier is trained, it updates the weight of each training example with the following formula:

$$D_{t+1}(i) = \frac{D_t(i) \exp\left(-\alpha_t y_i h_t\left(x_i\right)\right)}{Z_t} \tag{3.8}$$

Rationale for Model Selection

AdaBoosting models were selected for a number of factors. AdaBoosting first tackles hard-to-classify data by giving instances that were incorrectly categorized more weight. With this adaptive strategy, the model may concentrate on difficult situations, enhancing its capacity to manage intricate patterns and achieving greater accuracy. AdaBoosting's focus on demanding situations is especially pertinent in the area of accurate cancer therapy prediction since reliable forecasts for problematic cases are essential.
Secondly, AdaBoosting is effective when combined with weak learners, such as decision trees. Weak models are trained iteratively, with each model focusing on the misclassified instances from the previous iteration. By leveraging the strengths of multiple weak models, AdaBoosting creates a powerful ensemble that can capture intricate relationships and make accurate predictions [12].
Furthermore, AdaBoosting is known for its ability to handle imbalanced datasets. In cancer prediction, where the occurrence of specific cancer types may be rare, imbalanced data can pose a challenge. AdaBoosting's weighting mechanism allows it to assign higher importance to minority classes, ensuring that the model pays sufficient attention to the underrepresented cases and produces reliable predictions.
In summary, AdaBoosting is a powerful ensemble method for precise cancer treatment prediction. Its adaptive approach to focusing on difficult-to-classify samples, combination with weak learners, and ability to handle imbalanced datasets make it a valuable tool. By iteratively refining the predictions and weighing the contributions of each weak model, AdaBoosting achieves higher accuracy and provides insights into challenging cases, contributing to the advancement of cancer treatment strategies.

3.4.8 HistGradientBoosting

Histologic Gradient Boosting is an advanced variant of gradient boosting that incorporates histogram-based algorithms. It excels at efficiently handling large-scale datasets with high-dimensional features. By leveraging histograms, this model constructs tree models that capture intricate patterns and interactions, making it a valuable tool in cancer prediction tasks.
In the working process of Histologic Gradient Boosting, the model employs histogram-based algorithms to efficiently construct tree models. It starts by creating histograms of feature values, which provide insights into the distribution and patterns within the data. These histograms serve as a summary representation of the data, capturing important information without relying on individual data points.

The histograms guide the tree construction process by enabling efficient splits based on the histogram information.
The construction of tree models using histograms allows Histologic Gradient Boosting to overcome the computational challenges associated with large-scale datasets and high-dimensional features. By discretizing feature values into bins and constructing histograms, the model reduces the complexity of the data representation. Avoiding the need to analyze individual feature values for each split, which may be computationally costly for big datasets, enables more effective and scalable tree building [29].

Model Equations and Formulations

The mathematical equations for Histologic Gradient Boosting may vary depending on the specific implementation and library used. These equations involve the optimization process and the construction of tree models using histogram-based algorithms. The exact equations depend on the loss function being optimized, the specific histogram construction techniques, and the algorithmic details implemented in the library. It is important to consult the documentation and resources of the chosen implementation to gain a more detailed understanding of the equations and algorithmic specifics.
The equations in Histologic Gradient Boosting involve the optimization of the loss function through gradient descent and the construction of tree models that recursively partition the data based on histogram-guided splits. Each tree model contributes to the ensemble prediction by providing a weighted contribution based on its performance and importance.

Rationale for Model Selection

Histologic Gradient Boosting models were chosen for their ability to handle large-scale datasets with high-dimensional features efficiently. The incorporation of histogram-based algorithms allows for a more scalable and computationally efficient approach to constructing tree models. By summarizing the data using histograms, the model reduces the computational burden associated with evaluating individual feature values for each split, making it well-suited for large-scale datasets.
In the context of cancer prediction tasks, where complex patterns and interactions may exist within the data, Histologic Gradient Boosting's ability to capture intricate patterns through histogram-based tree construction is particularly valuable. The model can efficiently handle high-dimensional feature spaces and uncover fine-grained relationships, leading to more accurate predictions.
Furthermore, Histologic Gradient Boosting's focus on histograms allows it to handle diverse data types, including both categorical and continuous features. This flexibility enables the model to effectively capture the unique characteristics of cancer-related datasets, where a combination of different types of features may be present.
In summary, Histologic Gradient Boosting is an advanced variant of gradient boosting that leverages histogram-based algorithms for efficient and accurate modeling. By efficiently handling large-scale datasets with high-dimensional features, Histologic Gradient Boosting provides valuable insights into cancer-related outcomes. The construction of tree models guided by histograms enables the model to capture intricate patterns and interactions, contributing to precise cancer prediction.

3.4.9 Multilayer Perceptron

The Multi-Layer Perceptron (MLP) is an artificial neural network that comprises multiple layers of interconnected nodes, also known as neurons. It is a powerful model capable of learning complex patterns and capturing nonlinearity from data. An MLP has an input layer, one or more hidden layers, and an output layer among its layers. Each neuron in the network is connected to neurons in adjacent layers through weighted connections. During the training process, to reduce the discrepancy between projected and actual outputs, these weights are modified based on input data.
The hidden layers shown in figure 3.8 in MLP serve as intermediate layers between the input and output layers. Each neuron in the hidden layers first applies a weighted total of inputs, and then it applies an activation function. This nonlinearity introduced by the activation function allows the model to capture complex relationships and make nonlinear predictions. The choice of activation function can vary depending on the problem domain, with popular options including sigmoid, tanh, and ReLU [5].

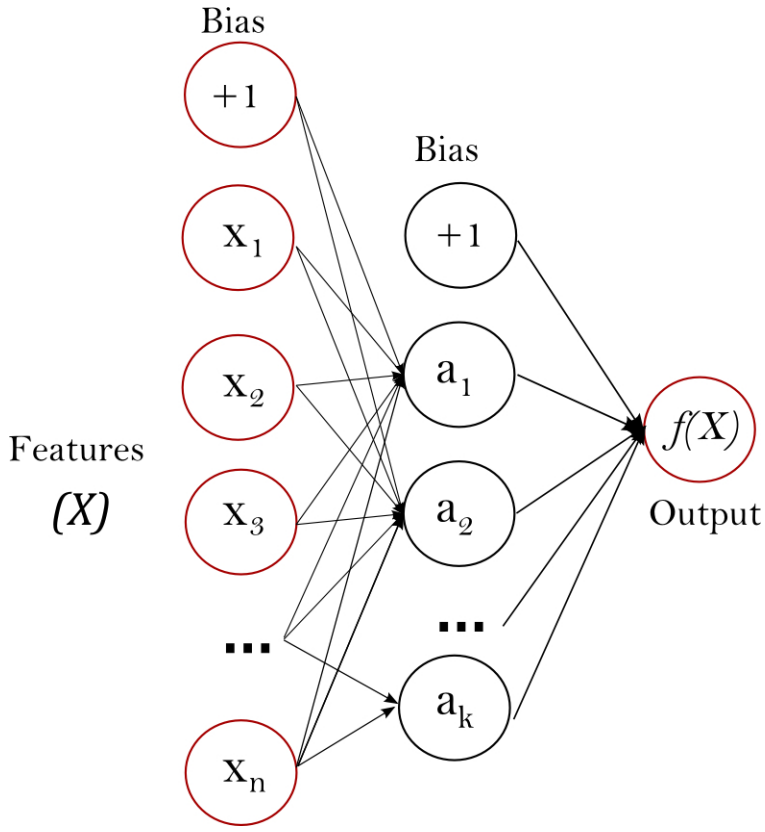

Figure 3.8: One Hidden Layer of MLP

Model Equations and Formulations

Given a set of training examples $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ where $x_i \in \mathbf{R}^n$ and $y_i \in \{0, 1\}$, a one hidden layer one hidden neuron MLP learns the function $f(x) = W_2 g\left(W_1^T x + b_1\right) + b_2$ where $W_1 \in \mathbf{R}^m$ and $W_2, b_1, b_2 \in \mathbf{R}$ are model parameters. W_1, W_2 represent the weights of the input layer and hidden layer, respectively; and b_1, b_2 represent the bias added to the hidden layer and the output layer, respectively.

$g(\cdot) : R \rightarrow R$ is the activation function, set by default as the hyperbolic tan. It is given as,

$$g(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}} \tag{3.9}$$

For binary classification, $f(x)$ passes through the logistic function $g(z) = 1/(1 + e^{-z})$ to obtain output values between zero and one.
The process of propagating inputs through the layers of neurons is performed using matrix multiplications and element-wise activation function applications. This allows for efficient computation and parallel processing, especially when working with large datasets.

Rationale for Model Selection

MLP models were chosen for their remarkable ability to learn complex patterns and capture non-linearity from data. As neural networks, MLP models can automatically discover relevant features and patterns, making them highly valuable in cancer prediction tasks, particularly when dealing with large and high-dimensional datasets. The multi-layer structure of MLP allows for the representation of complex relationships between input variables and output predictions.
In order to reduce the discrepancy between expected and actual outputs, the iterative training process of MLP, which frequently employs back-propagation, modifies the weights and biases throughout training. The weights and biases are updated depending on the calculated gradients using gradient-based optimization algorithms, such as stochastic gradient descent (SGD). The convergence and performance of the model can be affected by the optimization algorithm selection and associated hyperparameters.
Additionally, MLP models offer flexibility in terms of the choice of activation functions, allowing for customization based on the problem domain. This enables the model to capture and model the nonlinear relationships that may exist in cancer-related data. The ability of MLP models to handle nonlinearity and automatically learn relevant features from the data makes them a powerful tool in cancer prediction and treatment, where the identification of complex relationships and patterns is crucial for precise treatment strategies.
In summary, the Multi-Layer Perceptron (MLP) is a versatile and powerful model used for cancer prediction. With its ability to learn complex patterns, capture non-linearity, and automatically discover relevant features, MLP models provide valuable insights and accurate predictions for cancer treatment outcomes. By utilizing the mathematical equations, training process, and deep knowledge of MLP models, researchers, and practitioners can leverage its capabilities to advance the field of cancer prediction and contribute to the development of precise treatment strategies.

Chapter 4

Result & Analysis

This section contains the thorough findings and in-depth analysis of our study on hypothesizing the effectiveness of cancer treatments based on patient survival data. We used nine different models from three different learning techniques, including Logistic Regression, Decision Tree Classifier, Random Forest, Gaussian Naive Bayes, SVM, Gradient Boosting, AdaBoost, HistGradientBoosting, and Multilayer Perceptron by using the capability of machine learning. These models' accuracy was carefully assessed to determine how well they could estimate treatment outcomes. We learn a lot about the benefits and drawbacks of each strategy by contrasting the outcomes of various therapies. Our findings not only provide insight into the effectiveness of these models but also highlight crucial elements that affect treatment assumptions. The results of this study have important significance for the development of personalized treatment for cancer, giving healthcare professionals the information they need to choose the best course of action and ultimately improving patient outcomes and the battle against cancer.

4.1 Model Evaluation

The graph 4.1 provides valuable insights into the performance of various machine learning models for predicting precise cancer treatment based on the type of treatment. Each model's accuracy is presented, highlighting its strengths and weaknesses in different treatment scenarios. Understanding the implications of these results can aid in selecting the most suitable model for specific cancer treatment predictions, ultimately enhancing personalized and effective patient care.

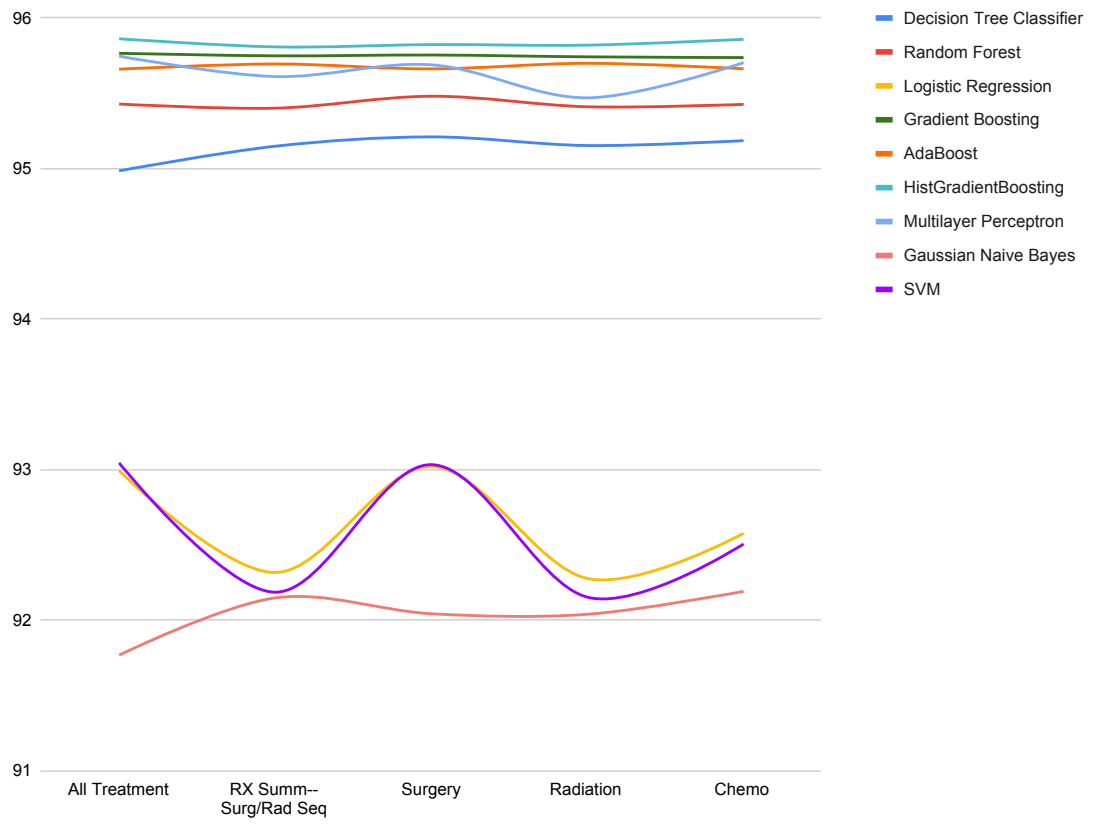

Figure 4.1: Accuracy of the models in regards to Treatment

Figure 4.1 displays the accuracy of various machine learning models in predicting precise cancer treatment, categorized by the type of treatment. Among the models, HistGradient Boosting exhibits the highest accuracy (95.86%) for all treatment types, closely followed by the gradient boosting model (95.77%) and the MLP model (95.75%).

In contrast, the decision tree model achieves an accuracy of 95.21% for predicting the survivability of surgery patients, and the logistic regression model has an accuracy of 93.02%. The Gaussian Naive Bayes model has an accuracy of 92.19% for predicting the survivability of patients undergoing chemotherapy, and the SVM model achieves an accuracy of 93.04% for predicting survivability across all treatments.

Interestingly, the accuracy of the models varies depending on the type of treatment being predicted. For instance, while HistGradient Boosting performs exceptionally well in predicting surgery outcomes, its accuracy drops when predicting radiation therapy outcomes. Similarly, the AdaBoost model excels in predicting radiation therapy survivability but performs relatively lower in surgery predictions.

This indicates that different models are better suited for predicting specific types of cancer treatments. The choice of the optimal model for a given treatment prediction relies on understanding the strengths and limitations of each model concerning the treatment characteristics. However, the dataset used during model training plays a crucial role in performance. Since the models were trained on text documents representing various cancer treatments, their accuracy may vary when predicting treatments for cancer types not well-represented in the dataset.

In conclusion, the graph showcases the potential of machine learning in accurate

cancer treatment prediction, with Boosting algorithms and Multilayer Perceptron standing out for their performance. Notably, the Boosting algorithms, particularly HistGradient Boosting, and gradient boosting, demonstrate the highest accuracy, outperforming other models. The Multilayer Perceptron also shows competitive accuracy akin to the Boosting models. Nevertheless, careful consideration of diverse cancer types and datasets is essential for maximizing the models' effectiveness in clinical applications.

4.1.1 Logistic Regression

The Logistic Regression model is a statistical model that is used to predict the probability of an event occurring. In the case of cancer treatment survival prediction, the Logistic Regression model can be used to predict the probability of a patient surviving their cancer treatment.
The Logistic Regression model works by fitting a line to the data. The line is called the logistic curve, and it represents the relationship between the features in the data and the probability of the event occurring.
The decision trees in the forest are trained independently of each other, which helps to prevent overfitting. Overfitting is a problem that occurs when a machine learning model learns the noise in the data instead of the underlying patterns. This is achieved by using regularization. Regularization is a technique that penalizes the model for making large predictions. This helps to prevent the model from overfitting the data.
To make a final prediction, the Logistic Regression model calculates the probability of the event occurring for each data point. The data point with the highest probability is the one that the model predicts will occur.
The Logistic Regression model is also interpretable, which means that the coefficients of the model can be interpreted. This can help to understand the factors that influence the probability of the event occurring. A coefficient that is positive indicates that the feature is associated with an increased probability of the event occurring, while a coefficient that is negative indicates that the feature is associated with a decreased probability of the event occurring. For example, the coefficient for treatment A is positive. This means that patients who are treated with treatment A are more likely to survive their cancer treatment than patients who are not treated with treatment A.
The Logistic Regression model has been shown for cancer treatment survival prediction with our dataset. The model was able to achieve an overall accuracy of 92.99%. This means that the model was able to correctly predict whether a patient would survive their cancer treatment with 92.99% accuracy.
Overall, the Logistic Regression model is a powerful and versatile algorithm for cancer treatment survival prediction. It is interpretable, and it can be used to predict the probability of the event occurring for each data point. These make it an ideal choice for this application.

The accuracy scores obtained in each fold are shown on the Y-axis in figure 4.2, while the cross-validation's multiple folds are shown on the X-axis. The accuracy rating is 0.9470 for all folds combined. The maximum accuracy score, which is over 0.950 for Fold 5, is distinguished from the lowest accuracy value, which is less than

0.945 for Fold 5.

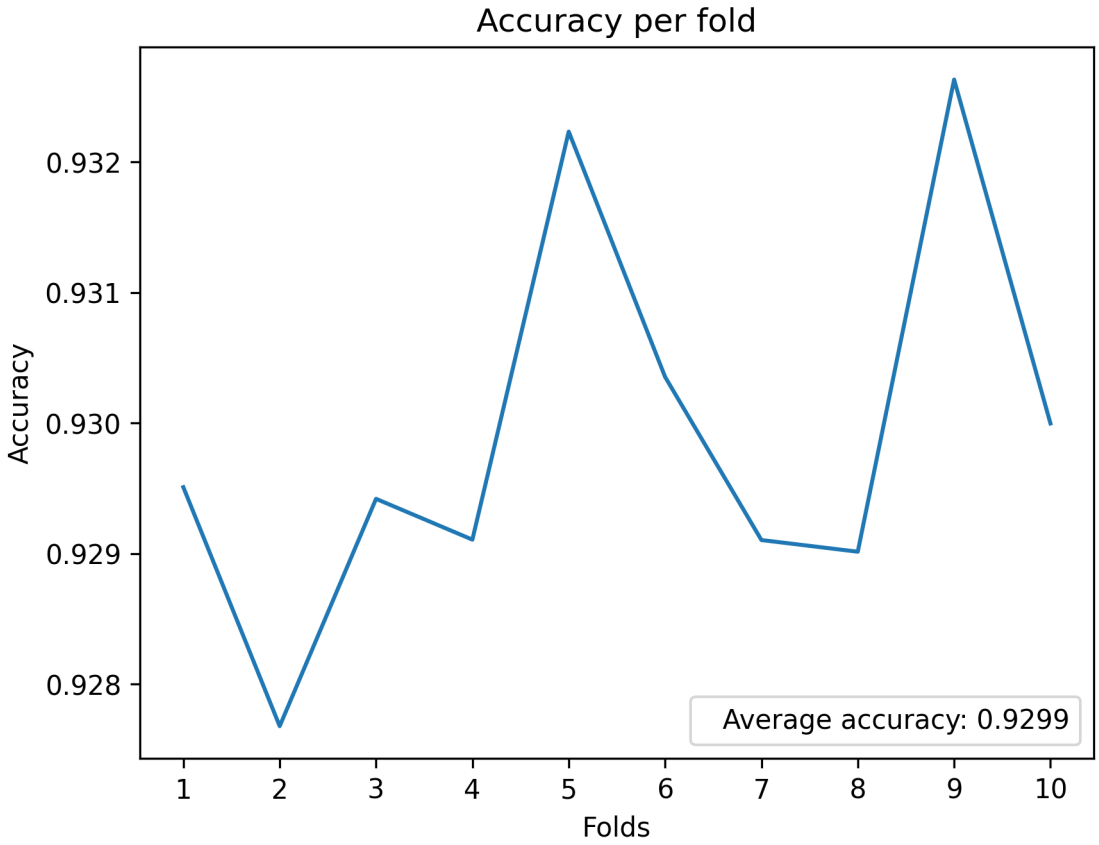

Figure 4.2: Accuracy of Logistic Regression

In Figure 4.3, the X-axis represents the cross-validation folds that were used, and the Y-axis represents the F1 scores that were acquired for each fold. The average F1 score for all folds combined is 0.95. The accuracy score for Fold 10 in particular is close to 0.9520 and it's the height score, whereas the lowest accuracy value for Fold 2 is nearly 0.9485.

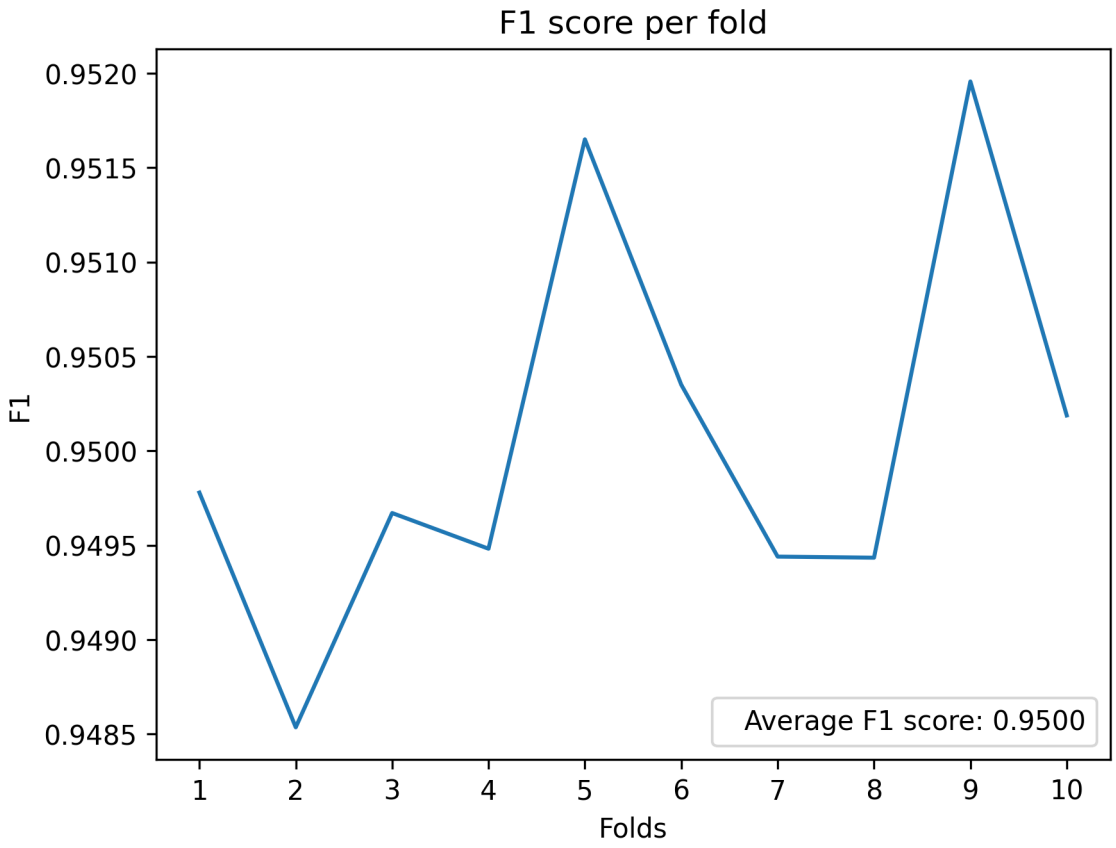

Figure 4.3: F1 Score of Logistic Regression.

All 10 of the ROC curves are located in the upper-left corner of Figure 4.4. Each fold's AUC, which ranges from 0.9731 to 0.9755, is similarly quite high. This further demonstrates the effectiveness of the logistic regression model.

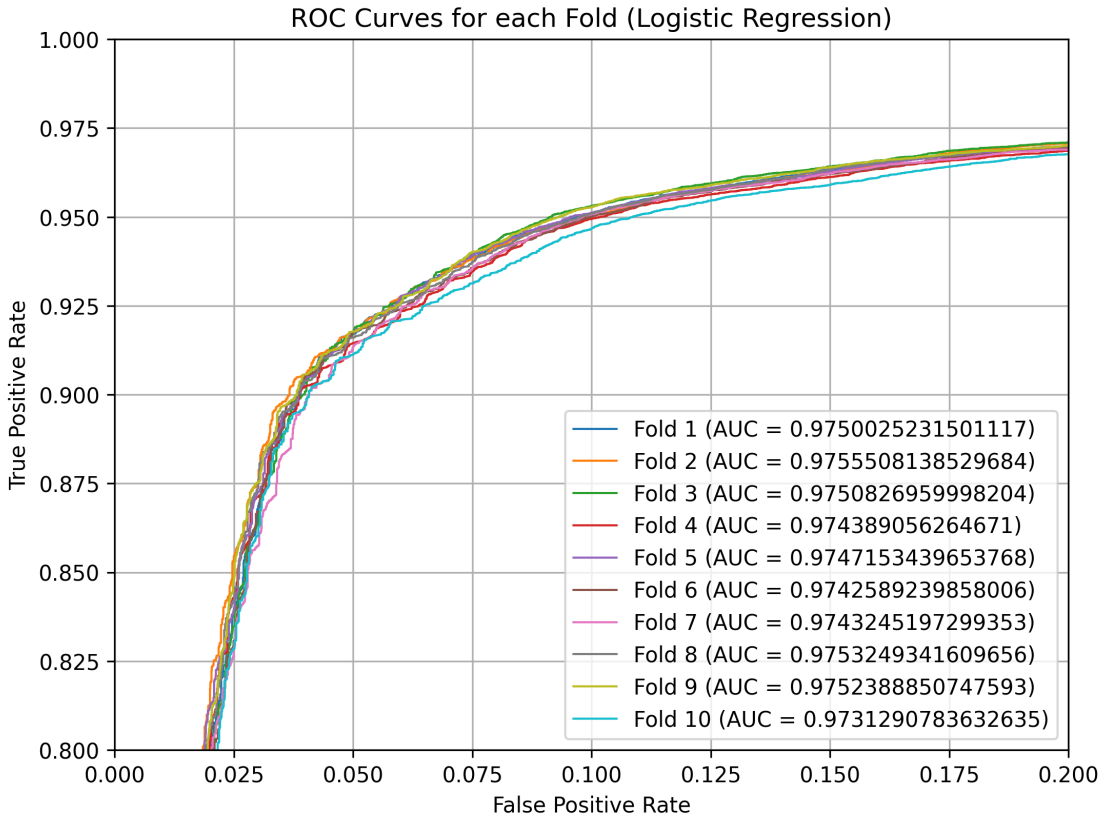

Figure 4.4: ROC Curve of Logistic Regression

This graph 4.5 displays the confusion matrix. We can evaluate the effectiveness of our categorization model with the aid of this example.

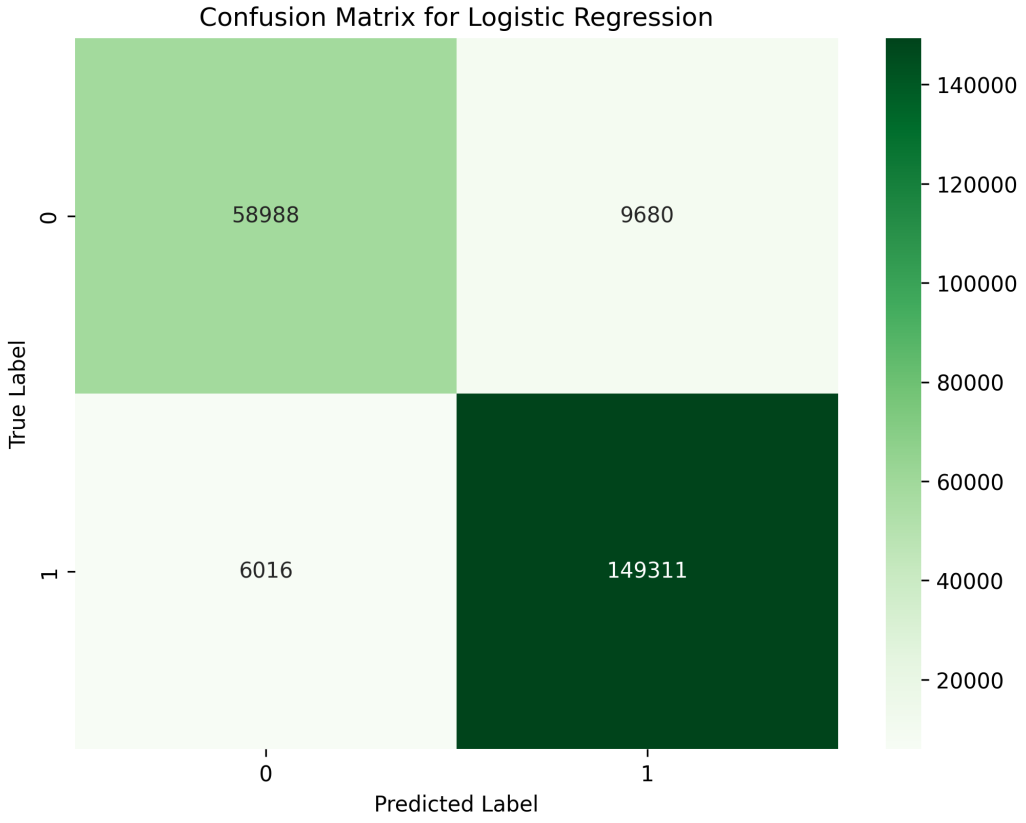

Figure 4.5: Confusion Matrix of Logistic Regression

The graph 4.6 displays a graphical representation in which the X-axis lists numerous cancer therapy types and the Y-axis provides the related accuracy ratings for each kind of therapy. The decision tree's efficiency is assessed for each treatment category. Notably, the surgical procedure has an extraordinary level of precision that approaches over 93. On the other hand, it may be proven that radiation therapy has a lower degree of accuracy that is near the threshold of 92.25.

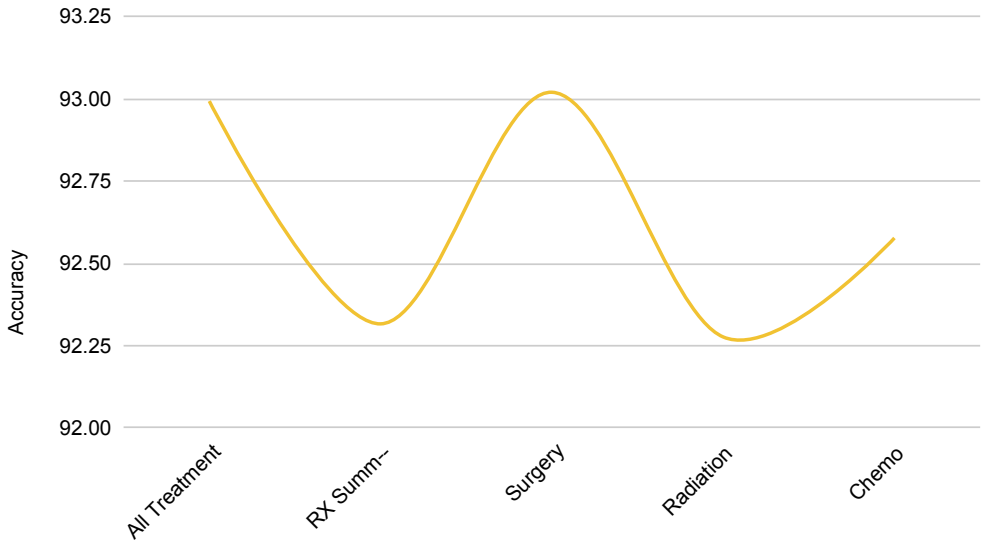

Figure 4.6: Accuracy of Logistic Regression in Regards Treatment

4.1.2 Decision Tree Classifier

The root node of the Decision Tree algorithm is created by selecting the most significant feature from the dataset. It then partitions the data recursively into subsets based on the values of the selected feature, creating child nodes that represent the various branches of the tree. This procedure is repeated until the algorithm reaches a predetermined stopping criterion, such as a maximum depth, a minimal number of samples per leaf, or when all samples belong to the same class. During node splitting, the algorithm evaluates multiple criteria, including Gini impurity and entropy, to determine the optimal feature and value for splitting, to minimize impurity and maximize information gain in each split. Once the tree has been constructed, it employs the learned decision rules to classify new data points, traversing from the root to a leaf node to represent the predicted outcome for each data point. The Decision Tree is a valuable algorithm for our cancer treatment survival prediction due to its simplicity, interpretability, and ability to manage both numeric and categorical features.

Common in medical datasets are numerical and categorical features, which can be handled by the Decision Tree algorithm. Decision Trees are also relatively simple to interpret and visualize, making them valuable tools for obtaining insight into the factors that influence cancer treatment survival prediction predictions. In addition, Decision Trees can manage non-linear relationships and interactions between features, enabling them to capture complex data patterns.

As medical professionals can readily interpret the rules governing the model's decisions, the decision tree is an attractive option for medical applications due to its simplicity and transparency. Additionally, Decision Trees are computationally efficient and perform well on comparatively small to medium-sized datasets, which is advantageous for small sample-size medical datasets.
Nonetheless, Decision Trees are susceptible to overfitting, especially when the tree depth is not constrained or when the dataset is unbalanced. To prevent overfitting, we employed hyperparameter tuning and appropriate stopping criteria to achieve the optimal equilibrium between model complexity and generalizability.
From our working with this model in our dataset, we got 94.98% accuracy on overall treatment types which is quite a good result to come with an output.
Overall, the Decision Tree algorithm was an integral part of our analysis, providing valuable insights into the factors influencing cancer treatment survival and providing a reliable baseline model against which to compare other more complex machine learning models. Its interpretability, versatility, and user-friendliness made it an ideal choice for the initial phases of our project, complementing the more complex models in our exhaustive analysis.

The X-axis of Figure 4.7 depicts the various folds used in cross-validation, whereas the Y-axis displays the accuracy scores obtained in each fold. The aggregate average accuracy score for all folds is 0.9470. In particular, fold 5 shows the maximum accuracy score, approaching 0.950, while the most minimal possible level of accuracy falls below 0.945.

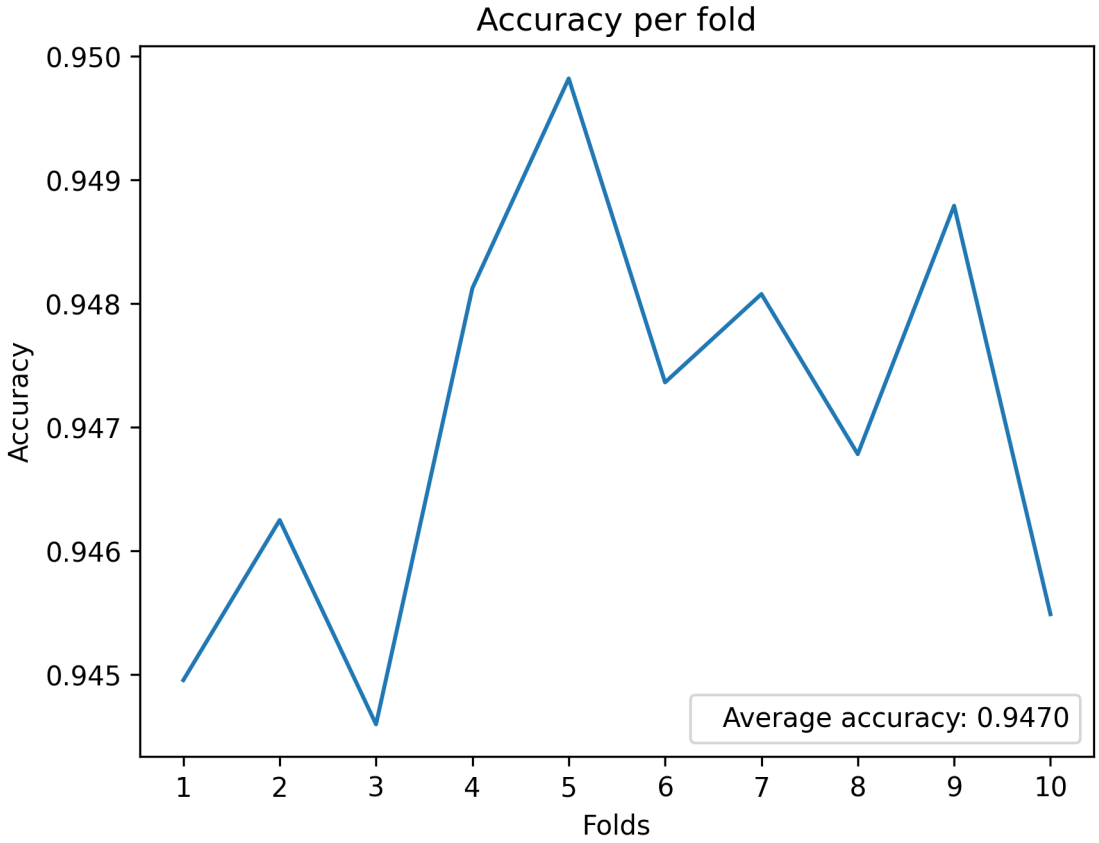

Figure 4.7: Accuracy of Decision Tree
The X-axis of Figure 4.8 depicts the various folds used in cross-validation, whereas

the Y-axis displays the F1 scores obtained in each fold. The aggregate average F1 score for all folds is 0.9615. In particular, fold 5 shows the maximum accuracy score, approaching 0.9635, while the most minimal possible level of accuracy falls below 0.9600.

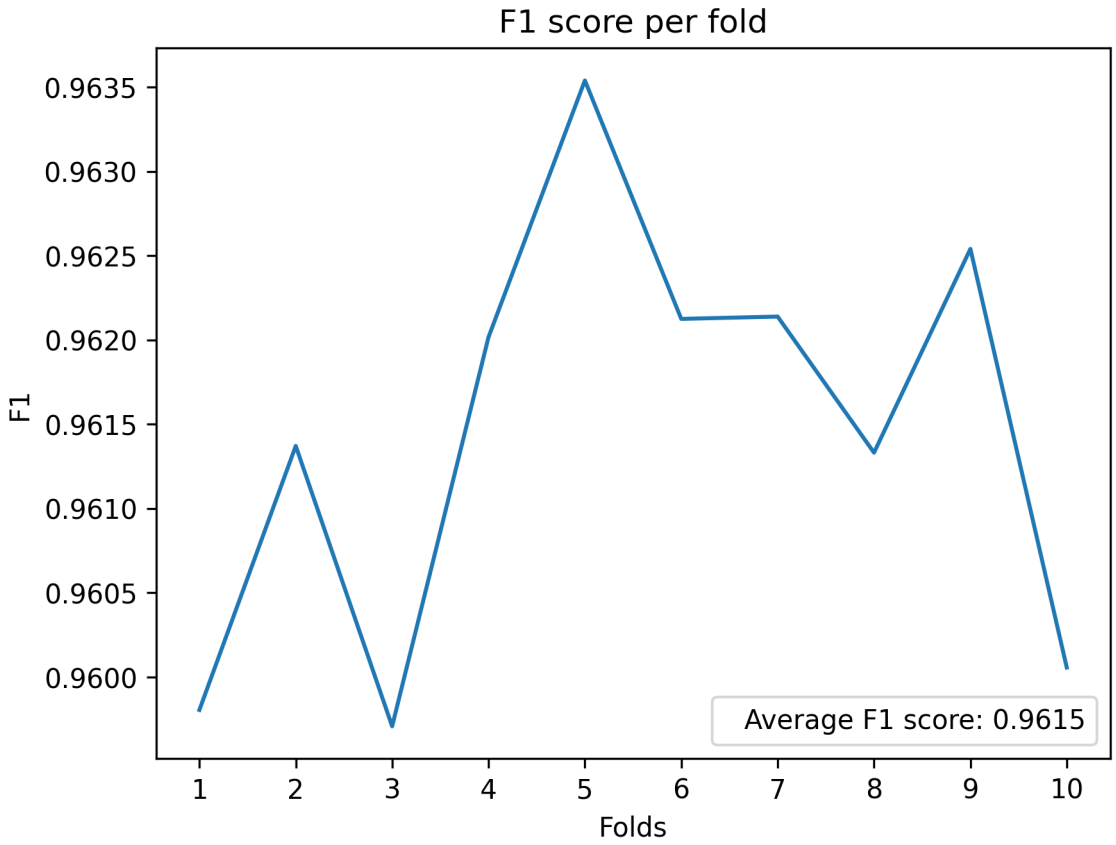

Figure 4.8: F1 score of Decision Tree

In Figure 4.9, the ROC curve shows how a decision tree did on 10 folds of data. In the legend, the AUC (area under the curve) for each fold is shown. All 10 of the ROC curves are very close to the top left corner. This shows that the decision tree works effectively across all of the data folds. The AUC for every fold is also exceptionally high, ranging between 0.9589 and 0.9622.

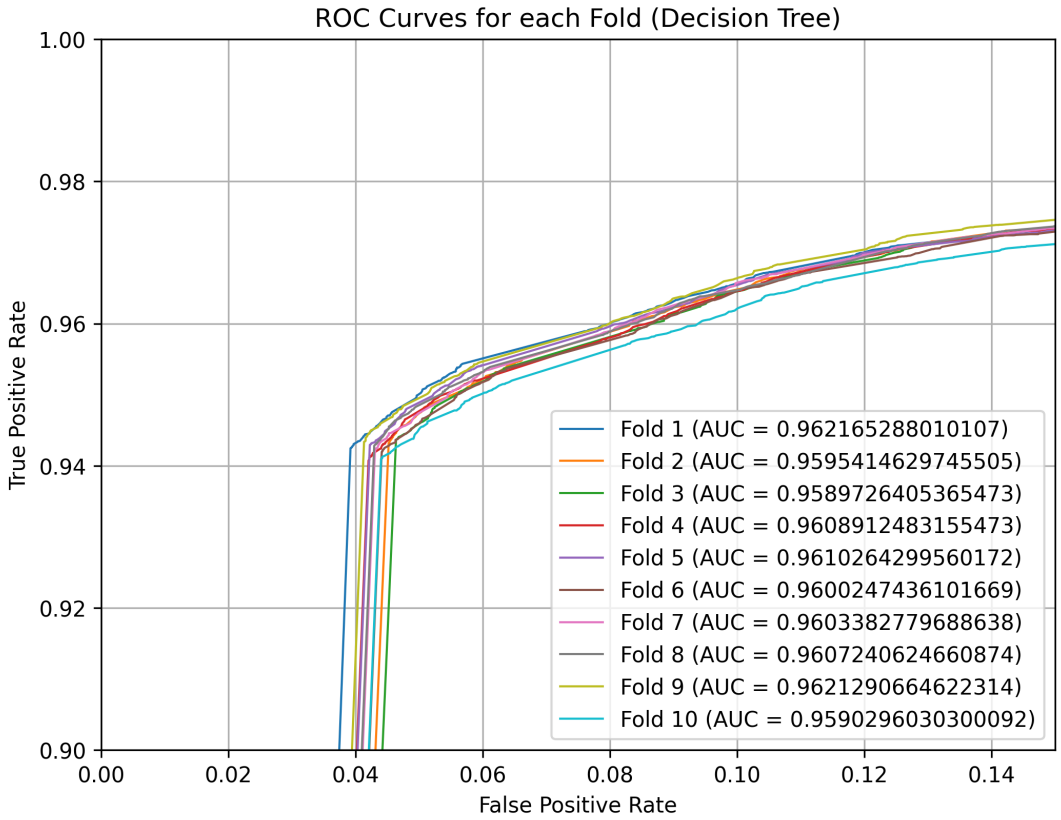

Figure 4.9: ROC Curve of Decision Tree

We can see the confusion matrix in figure 4.10. This is an illustration that enables us to evaluate the effectiveness of our categorization model.

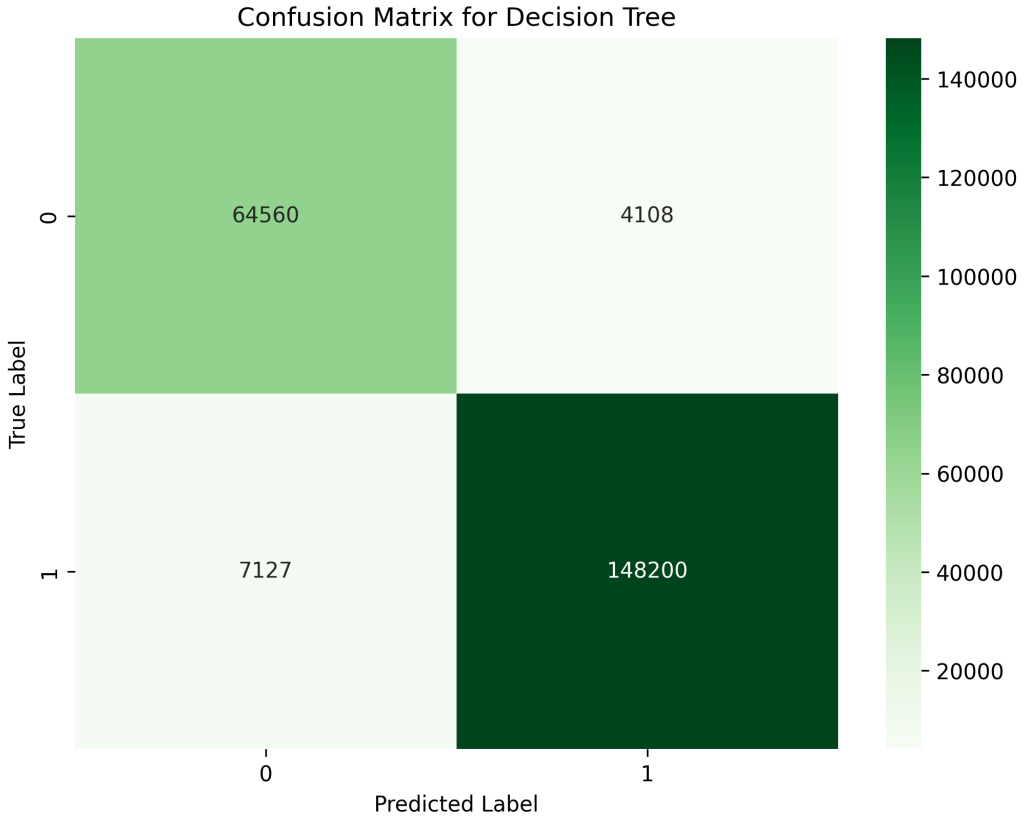

Figure 4.10: Confusion Matrix of Decision Tree

Figure 4.11 displays a graphical representation in which the X-axis indicates various types of treatments utilized in cancer prediction, and the Y-axis reflects the related accuracy ratings. The performance of the decision tree is assessed for every therapy category. Significantly, the surgical intervention demonstrates a remarkable level of accuracy, exceeding 95.2. In contrast, it can be shown that all alternative therapies demonstrate a diminished level of accuracy, descending below the limit of 95.

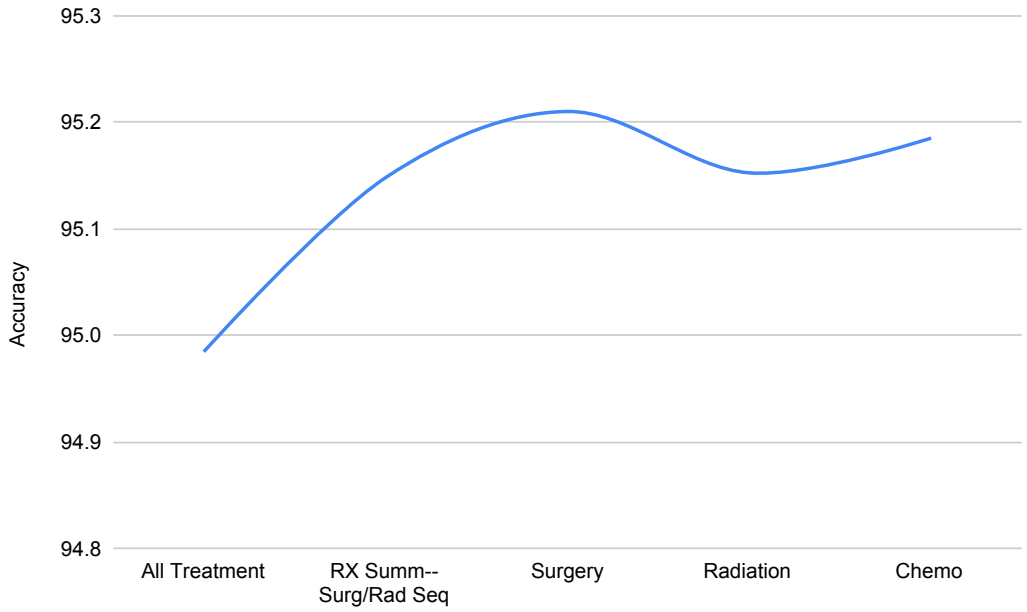

Figure 4.11: Accuracy of Decision Tree in regards Treatment

4.1.3 Random Forest

The Random Forest model works by randomly selecting a subset of features from the dataset and then building a decision tree using only the selected features. This process is repeated multiple times, resulting in a collection of decision trees.

The decision trees in the forest are trained independently of each other, which helps to prevent over-fitting. Over-fitting is a problem that occurs when a machine learning model learns the noise in the data instead of the underlying patterns. This is achieved by using bootstrap sampling to create training sets for the individual decision trees. Bootstrap sampling is a resampling method that randomly samples data with replacement. This helps to ensure that the decision trees are trained on a representative sample of the data, which helps to prevent over-fitting.

To make a final prediction, the Random Forest model averages the predictions of the individual decision trees. This helps to reduce the variance of the predictions, resulting in a more accurate model. The variance of a model is a measure of how much the model's predictions vary. A low variance model is a model that makes consistent predictions.

The Random Forest model is also interpretable, which means that the individual decision trees can be visualized. This can help to understand the factors that influence cancer treatment survival. For example, a decision tree might show that patients with a certain type of tumor are more likely to survive than patients with a different

type of tumor. This can help medical professionals to make better decisions about cancer treatment.
In our dataset, the model accuracy was quite impressive and it shows almost 95.42% accuracy.
Overall, the Random Forest model is a powerful and versatile algorithm for cancer treatment survival prediction. It is robust to overfitting, interpretable, and versatile. These make it an ideal choice for this application.

The Y-axis in Figure 4.12 shows the accuracy scores attained in each fold, whereas the X-axis in Figure shows the numerous folds utilized in cross-validation. For all folds combined, the accuracy score is 0.9528. Fold 5 in particular has the highest accuracy score, which is close to 0.956, while the lowest accuracy score is less than 0.951.

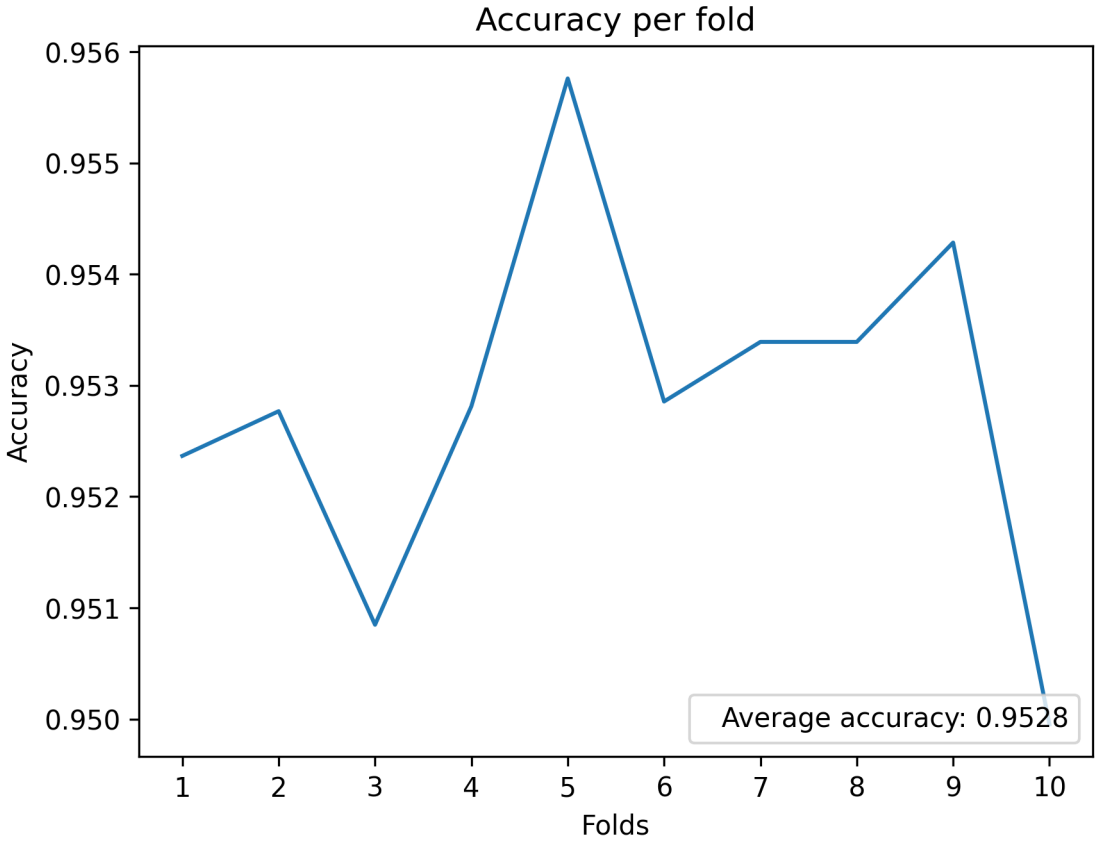

Figure 4.12: Accuracy of Random Forest
The cross-validation folds utilized are shown on the X-axis of figure 4.13, and the F1 scores obtained for each fold are shown on the Y-axis. For all folds combined, the average F1 score is 0.9528. Fold 5 in particular has the highest accuracy score, which is close to 0.956, while the lowest accuracy score is less than 0.951.

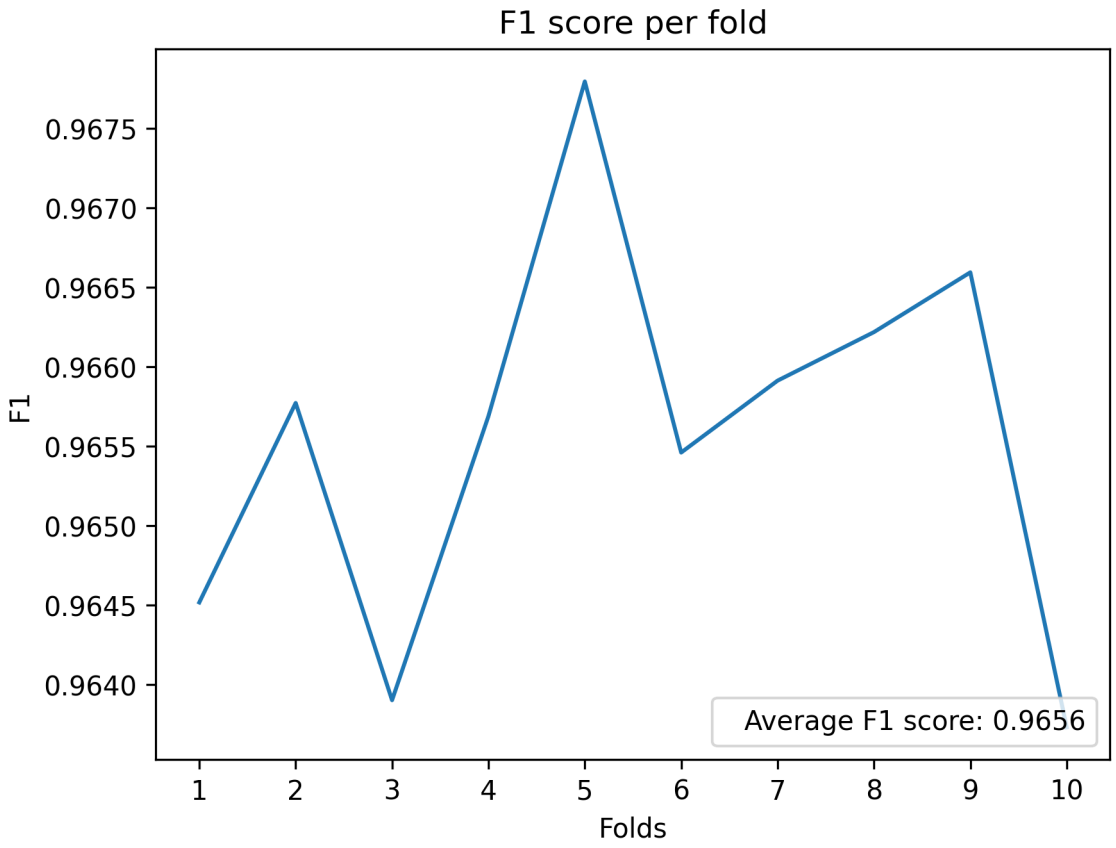

Figure 4.13: F1 Score of Random Forest

All 10 of the ROC curves are very close to the upper left corner of Figure 4.14. This shows that the random forest classifier does a great job with all types of data. The AUC for each fold is also quite high, running from 0.9882 to 0.9894. This is more proof that the random forest algorithm is doing excellently.

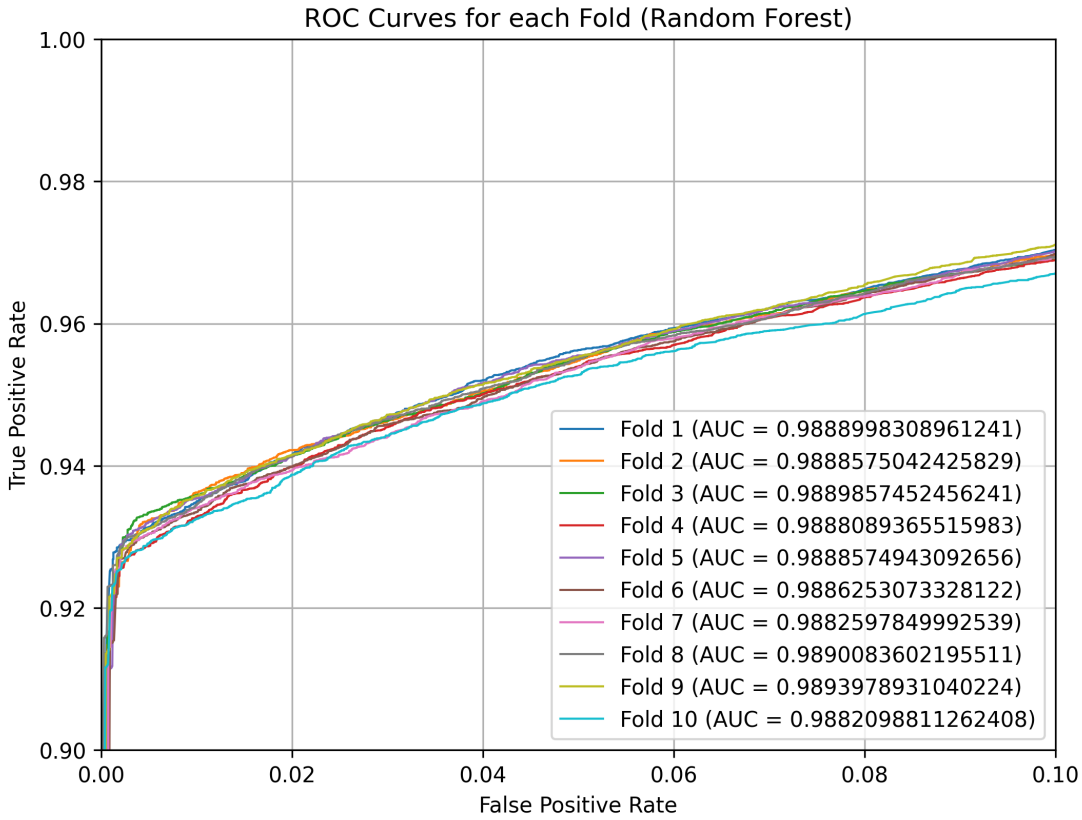

Figure 4.14: ROC Curve of Random Forest

The confusion matrix is shown in figure 4.15. With the help of this example, we can assess how well our classification model works.

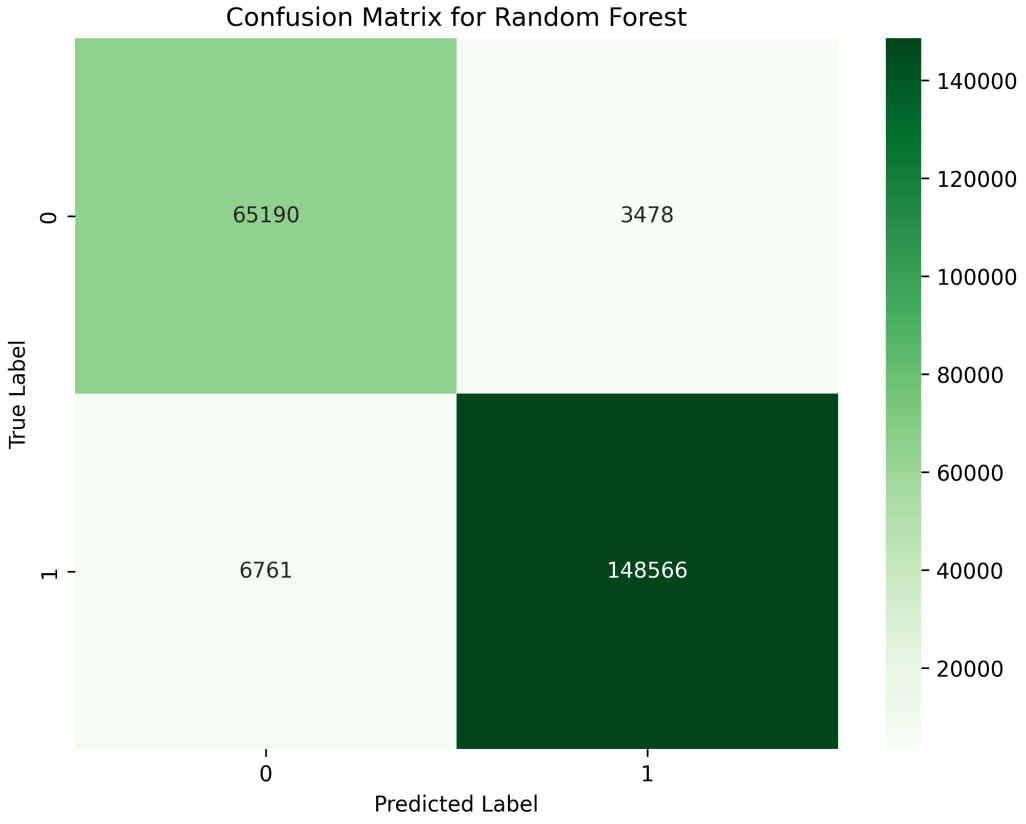

Figure 4.15: Confusion Matrix of Random Forest

The graph 4.16 shows a graphical depiction in which the Y-axis represents the corresponding accuracy ratings and the X-axis indicates various therapy kinds used to forecast cancer. For each therapeutic category, the decision tree's effectiveness is evaluated. Notably, the surgical intervention exhibits an exceptional degree of accuracy that exceeds 95.45. Contrarily, it can be demonstrated that the Surgery Radiation Sequence has a decreased degree of accuracy that falls to the point of 95.40.

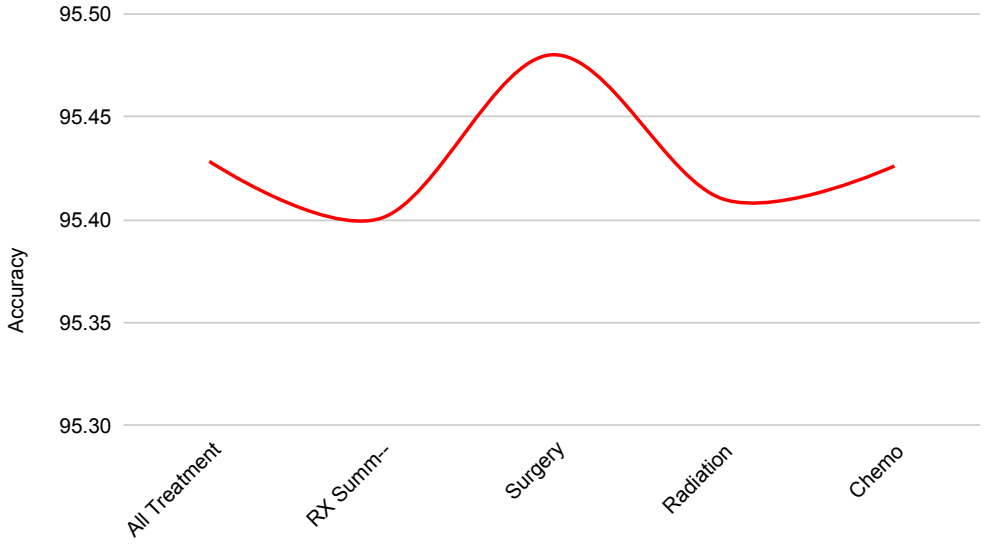

Figure 4.16: Accuracy of Random Forest in Regards Treatment

4.1.4 Gaussian Naive Bayes

The Gaussian Naive Bayes (GNB) model is a supervised learning algorithm. Naive Bayes algorithms are based on the Bayesian theorem, which is a mathematical formula that can be used to calculate the probability of an event occurring given the occurrence of other events.

The GNB model makes the naive assumption that the features in the data are independent of each other. This means that the GNB model assumes that the probability of one feature occurring does not affect the probability of another feature occurring.

The GNB model is a non-parametric model, which means that it does not make any assumptions about the distribution of the features in the data. This makes the GNB model a robust model that can be used to predict the probability of an event occurring in a variety of different domains.

The GNB model is a fast model, which means that it can be trained and used to make predictions quickly. This makes the GNB model a good choice for applications where speed is important.

The GNB model is trained on a dataset of known outcomes. The model learns the distribution of each feature in the dataset. This means that the model learns the probability that a feature will have a certain value.

The GNB model can then be used to predict the probability of an event occurring for a new data point. The model does this by calculating the posterior probability of

the event occurring. The posterior probability is calculated using the Bayes theorem and the distribution of the features in the new data point.
The GNB model is interpretable to some extent. The distribution of each feature in the model can be interpreted as the importance of the feature in the prediction. For example, a feature with a wide distribution is more important than a feature with a narrow distribution.
The GNB model has been shown to be effective for cancer treatment survival prediction with our dataset. The model was able to achieve an overall accuracy of 91.76%.
Overall, the GNB model is a powerful and versatile algorithm for cancer treatment survival prediction. It is interpretable to some extent, and it can be used to predict the probability of the event occurring for each data point. These make it an ideal choice for this application.

The X-axis of figure 4.17 represents the many cross-validation folds, while the Y-axis displays the accuracy scores attained in each fold. The total average accuracy of the folds is 0.9184. Fold 9's accuracy is nearly 0.9521, whereas Fold 2's is at the lowest accuracy which is 0.916.

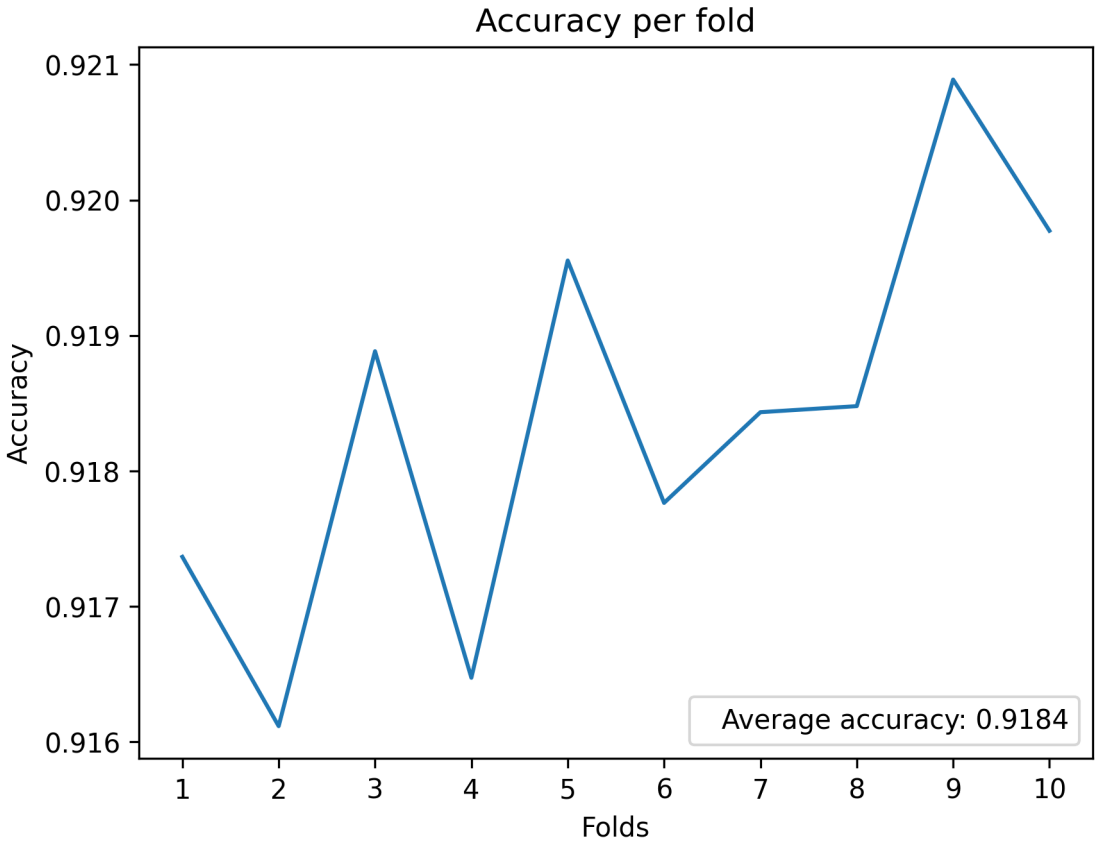

Figure 4.17: Accuracy of Gaussian Naive Bayes
In this figure 4.18 F1 Folds scores are plotted along the Y axis, while the number of cross-validation folds is shown along the X axis. A total F1 score of 0.9416 is achieved by f1 fold. Fold 9 has a height score over 0.9430 whereas Fold 2 has a number near 0.9400.

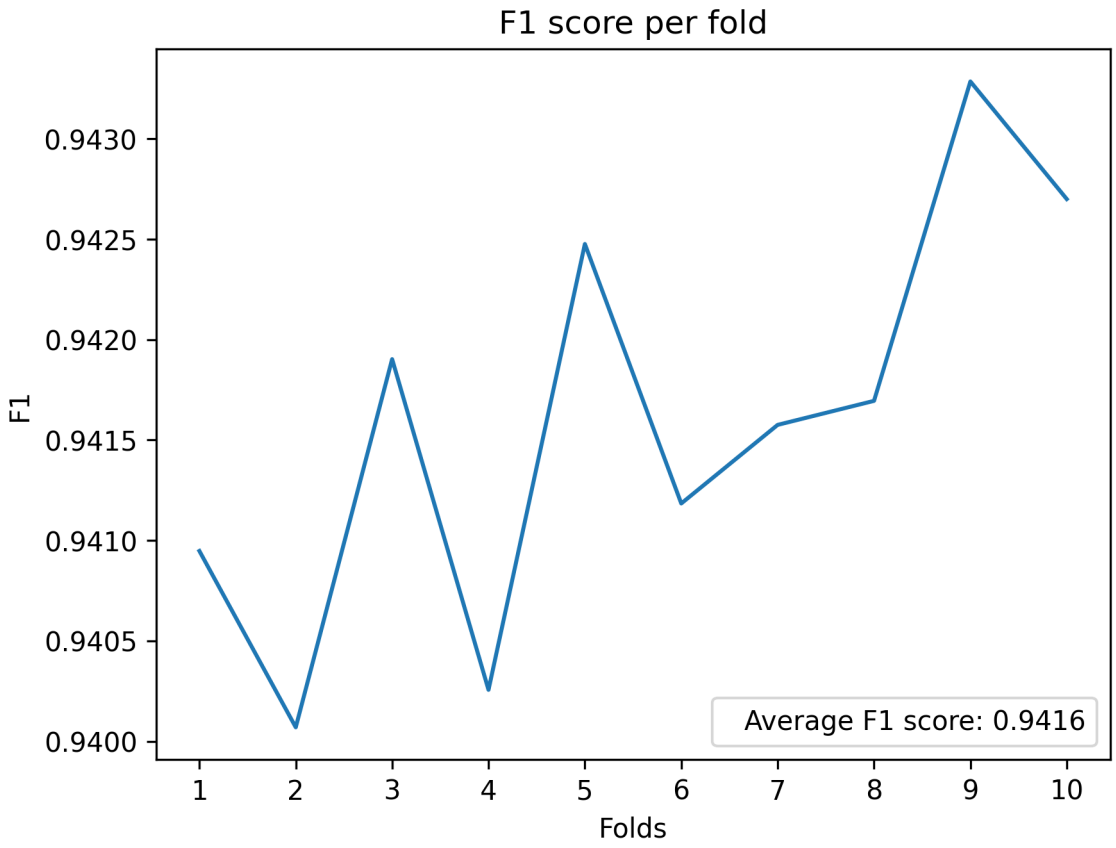

Figure 4.18: F1 Score of Naive Bayes

The provided FIgure 4.19 illustrates the Receiver Operating Characteristic (ROC) curves for ten separate folds of Gaussian Naive Bayes classifiers. The receiver operating characteristic (ROC) curves have a proximity to the upper-left corner, suggesting that the Gaussian Naive Bayes models demonstrate strong performance across all folds of the data. The area under the receiver operating characteristic curve (AUC) for each fold exhibits a high level of performance, with values ranging from 0.9355 to 0.9396

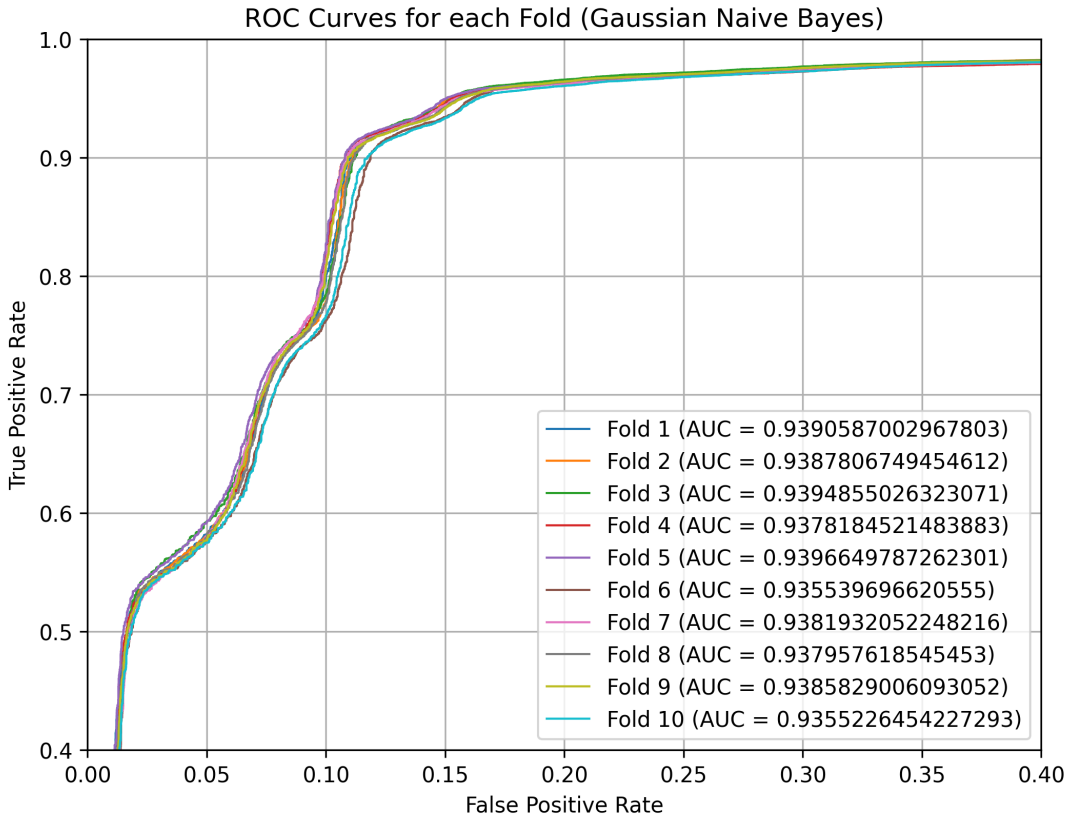

Figure 4.19: ROC Curve of Gaussian Naive Bayes

Here in figure 4.20, we see a graphical representation of the confusion matrix.

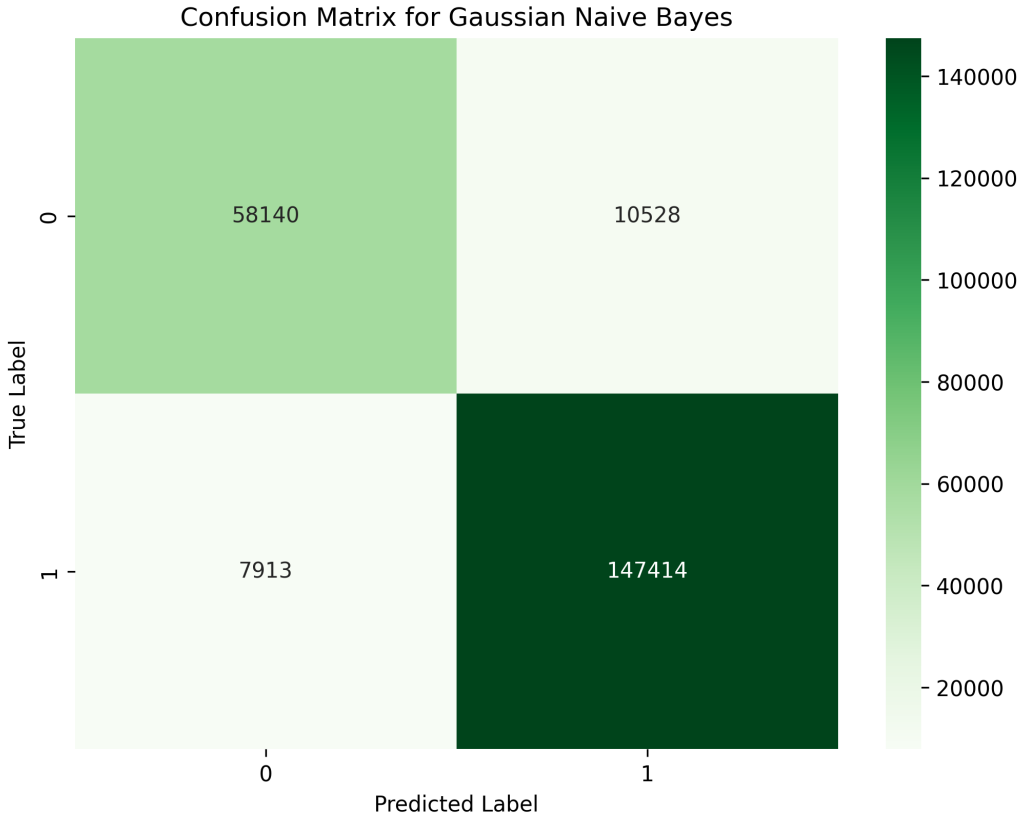

Figure 4.20: Confusion Matrix of Gaussian Naive Bayes

Figure 4.21 presents a number of cancer treatment choices along the X-axis and associated accuracy scores along the Y-axis. The effectiveness of each potential therapy is calculated. In particular, the overall effectiveness rate of Chemotherapy is at a height level approaching 92.2 percent. However, evidence of all available treatment's effectiveness is lower than 91.8 percent.

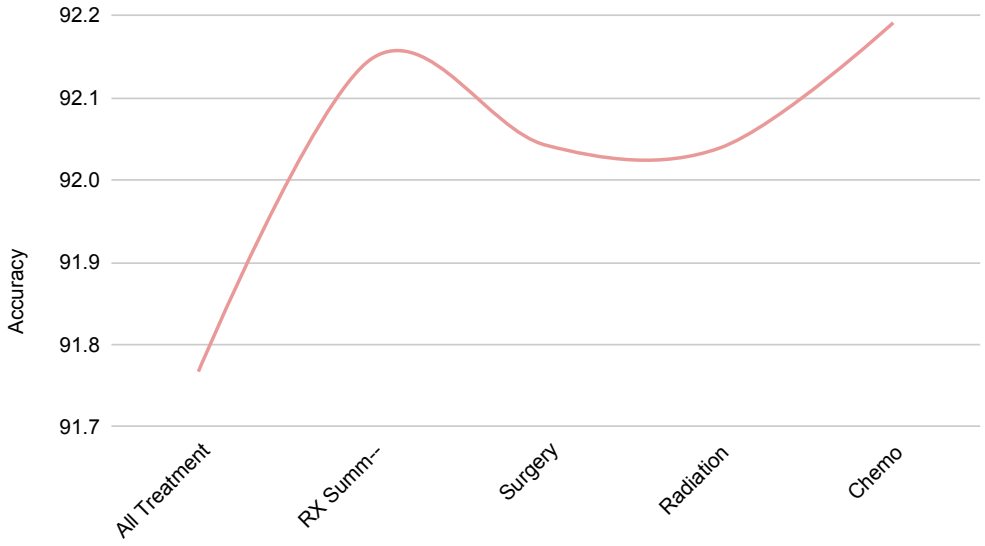

Figure 4.21: Accuracy of Gaussian Naive Bayes in Regards Treatment

4.1.5 Support Vector Machine

The Support Vector Machine (SVM) model is a supervised learning algorithm that is used to classify data points into two or more categories. The SVM model works by finding a hyperplane that best separates the two categories of data points. The hyperplane is a line or a plane that passes through the data points in such a way that the data points on one side of the hyperplane are all of one category and the data points on the other side of the hyperplane are all of the other category.
The SVM model is a non-parametric model. It's also a linear model, which means that the hyperplane that separates the two categories of data points is a straight line. This makes the SVM model easy to understand and interpret. However, the SVM model can also be used to classify data points in non-linear domains by using a kernel function. The kernel function transforms the data points into a higher-dimensional space where the hyperplane can be a non-linear line.
The SVM model is trained on a dataset of known outcomes. The model learns the parameters of the hyperplane that best separates the two categories of data points. The SVM model can then be used to predict the category of a new data point. The model does this by classifying the new data point to the side of the hyperplane that has the most data points in its category.
The SVM model is interpretable to some extent. The parameters of the hyperplane can be interpreted as the importance of the features in the prediction. For example, a feature with a large weight in the hyperplane is more important than a feature with a small weight.
The SVM model is a fast model, which means that it can be trained and used to

make predictions quickly. This makes the SVM model a good choice for applications where speed is important.
The SVM model has been shown to be effective for cancer treatment survival prediction with our dataset. The model was able to achieve an overall accuracy of 93.04%.
The SVM model is a powerful and versatile algorithm for cancer treatment survival prediction. It is interpretable to some extent, and it can be used to predict the category of a new data point with high accuracy. These make it an ideal choice for this application.

The X-axis of the graph 4.22 shows how many cross-validations folds there are, and the Y-axis shows how accurate each fold was. The splits are right about 0.9306 of the time, on average. Fold 6 is more accurate than 0.9310, while Fold 3, which is 0.9290, is the least exact.

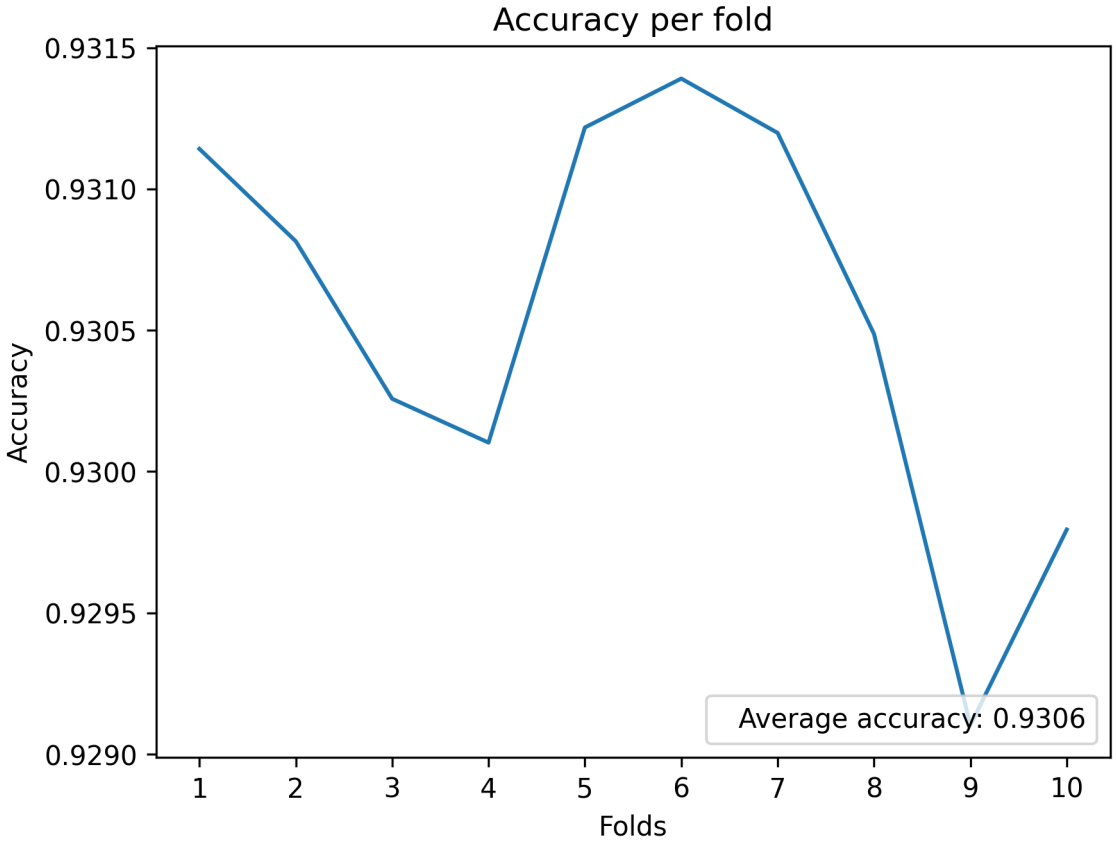

Figure 4.22: Accuracy of SVM

Figure 4.23 has the F1 scores on the Y axis and the number of cross-validation breaks on the X axis. The average F1 score for pleats is 0.9306. Fold 6 has a height number of more than 0.9310, while Fold 9, which is almost at 0.9290, is the least accurate.

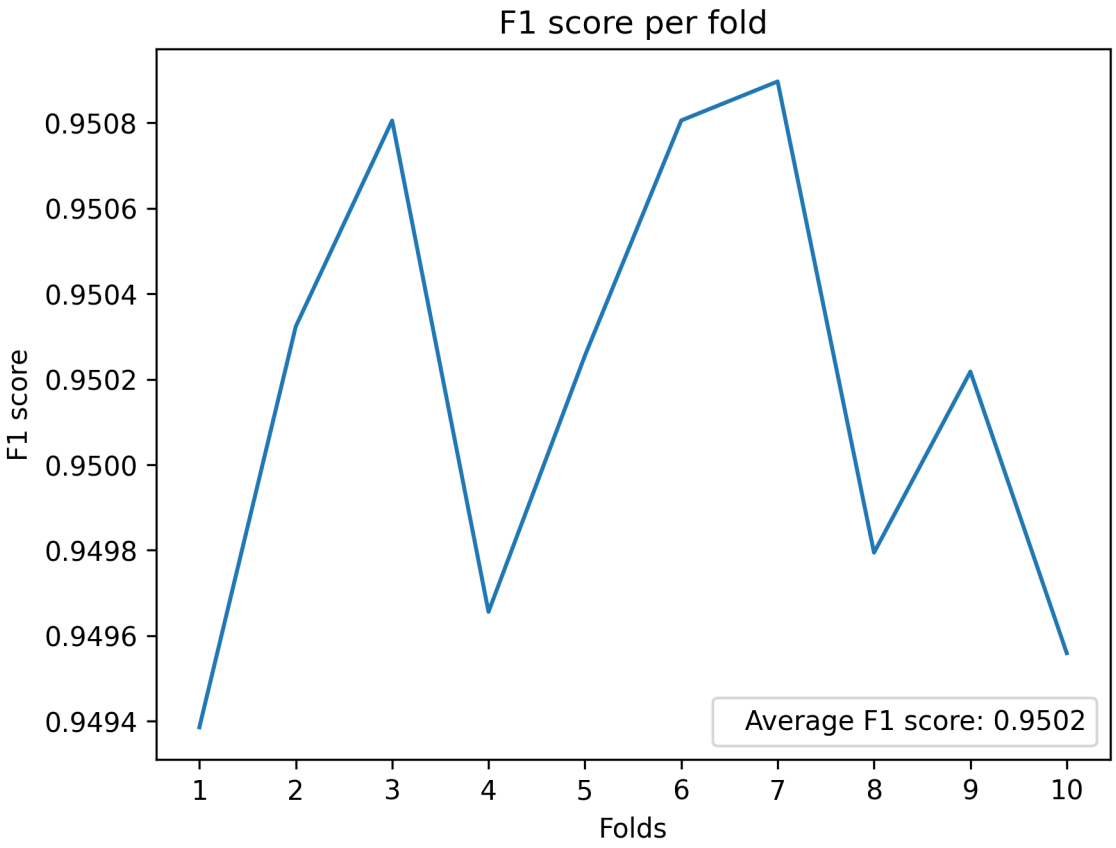

Figure 4.23: F1 Score of SVM

The graph 4.24 depicts several cancer therapies, with the X-axis identifying them and the Y-axis offering accuracy ratings for each type of therapy. The decision tree's effectiveness is assessed for each therapeutic category. Notably, both the surgical procedure and all treatment methods are incredibly precise, both have a score of 93. Both Radiation and Surgery Radiation Sequence has been demonstrated to have a lower degree of accuracy, with an accuracy descending towards 92.

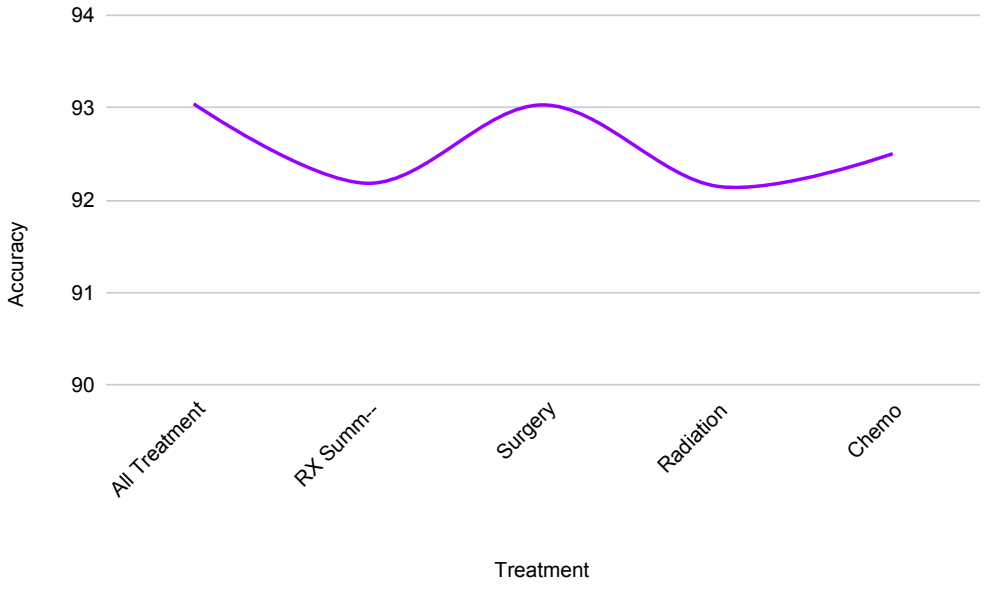

Figure 4.24: Accuracy of SVM in regards Treatment.

4.1.6 Gradient Boosting

The Gradient Boosting model is a machine learning algorithm that is used to predict the probability of an event occurring. In the case of cancer treatment survival prediction, the Gradient Boosting model can be used to predict the probability of a patient surviving their cancer treatment.
The Gradient Boosting model is a boosting algorithm, which means that it builds a series of decision trees that are trained to correct the mistakes of the previous trees. This helps to prevent the model from overfitting the data.
The Gradient Boosting model works by building a series of decision trees. Each decision tree is trained to predict the probability of the event occurring. The predictions from the individual decision trees are then combined to make a final prediction.
The Gradient Boosting model is a non-linear model, which means that the relationship between the features in the data and the probability of the event occurring may not be a straight line. This means that the coefficients in the model may not be directly interpretable in the same way as the coefficients in a Logistic Regression model.
To make a final prediction, the Gradient Boosting model calculates the weighted sum of the predictions from the individual decision trees. The weights of the decision trees are determined by the gradient descent algorithm.
The Gradient Boosting model is also interpretable, but not in the same way as the Logistic Regression model. The coefficients in the Gradient Boosting model can be interpreted as the contribution of each feature to the final prediction. For example, a coefficient that is positive indicates that the feature is associated with an increase in the probability of the event occurring, while a coefficient that is negative indicates that the feature is associated with a decrease in the probability of the event occurring.
The Gradient Boosting model has been shown for cancer treatment survival prediction with our dataset. The model was able to achieve an overall accuracy of 95.76%.
Overall, the Gradient Boosting model is a powerful and versatile algorithm for cancer treatment survival prediction. It is interpretable, and it can be used to predict the probability of the event occurring for each data point. These make it an ideal choice for this application.

The accuracy scores acquired in each fold are shown on the Y-axis in figure 4.25, while the many folds of cross-validation are shown on the X-axis. The accuracy value for all of the folds combined is 0.9577. The greatest accuracy score for Fold 6 is greater than 0.959, while the lowest accuracy number for Fold 3 is less than 0.956.

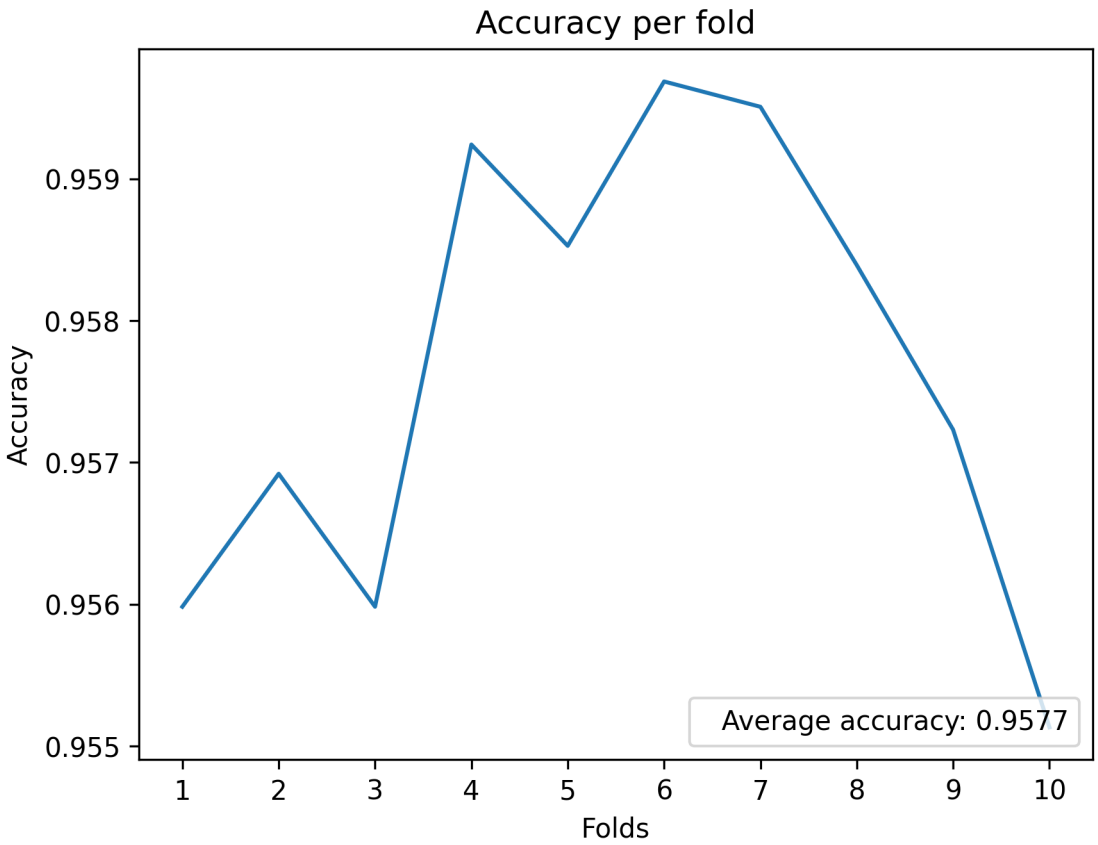

Figure 4.25: Accuracy of Gradient Boosting

The X-axis in figure 4.26 indicates the cross-validation folds employed, and the Y-axis indicates the F1 scores obtained for each fold. The overall average F1 score for folds is 0.9690. Particularly for Fold 6, the accuracy score is close to 0.9705 and it is the height score, while Fold 3 has the lowest accuracy value below 0.9680.

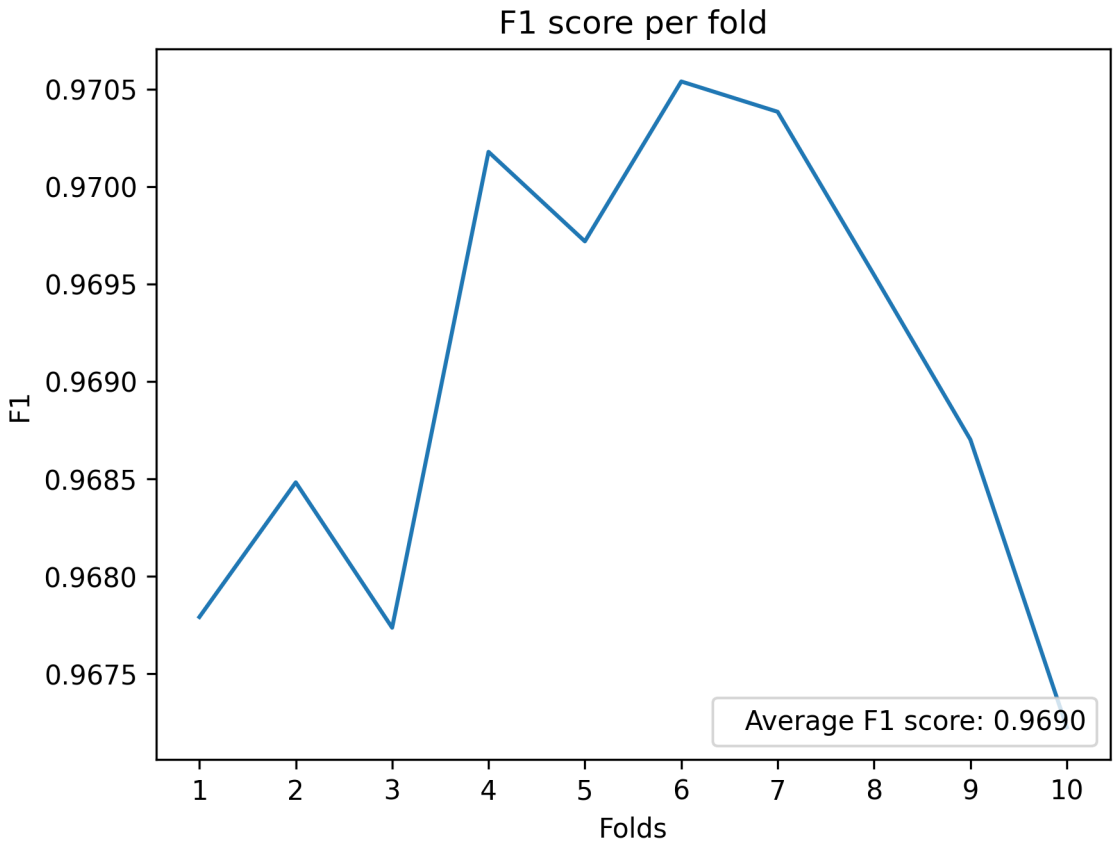

Figure 4.26: F1 Score of Gradient Boosting

Figure 4.27 here are ROC curves representing 10 gradient boost folds. The ROC curves are grouped in the upper left corner, which means the gradient-boosting algorithm is effective across all data subsets. High AUC values for each fold range from 0.9911 to 0.9918. The effectiveness of the gradient-boosting concept gets strengthened further.

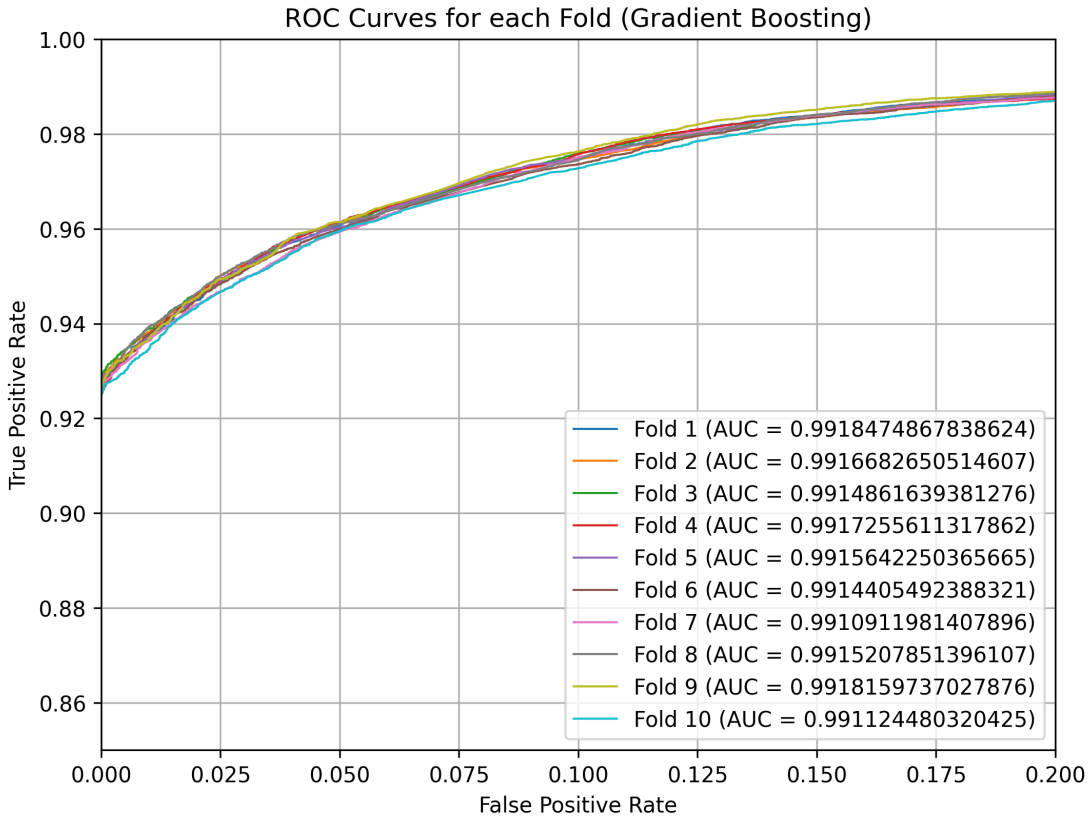

Figure 4.27: ROC Curve of Gradient Boosting

The confusion matrix is depicted in figure 4.28. Using this example, we can assess the efficacy of our categorization approach.

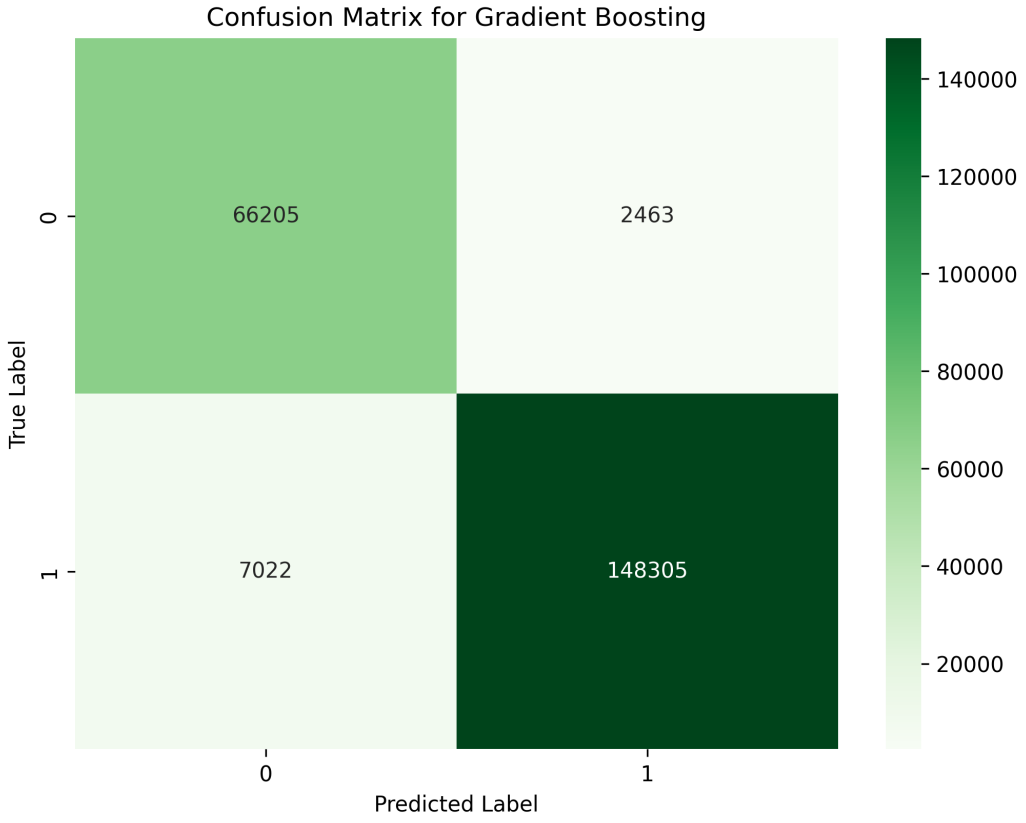

Figure 4.28: Confusion Matrix of Gradient Boosting

The graph 4.29 depicts a graphical depiction in which the X-axis lists various cancer therapies and the Y-axis provides accuracy ratings for each type of therapy. The effectiveness of the decision tree is evaluated for each therapy category. Notably, All treatment process has an extremely high level of precision, more than 95.76. Chemotherapy, on the other hand, may be shown to have a lower degree of accuracy that is below 95.74.

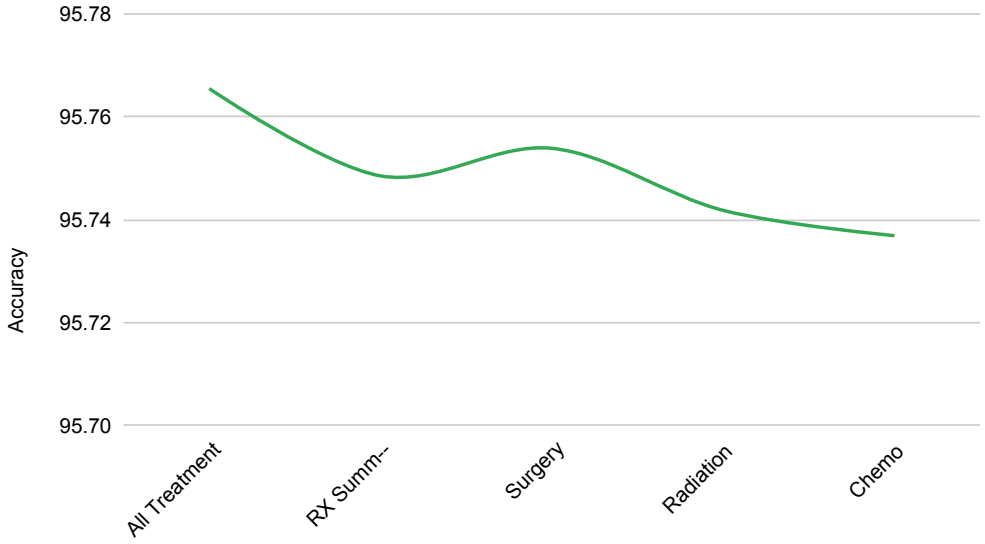

Figure 4.29: Accuracy of Gradient Boosting in Regards Treatment

4.1.7 AdaBoost

AdaBoost stands for Adaptive Boosting. It is a boosting algorithm, which means that it builds a series of weak learners that are combined to create a strong learner. The AdaBoost model works by training a series of decision trees. Each decision tree is trained to predict the probability of the event occurring. The predictions from the individual decision trees are then weighted. The weights of the decision trees are determined by the AdaBoost algorithm.

The AdaBoost algorithm works by iteratively training decision trees and then weighting them according to their accuracy. The algorithm starts by assigning equal weights to all of the decision trees. Then, it trains a new decision tree and weights it according to its accuracy. The algorithm then repeats this process, training new decision trees and weighting them according to their accuracy. The final prediction is made by combining the predictions from the individual decision trees.

The AdaBoost model has been shown for cancer treatment survival prediction with our dataset. The model was able to achieve an overall accuracy of 95.65%.

Overall, the AdaBoost model is a powerful and versatile algorithm for cancer treatment survival prediction. It is interpretable, and it can be used to predict the probability of the event occurring for each data point. These make it an ideal choice for this application.

Figure 4.30 shows the accuracy scores obtained in each fold on the Y-axis and the

numerous folds of cross-validation on the X-axis. The total accuracy for all folds combined is 0.9562. Fold 6 has the highest accuracy score of 0.9705, while Fold 3 has the lowest accuracy score approaching 0.954.

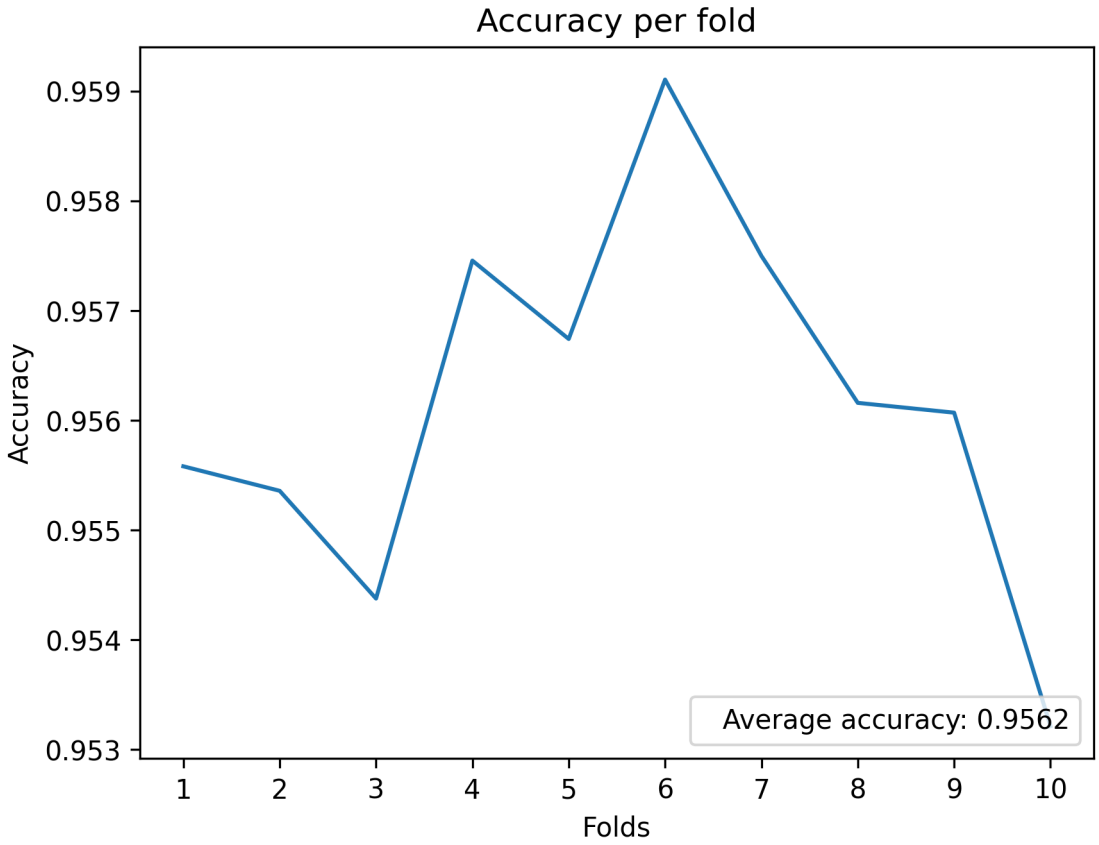

Figure 4.30: Accuracy of AdaBoost

In Figure 4.31, the X-axis represents the cross-validation folds used, and the Y-axis represents the F1 scores achieved for each fold. Folds have an average F1 score of 0.9680. Fold 6 has an accuracy score close to 0.970 and is the height score, whilst Fold 3 has the least accuracy score below 0.967.

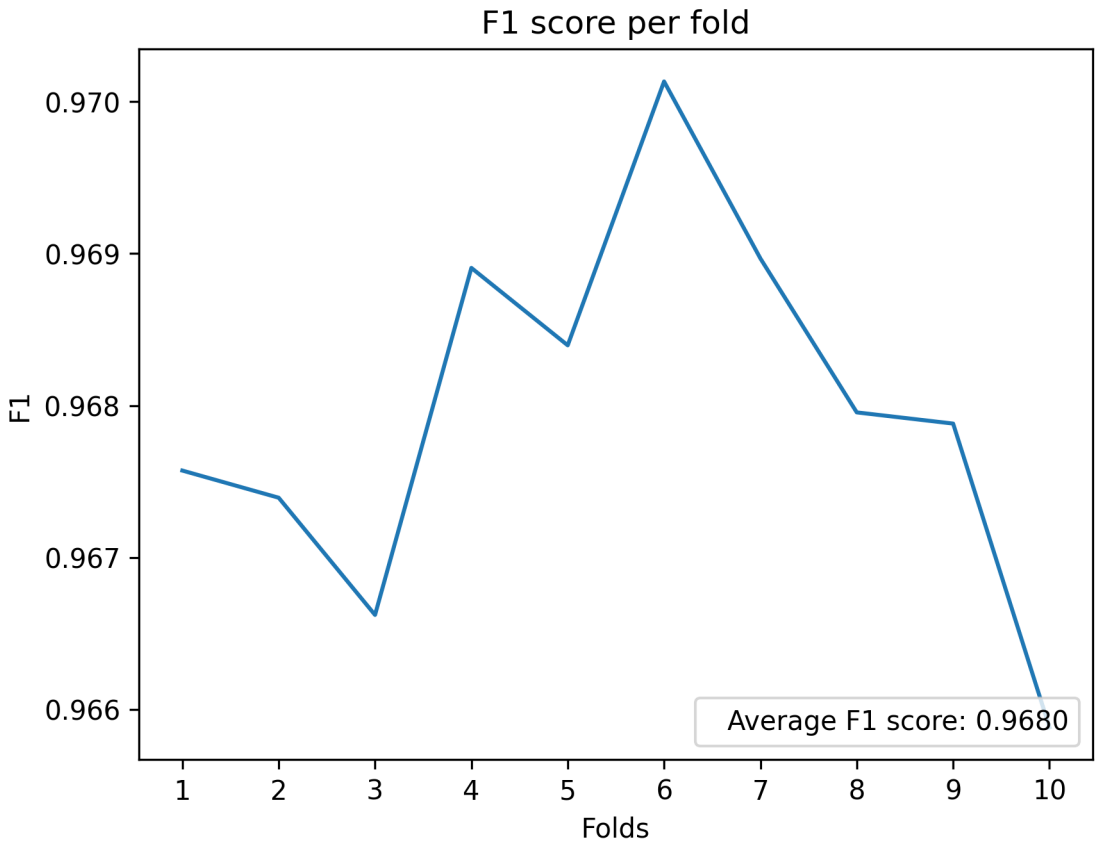

Figure 4.31: F1 Score of AdaBoost

ROC curves for 10 different AdaBoost iterations are plotted below in Figure 4.32. The ROC curves are clustered in the upper left corner, which means the AdaBoost model is performing well across all data sets. Each fold has a high AUC, with values between 0.9918 and 0.9923. This demonstrates once again how well-performing the AdaBoost model is.

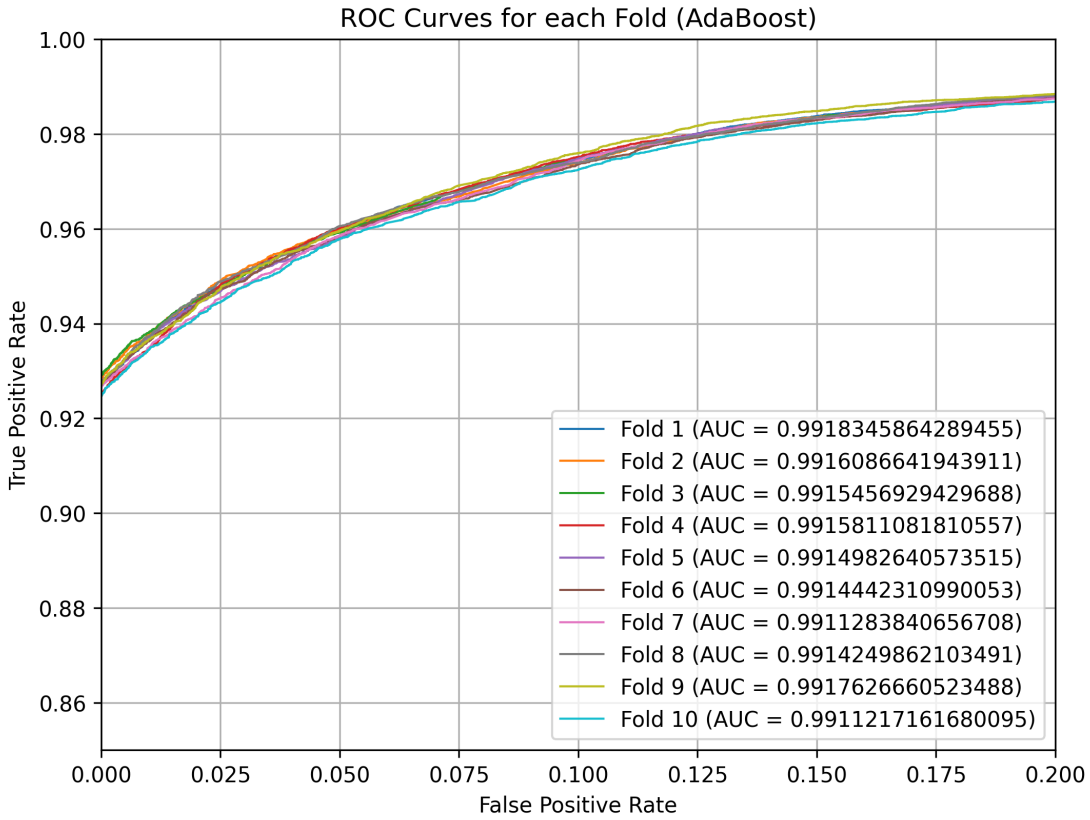

Figure 4.32: ROC Curve of AdaBoost

Figure 4.33 depicts the confusion matrix. Using this example, we can evaluate the effectiveness of our categorization method.

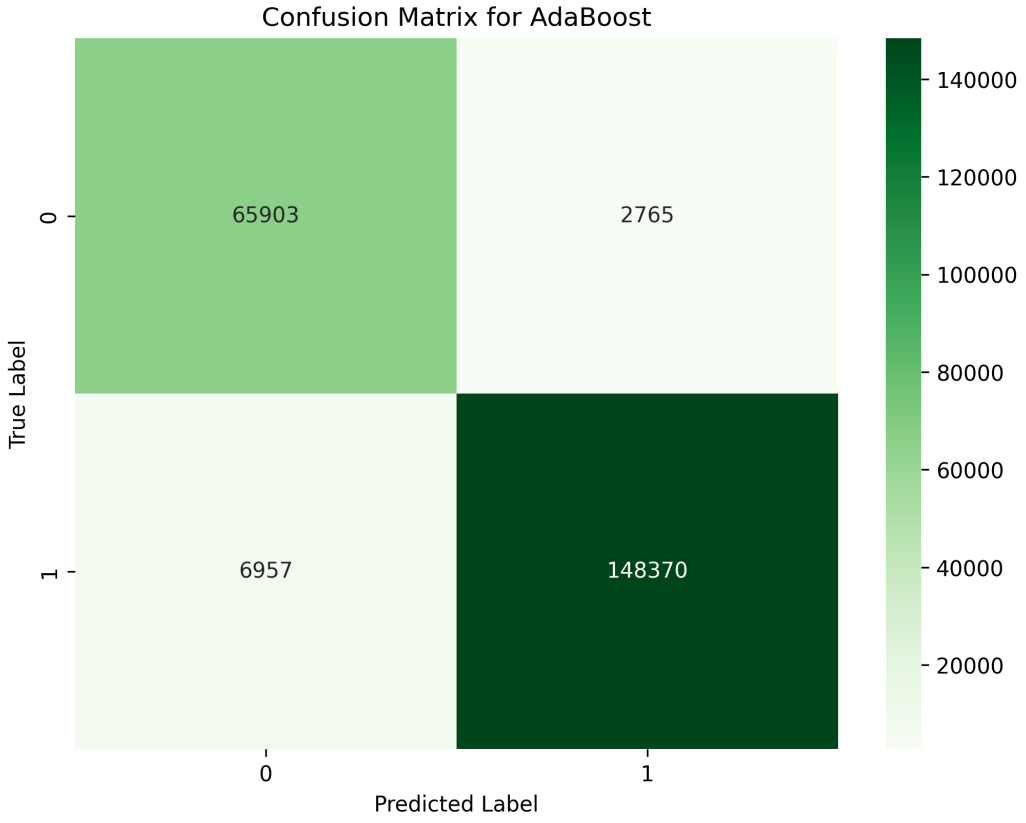

Figure 4.33: Confusion Matrix of AdaBoost

Figure 4.34 depicts several cancer therapies, with the X-axis identifying them and the Y-axis offering accuracy ratings for each type of therapy. The decision tree's effectiveness is assessed for each therapeutic category. Notably, both the surgical procedure and the Surgery Radiation Sequence are incredibly precise, but the Radiation technique has a height score of 95.70. All therapies, on the other hand, have been demonstrated to have a lower degree of accuracy, with a score of 95.65.

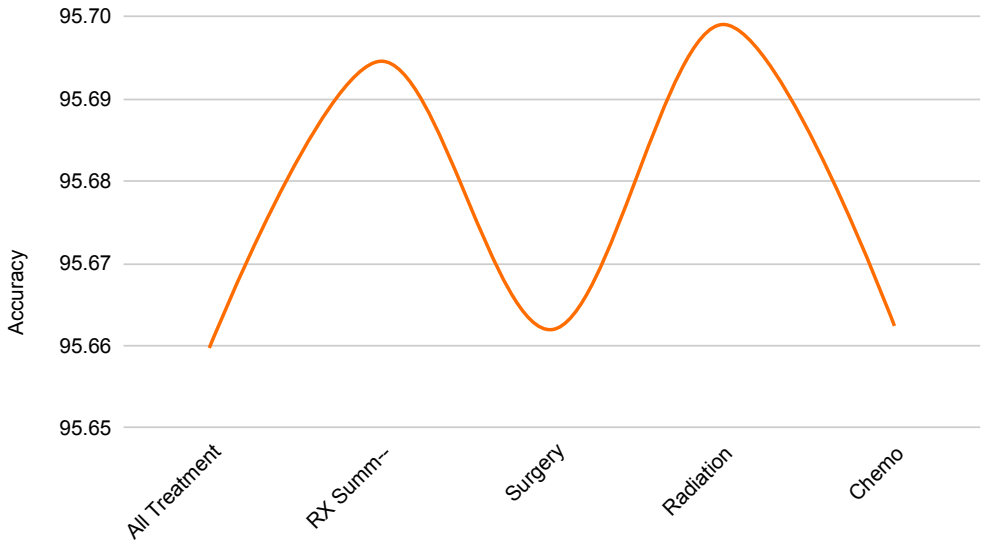

Figure 4.34: Accuracy of AdaBoost in regards Treatment

4.1.8 Histogram-based Gradient Boosting

Histogram-based Gradient Boosting (HGB) is a gradient boosting algorithm, which means that it builds a series of decision trees that are trained to correct the mistakes of the previous trees. This helps to prevent the model from overfitting the data.
The HGB model works by dividing the features in the data into histogram bins. Once the features have been divided into histogram bins, the HGB algorithm then builds a series of decision trees. Each decision tree in the model is trained to predict the probability of the event occurring for a specific histogram bin. The predictions from the individual decision trees are then combined to make a final prediction. The final prediction is made by taking the weighted average of the predictions from the individual decision trees. The weights of the decision trees are determined by the HGB algorithm.
The HGB model is interpretable. The coefficients in the HGB model can be interpreted as the contribution of each feature to the final prediction.
The HGB model has been shown to be very accurate for cancer treatment survival prediction with our dataset. The model was able to achieve an overall accuracy of 95.86% which is also the highest among the models. This means that the model was able to correctly predict more accurately whether a patient would survive their cancer treatment with 95.86% accuracy.
Overall, the HGB model is a powerful and versatile algorithm for cancer treatment survival prediction. It is interpretable, and it can be used to predict the probability

of the event occurring for each data point. These make it an ideal choice for this application.

On the Y-axis of figure 4.35 are the accuracy scores obtained in each fold, while the X-axis displays the many folds of cross-validation. The combined accuracy score for each of the folds is 0.9585. Fold 6 has the highest accuracy score, which is greater than 0.961, while Fold 3 has the lowest accuracy score, which is 0.957.

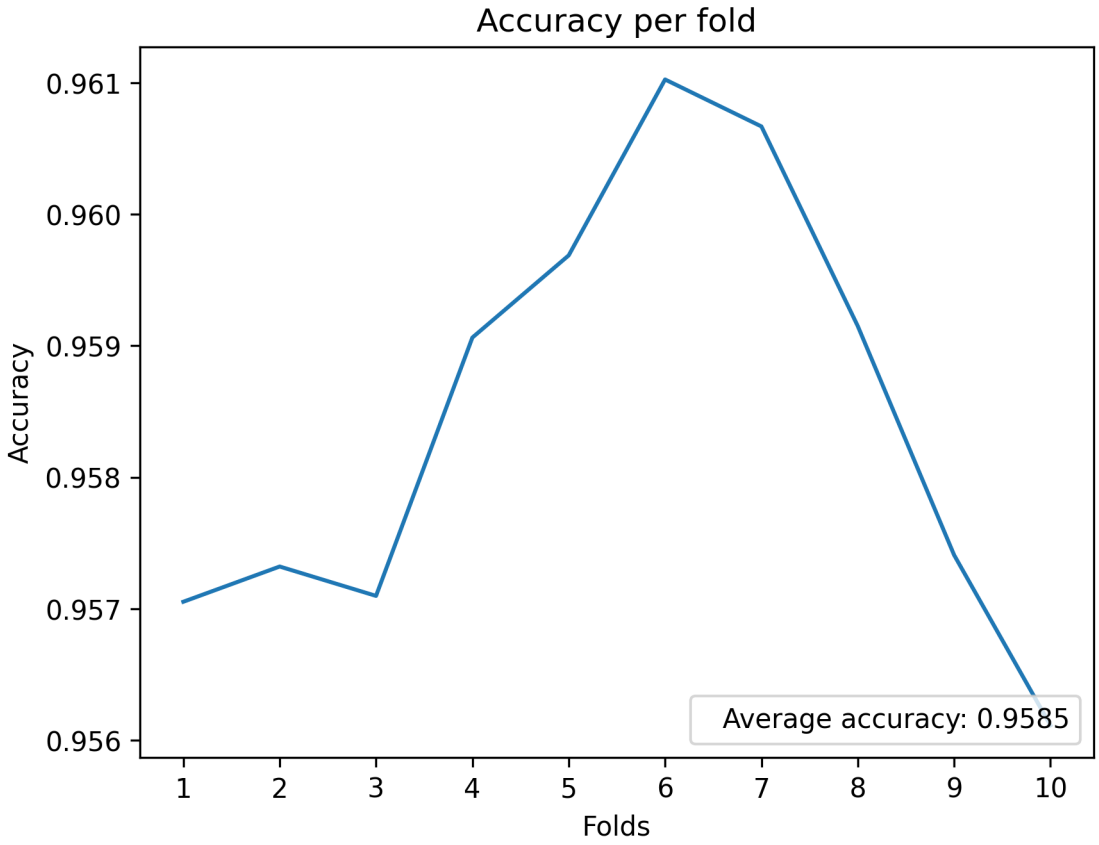

Figure 4.35: Accuracy of HistGradientBoosting

Figure 4.36 depicts the cross-validation folds employed along the X-axis, and the F1 scores derived for each fold along the Y-axis. The average F1 score for pleats is 0.9697 overall. Fold 6 has an accuracy score over 0.9710, which is the height score, whereas Fold 10 has the lowest accuracy score below 0.9680.

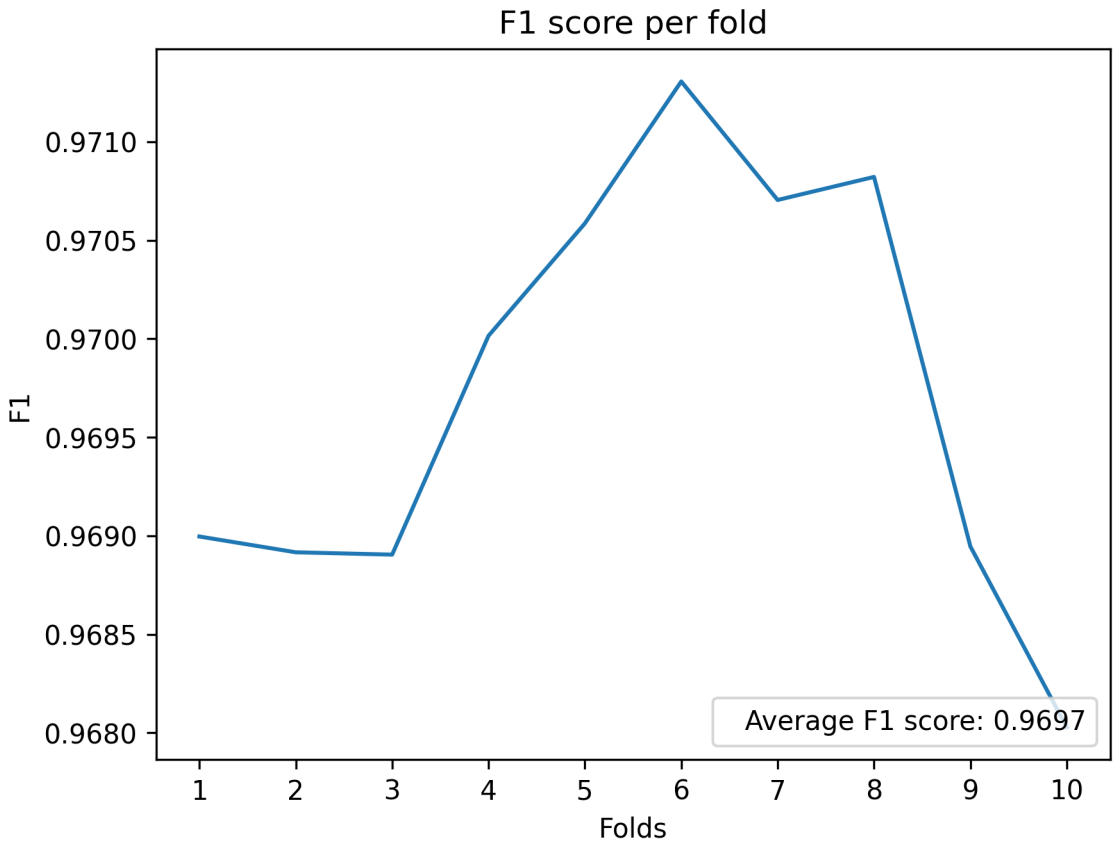

Figure 4.36: F1 Score of HistGradientBoosting

Figure 4.37 shows the ROC curves for HistGradientBoosting with 10 folds. The ROC curves are close to the top left corner, which means that the HistGradientBoosting model works well on all folds of data. The AUC is also high for each fold, running from 0.9923 to 0.9970. This is more proof that the HistGradientBoosting model works well.

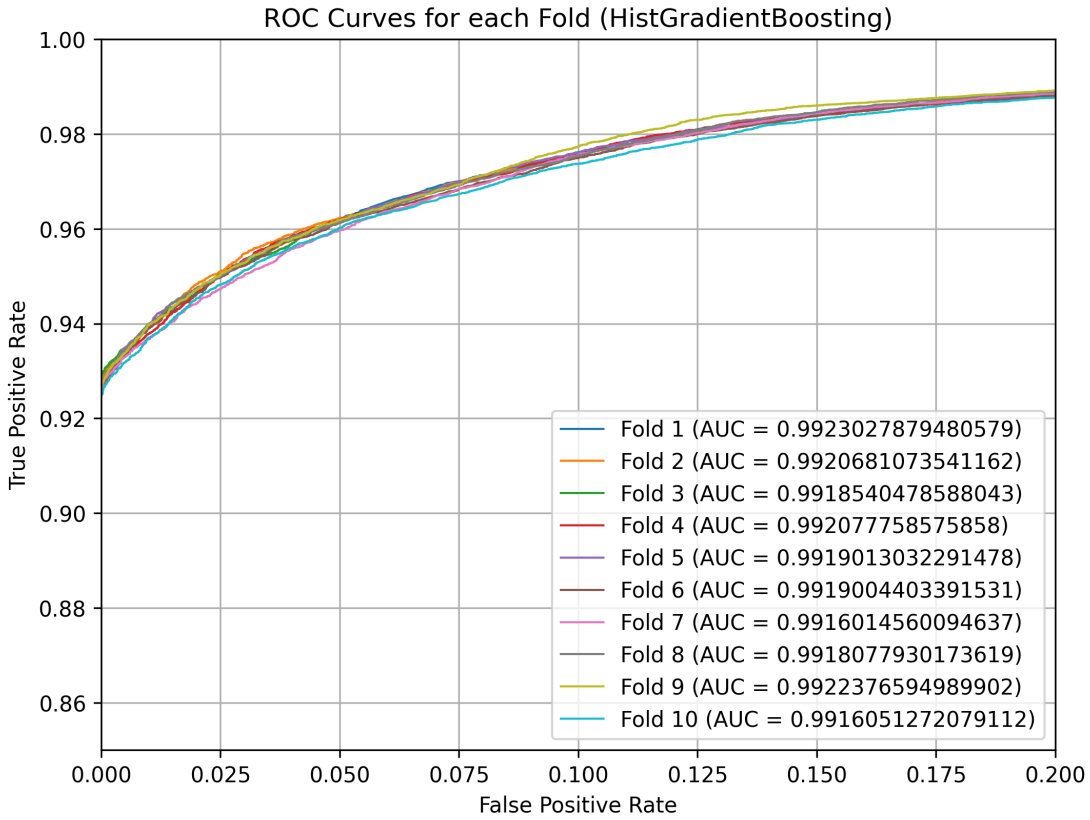

Figure 4.37: ROC Curve of HistGradientBoosting

This graph 4.38 represents the perplexity matrix. Using this instance, we can evaluate the effectiveness of our categorization strategy.

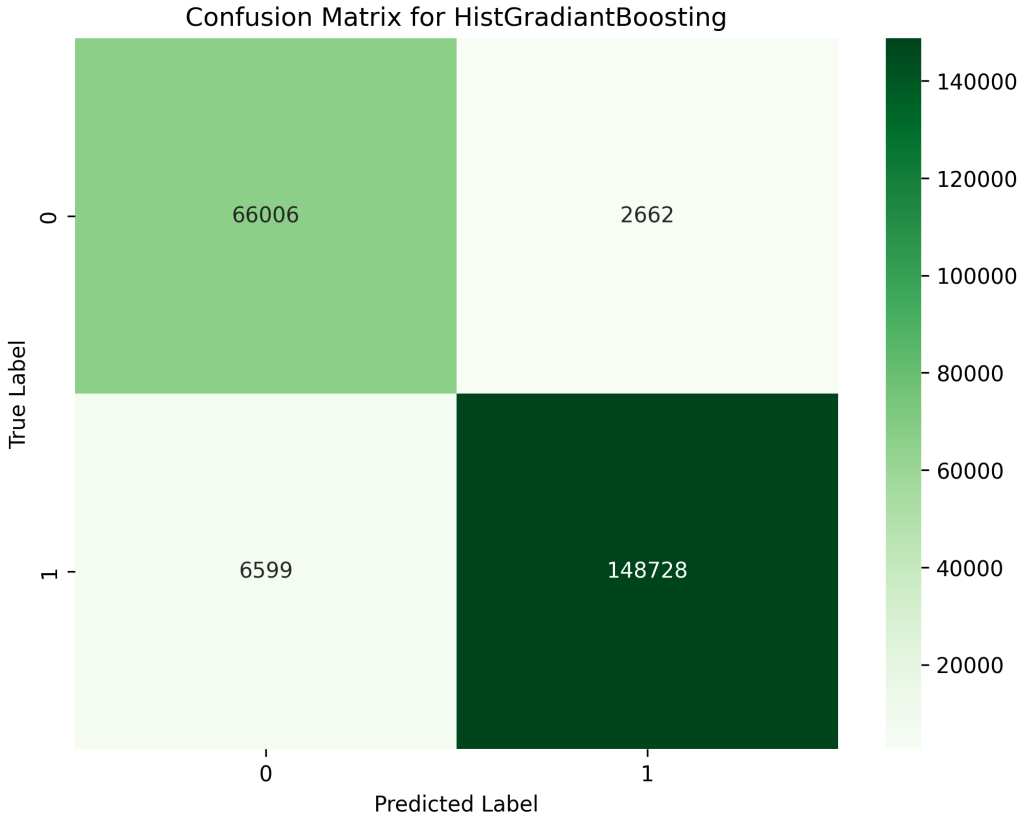

Figure 4.38: Confusion Matrix of HistGradientBoosting

The X-axis of the graph 4.39 lists various cancer therapies, while the Y-axis offers accuracy scores for each form of treatment. Each therapy category's efficacy of the decision tree is determined. Particularly, the precision of each and every treatment procedure exceeds 95.86 percent. Surgery Radiation Sequence, on the other hand, may be shown to have an accuracy below 95.82 percent.

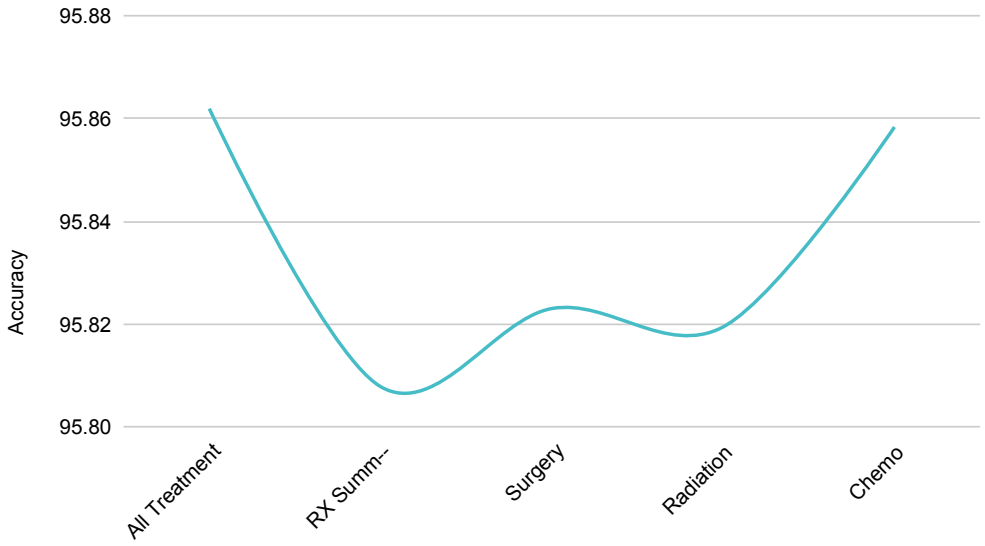

Figure 4.39: Accuracy of HistGradientBoosting in regards Treatment

4.1.9 Multi-layer Perceptron

The Multilayer Perceptron (MLP) model is a supervised learning algorithm. It is also a type of feedforward neural network algorithm. Neural networks are inspired by the way the human brain works. They are composed of interconnected nodes, which are organized into layers. This means that the data flows in one direction through the network, from the input layer to the output layer. The hidden layers are used to process the data and to learn the relationships between the features.

The number of hidden layers in an MLP model can vary. Some models have only one hidden layer, while others have multiple hidden layers. The number of hidden layers in a model is determined by the complexity of the problem that the model is trying to solve.

The MLP model is trained using a backpropagation algorithm. This algorithm works by adjusting the weights of the connections between the nodes in the network so that the model makes the correct predictions as often as possible. The backpropagation algorithm starts by calculating the error between the predicted value and the actual value. The error is then propagated back through the network, and the weights of the connections are adjusted accordingly.

The MLP model is interpretable to some extent. The weights of the connections between the nodes can be interpreted as the importance of each feature in the prediction. For example, a weight that is very close to 1 indicates that the feature is very important, while a weight that is very close to 0 indicates that the feature is not important.

The MLP model has been shown for cancer treatment survival prediction with our dataset. The model was able to achieve an overall accuracy of 95.74% which is very similar to the boosting models that we used.
Overall, the MLP model is a powerful and versatile algorithm for cancer treatment survival prediction. It is interpretable to some extent, and it can be used to predict the probability of the event occurring for each data point. These make it an ideal choice for this application.

Figure 4.40 demonstrates the numerous folds of cross-validation on the X-axis, while the accuracy scores achieved in each fold are shown on the Y-axis. Overall, the folds have an accuracy score of 0.9562. The accuracy score is nearly 0.959 between Fold 6 and Fold 7, whereas the accuracy value for Fold 10 and Fold 1 is below 0.953.

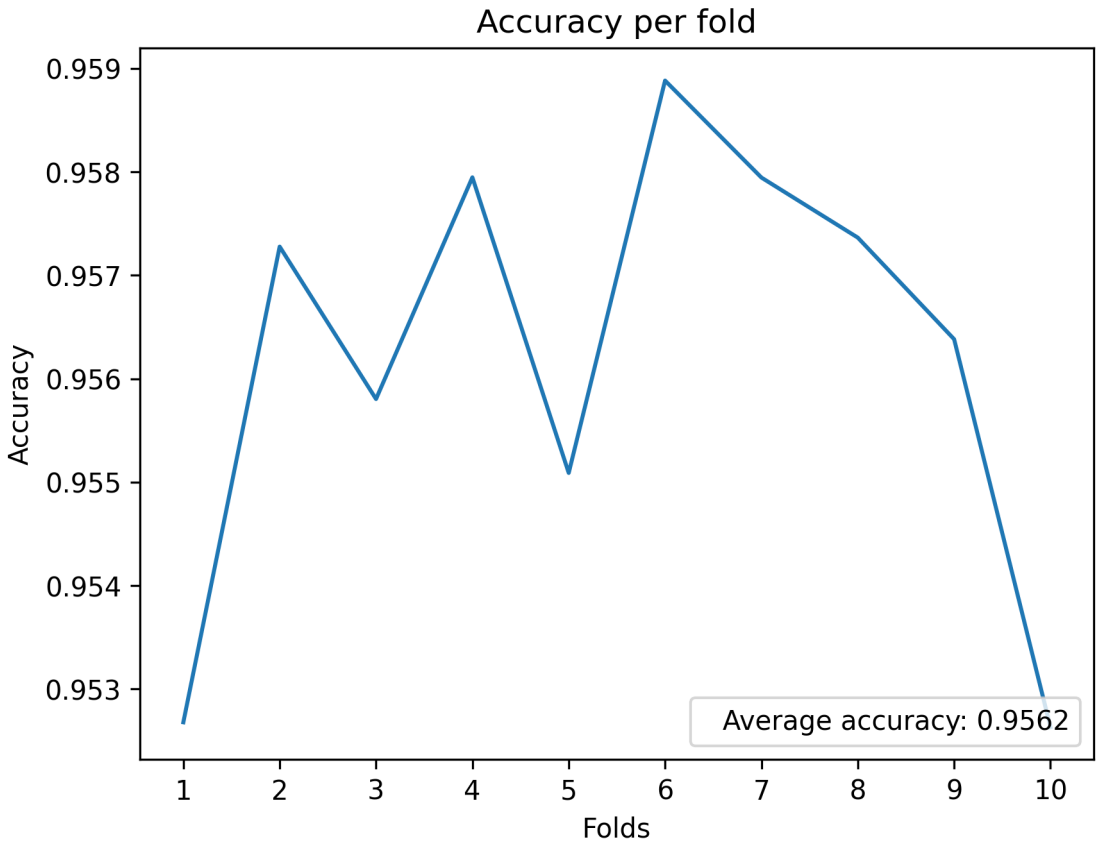

Figure 4.40: Accuracy of Multi-layer Perceptron

The X-axis of figure 4.41 shows the number of cross-validation folds used, while the Y-axis shows the corresponding F1 scores. Overall, pleats have an F1 score of 0.9684. The height score for Fold 6 is over 0.970, while the lowest value is nearly 0.966 for Fold 1.

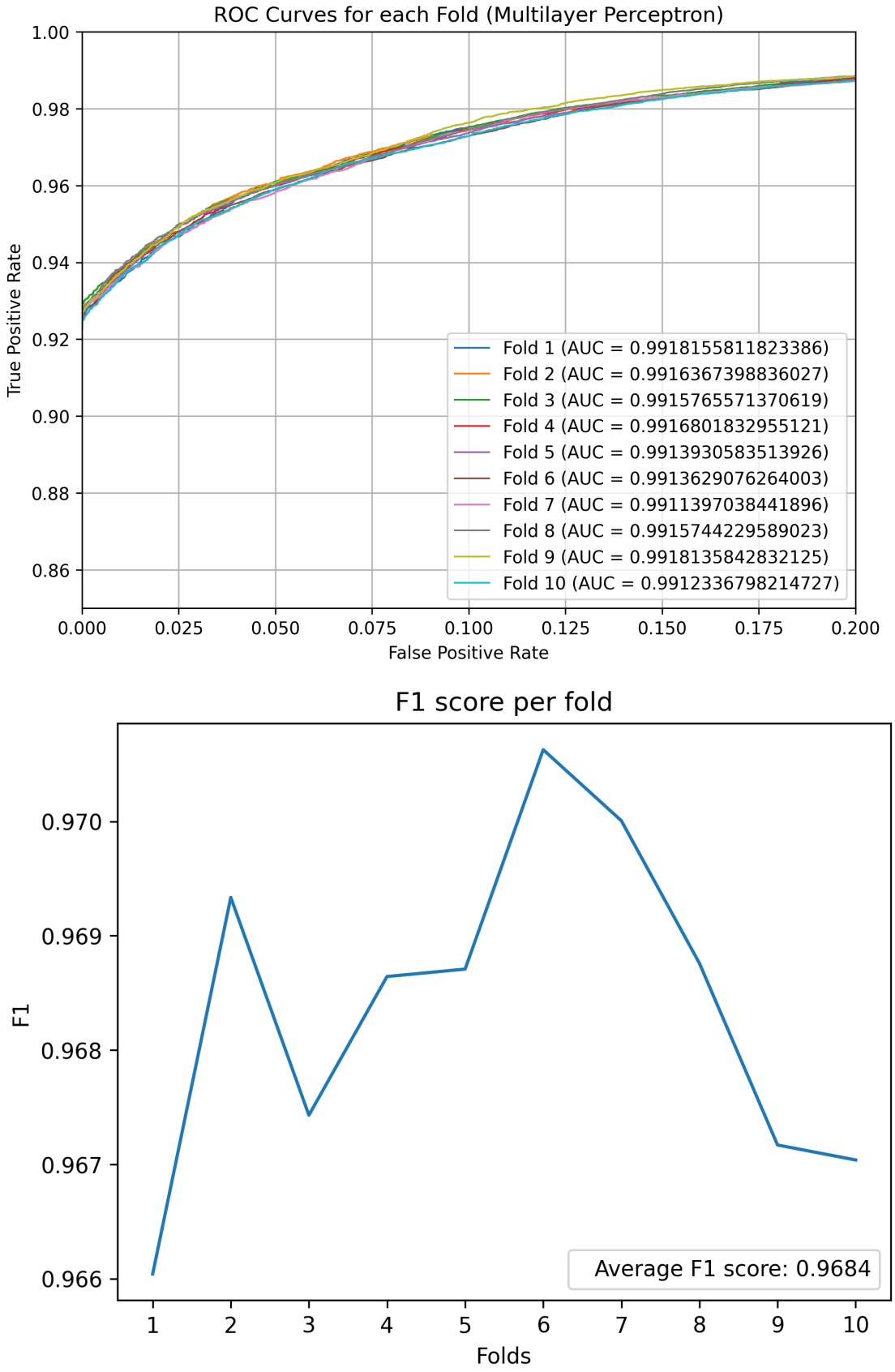

Figure 4.41: F1 Score of Multi-layer Perceptron

The multilayer perceptron (MLP) classifier ROC curves for ten folds are displayed in Figure 4.42. The MLP models are doing well on all folds of data since the ROC

curves are near to the top-left corner. Each fold's AUC, which ranges from 0.9918 to 0.9970, is similarly quite high.

Figure 4.42: ROC Curve of Multi-layer Perceptron

The confusion matrix is depicted here as a graph in Figure 4.43. We can test how well our classification scheme works by using this example.

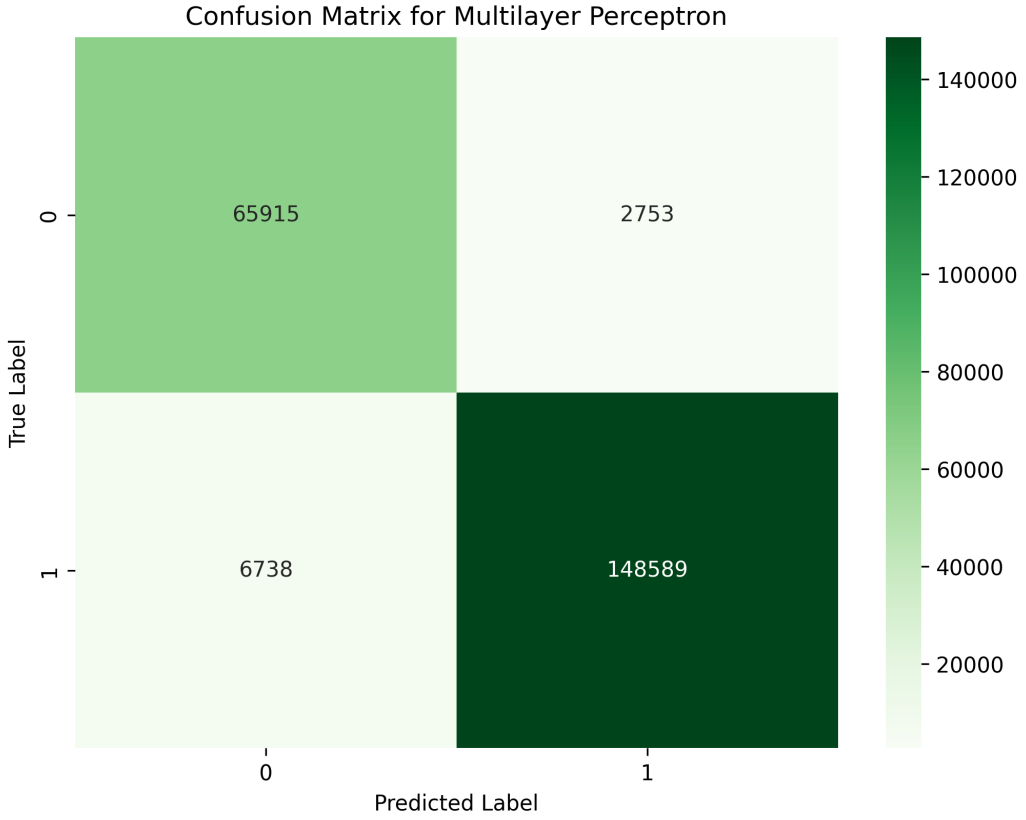

Figure 4.43: Confusion Matrix of Multi-layer Perceptron

On the X-axis of the graph 4.44, you'll find a list of cancer therapy options, and on the Y-axis, you'll find accuracy scores for those options. The decision tree evaluates the efficacy of each treatment option. In particular, the success rate of every single treatment method is more than 95.7 percent. However, it is possible that Radiation efficacy will be demonstrated to be lower than 95.5 percent.

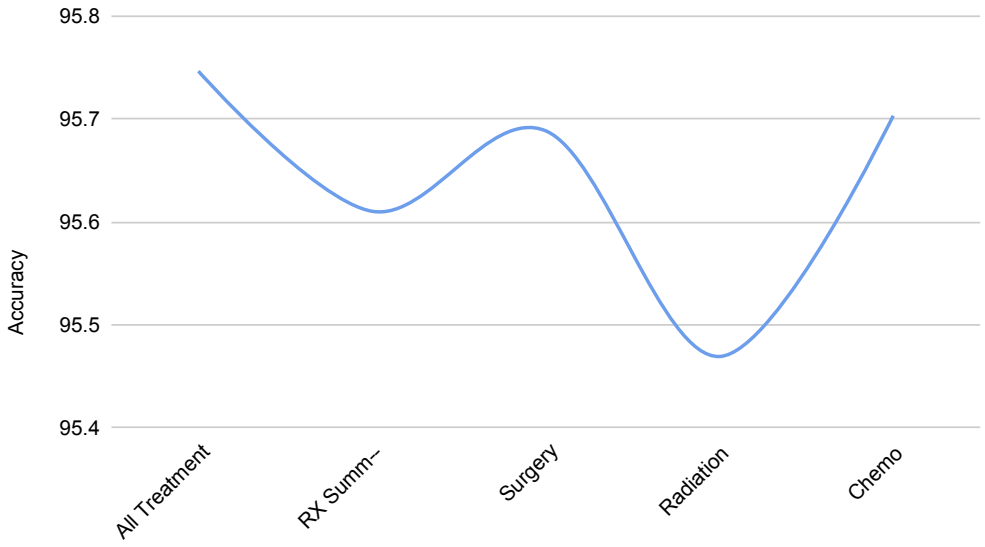

Figure 4.44: Accuracy of Multi-layer Perceptron in regards Treatment

4.2 Comparison

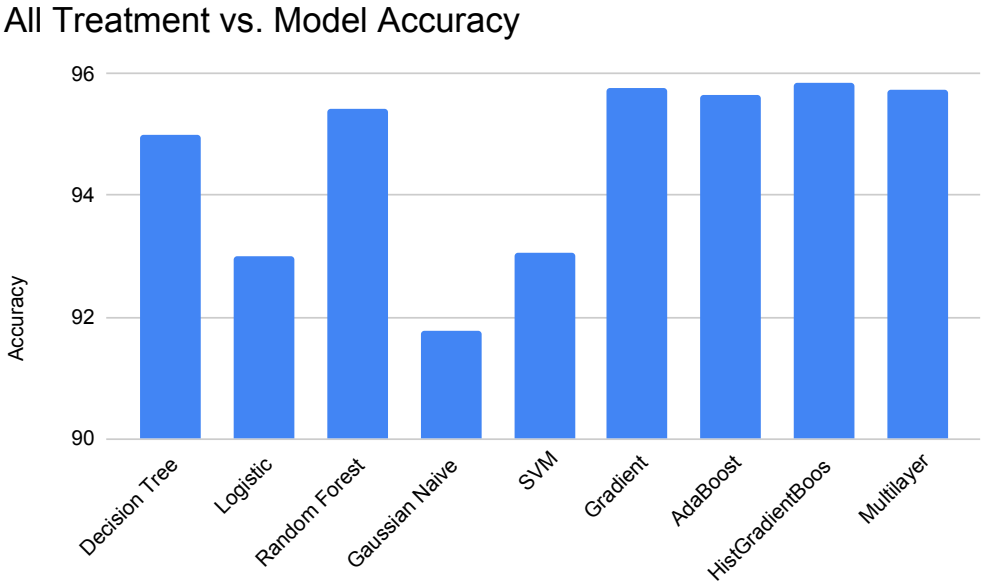

Figure 4.45: Accuracy of different models in regards All Treatment

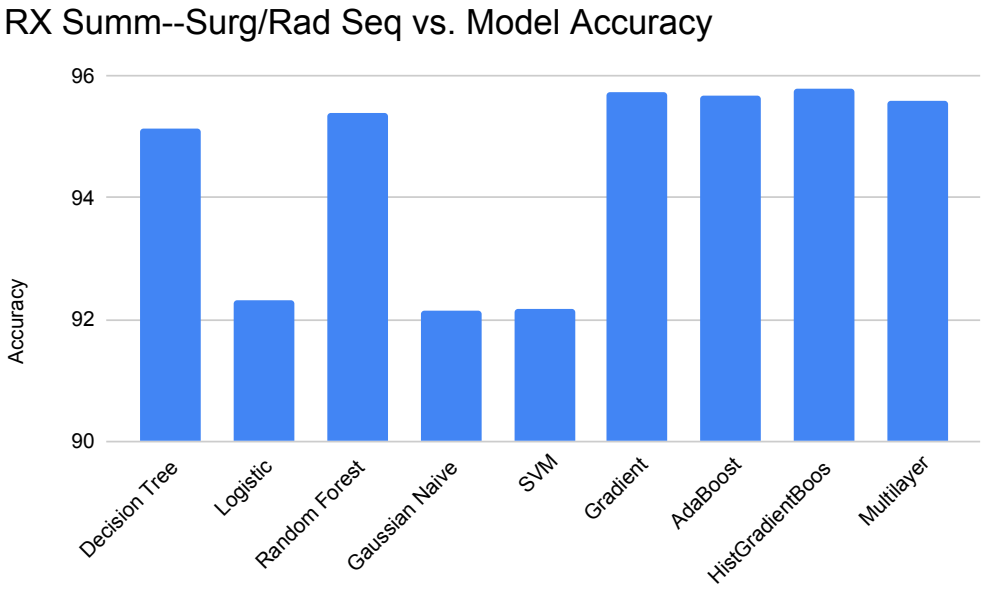

Figure 4.46: Accuracy of different models in regards RX Summ–Surg/Rad Seq

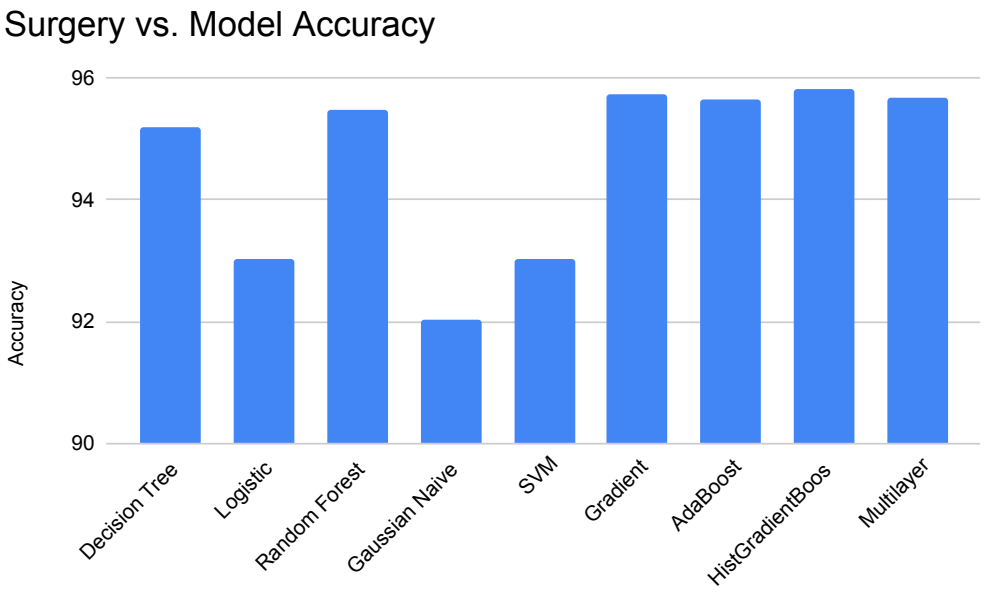

Figure 4.47: Accuracy of different models in regards Surgery

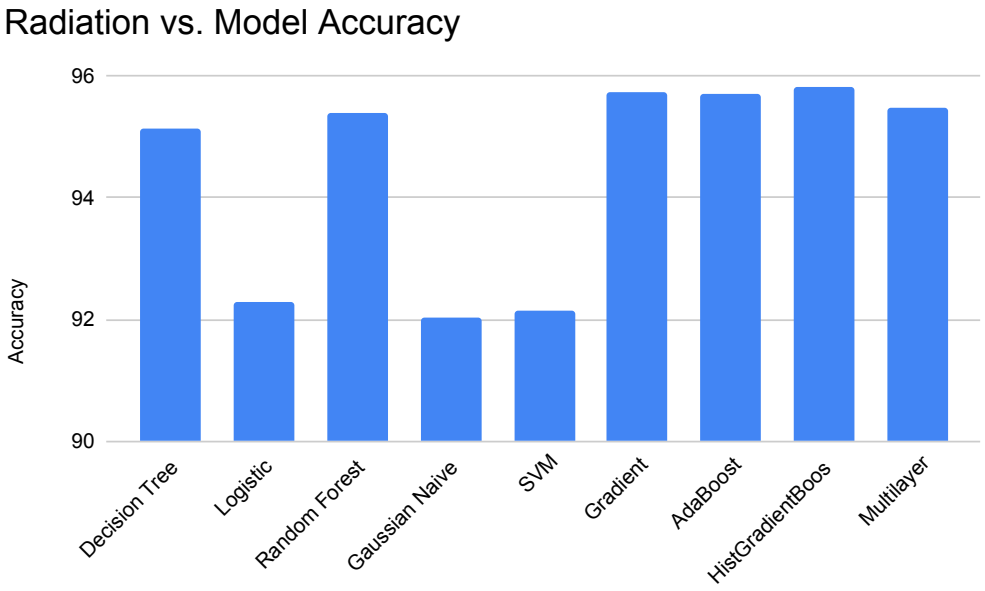

Figure 4.48: Accuracy of different models in regards Radiation

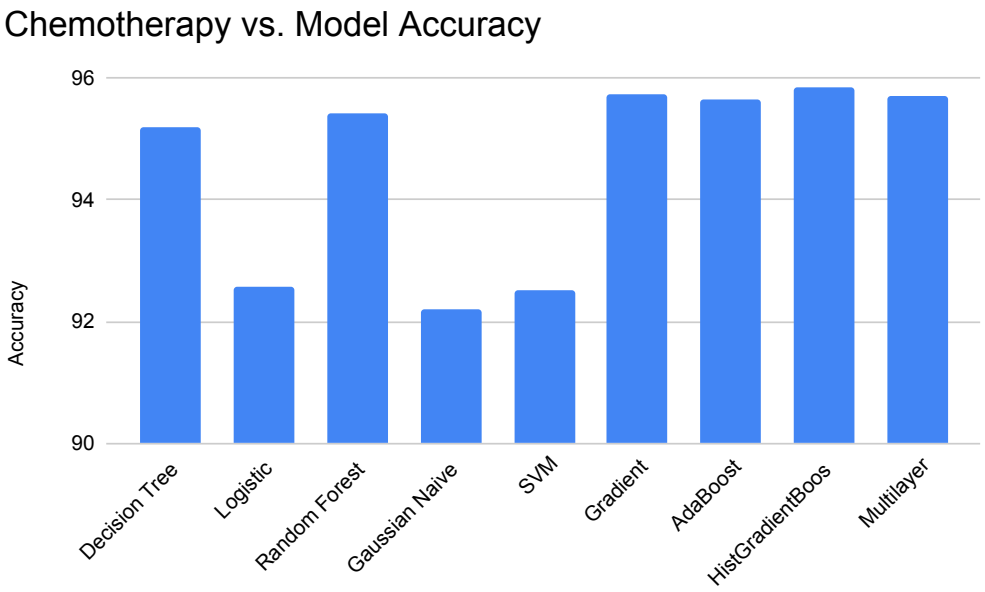

Figure 4.49: Accuracy of different models in regards Chemotherapy

Among the ML models scrutinized, HistGradientBoosting emerged as the top-performing model, achieving remarkable accuracy scores across all treatment categories. With an impressive accuracy of 95.86% for "All Treatment," 95.81% for "RX Summ–Surg/Rad Seq" 95.82% for "Surgery" 95.82% for "Radiation" and 95.86% for "Chemo" HistGradientBoosting showcased its exceptional predictive power.

Following closely behind, the Gradient Boosting and AdaBoost models shared the second and third position in accuracy scores. Both models achieved competitive accuracy, with Gradient Boosting obtaining accuracies of 95.77% for "All Treatment" 95.74% for "RX Summ–Surg/Rad Seq" 95.75% for "Surgery" 95.82%

for "Radiation" and 95.66% for "Chemo" while AdaBoost attained accuracies of 95.66% for "All Treatment" 95.69% for "RX Summ–Surg/Rad Seq" 95.68% for "Surgery" 95.47% for "Radiation" and 95.66% for "Chemo" Both models demonstrated consistent performance and are well-suited for predicting various treatment outcomes.

The middle position in terms of accuracy was secured by the Random Forest model. Random Forest demonstrated remarkable accuracy, achieving accuracy scores of 95.43% for "All Treatment" 95.40% for "RX Summ–Surg/Rad Seq" 95.48% for "Surgery" 95.41% for "Radiation" and 95.43% for "Chemo" The model's consistent performance across multiple treatment categories underscores its potential as a robust and reliable model for accurate cancer treatment prediction.

While the Decision Tree Classifier, Gaussian Naive Bayes, and SVM models achieved respectable accuracy scores, they obtained lower accuracy compared to the top-performing models. Decision Tree Classifier obtained accuracies of 94.98% for "All Treatment," 95.15% for "RX Summ–Surg/Rad Seq" 95.21% for "Surgery" 95.15% for "Radiation" and 95.19% for "Chemo" Gaussian Naive Bayes achieved accuracies of 91.77% for "All Treatment" 92.15% for "RX Summ–Surg/Rad Seq" 92.03% for "Surgery" 92.15% for "Radiation" and 92.19% for "Chemo" SVM obtained accuracies of 93.04% for "All Treatment" 92.19% for "RX Summ–Surg/Rad Seq" 93.03% for "Surgery" 92.15% for "Radiation" and 92.51% for "Chemo"

Analyzing the specific treatment categories, we observe that HistGradientBoosting performed remarkably well for "Chemo" achieving the highest accuracy of 95.86% for this treatment. For "RX Summ–Surg/Rad Seq" both HistGradientBoosting and Gradient Boosting demonstrated exceptional accuracy, securing the top two highest accuracy scores. Similarly, for "Surgery" and "Radiation" HistGradientBoosting and Gradient Boosting were the top-performing models, respectively, showcasing their effectiveness in predicting patient outcomes for these treatments.

Indeed, our observation is noteworthy and aligns with the current trend in the field of machine learning and deep learning in the domain of cancer treatment prediction. The comparison between traditional ML models boosted ML models and a DL model like MLP highlights the effectiveness of different approaches in achieving accurate predictions for cancer treatment outcomes.

The traditional ML models, including Decision Tree Classifier, Random Forest, Logistic Regression, Gaussian Naive Bayes, and SVM, have proven to be valuable tools in various machine learning tasks. However, when applied to cancer treatment prediction, their accuracy might be limited due to the complexity and high dimensionality of the data involved. Despite their respectable performance, these traditional models might struggle to capture intricate patterns and relationships within the data, especially in large and diverse datasets related to cancer treatment.

On the other hand, boosted ML models, such as Gradient Boosting, AdaBoost, and HistGradientBoosting, have demonstrated their potential to enhance the predictive power of traditional ML models. By combining multiple weak learners into a strong

ensemble model, boosted ML models can effectively improve accuracy and generalization. The boosting techniques provide a clever way to focus on the challenging samples, allowing the models to better adapt to the complexities of cancer treatment prediction.

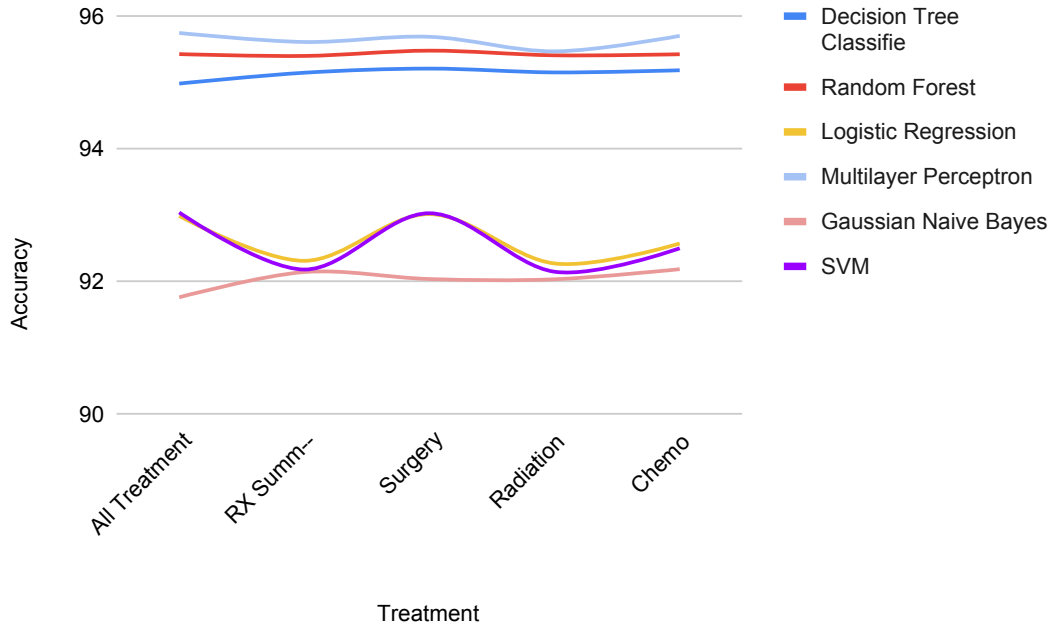

Figure 4.50: Traditional ML and DL Accuracy

From figure 4.50 we can say the DL model, MLP, has emerged as the best performer in terms of accuracy in your work comparing traditional ML and DL models. MLPs have the ability to learn intricate and non-linear relationships within the data, making them well-suited for cancer treatment prediction tasks. The deep architecture of MLP allows it to learn hierarchical representations, capturing high-level features and patterns in the data. This capability makes MLP a powerful model for handling complex and diverse data types encountered in cancer treatment prediction.

While MLP outperforms traditional ML models in accuracy, the boosted ML models provide the best overall accuracy. This could be attributed to the combined strength of boosted models in leveraging the advantages of traditional ML techniques while also addressing some of their limitations. The boosting process enhances the performance of the individual base models and adapts them to the specific characteristics of the cancer treatment data, resulting in improved accuracy.

In conclusion, the comparison of various ML models in your work has shed light on their individual strengths and weaknesses in cancer treatment prediction. The DL model, MLP, has demonstrated impressive accuracy, while the boosted ML models have proven to be the best performers overall. These findings emphasize the importance of exploring a diverse set of models and techniques to optimize accuracy in cancer treatment prediction. The combination of traditional ML models with boosting techniques and the incorporation of a powerful DL model like MLP can lead to more accurate and reliable predictions, offering valuable insights for personalized cancer treatment recommendations and improved patient outcomes.

4.3 Hypothesis

4.3.1 Feature Importance Ranking

In our quest to find precise cancer treatments through the analysis of the SEER dataset, we hypothesized that surgery, as a primary cancer treatment, may significantly impact patient survival rates. From our dataset, we have observed both positive and negative correlations between certain features and the "Survival Flag" which represents whether a patient will survive or not. Let's explore these correlations in detail and provide a comprehensive analysis:

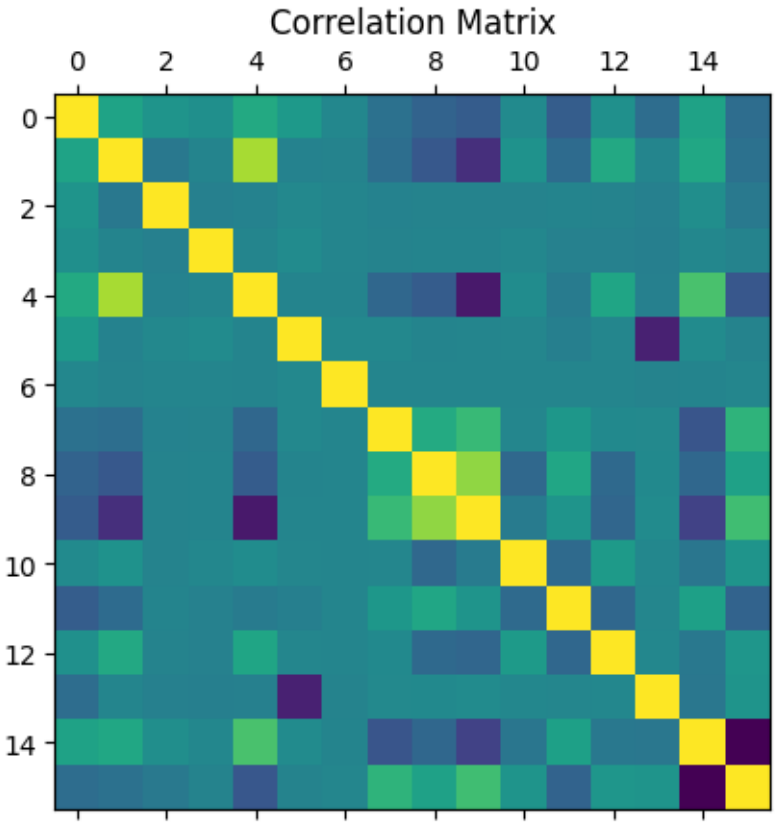

Figure 4.51: Correlation Matrix

Positive Correlations:

- **Summary Stage:** The positive correlation between the summary stage and the "Survival Flag" suggests that patients diagnosed with early-stage cancer have a higher likelihood of survival. Early-stage cancer is typically localized and confined to its site of origin, making it more amenable to curative treatments like surgery or radiation therapy. Patients diagnosed at an early stage can benefit from prompt and targeted interventions, resulting in improved treatment success and extended survival.

- **Surgery Radiation Sequence:** The positive correlation with the "Survival Flag" indicates that the sequence of treatments, specifically the combination of surgery and radiation, may impact patient outcomes positively. Patients who receive appropriate surgical and radiation treatments tailored to their

cancer type and stage might experience better disease control and prolonged survival.

- **Surgery Record:** The positive correlation implies that patients who did not undergo cancer-directed surgery might have chosen alternative treatments or were deemed unsuitable candidates for surgery. This finding highlights the importance of considering individualized treatment approaches based on patient preferences, health status, and tumor characteristics. It also underscores the need to explore and optimize alternative therapies for patients who are ineligible for surgery.

- **Radiation Record:** The positive correlation suggests that patients who received radiation therapy as part of their cancer treatment may have improved survival rates. Radiation therapy is effective in targeting and destroying cancer cells, either as a primary treatment modality or in combination with other therapies. Its positive association with patient survival reinforces the role of radiation in managing cancer and improving outcomes.

- **Chemotherapy Record :** The positive correlation with the "Survival Flag" indicates that patients who received chemotherapy might have better survival rates. Chemotherapy is a systemic treatment that circulates throughout the body to attack cancer cells. It is often used to treat aggressive or widespread cancers and can complement other treatments to enhance patient responses.

- **Months from diagnosis to treatment:** The positive correlation suggests that patients who received treatment sooner after diagnosis may have better survival rates. Reducing the time between diagnosis and treatment initiation is crucial in preventing cancer progression and metastasis, thereby increasing the chances of successful treatment outcomes.

- **First malignant primary indicator:** This feature's positive correlation with the "Survival Flag" implies that patients with primary tumors (initial occurrence) might have a higher likelihood of survival compared to those with recurrent or metastatic tumors. Early detection and treatment of primary tumors are vital in controlling cancer growth and improving overall survival.

Negative Correlations:

- **Age:** The negative correlation with the "Survival Flag" suggests that older patients may have a lower likelihood of survival. Advanced age is associated with increased comorbidities and reduced physiological resilience, potentially impacting the response to cancer treatments and overall survival.

- **Sex:** The negative correlation implies that male patients may have a lower chance of survival compared to female patients. This discrepancy could be attributed to biological differences, varying cancer types, or differences in treatment response between genders.

- **Marital status at diagnosis:** The negative correlation suggests that marital status at the time of cancer diagnosis may have some influence on patient outcomes. However, the extent of this impact requires further investigation, as it could be influenced by various social and psychological factors.

- **Site recode:** The negative correlation with the "Survival Flag" indicates that certain cancer sites might be associated with reduced survival rates. The anatomical location of the cancer could influence treatment options and response, impacting patient outcomes.

- **Chemotherapy recode:** While this feature showed a positive correlation, it also has a negative correlation with the "Survival Flag." This indicates that while chemotherapy can improve survival for some patients, it might not be effective for others, depending on the cancer type and individual response.

From the correlation data provided by the figure 4.51 correlation matrix, we can observe the following correlations for treatment types:

Treatment	**Correlation Value**
Surgery Record	0.434595
Surgery Radiation Sequence	0.213989
Radiation (Radiation Record)	0.112867
Chemotherapy (Chemotherapy Record)	-0.251130

Table 4.1: Correlation Values of Treatment

Surgery Record: The correlation between "Surgery Record" and the "Survival flag" suggests a positive correlation with a value of 0.434595. Correlation values range from -1 to +1, where +1 indicates a perfect positive correlation, 0 indicates no correlation, and -1 indicates a perfect negative correlation. There is a substantial and positive association between whether or not surgery was performed and the presence or absence of the "Survival flag." Based on this correlation, it appears that patients who did not get cancer-targeted surgery for a variety of reasons may have a decreased chance of surviving their condition. and those who had surgeries have had a higher rate of survival.

Surgery Radiation Sequence: The correlation value is 0.213989, indicating a positive correlation with the "Survival Flag". This suggests that patients who underwent surgery or a combination of surgery and radiation therapy tend to have a higher likelihood of survival.

Radiation (Radiation Record): The correlation value is 0.112867, suggesting a positive correlation with the "Survival Flag." This implies that patients who received radiation therapy as part of their treatment may have better survival outcomes.

Chemotherapy (Chemotherapy Record): The correlation value is -0.251130, indicating a negative correlation with the "Survival Flag." This implies that patients who received chemotherapy as part of their treatment may have lower survival rates, although it is essential to interpret this cautiously as correlation does not imply causation.

Our analysis of the correlation data revealed a notable finding, indicating that surgery record and surgery and RX Summ–Surg/Rad Seq demonstrated the highest positive correlation with the "Survival Flag." This means that there is a potential association between undergoing surgery as part of cancer treatment and better patient survival outcomes. The significance of this finding suggests that when considering various treatment options for cancer, surgery should be given strong consideration, especially if it is feasible and safe for the patient.
Surgery has long been recognized as a primary treatment approach for many types of cancer. It offers the potential to remove cancerous tissue, reduce tumor burden, and prevent the spread of cancer to other parts of the body. In certain cases, surgery can be curative, particularly when the cancer is detected at an early stage and localized to a specific area of the body.
However, it is crucial to emphasize that the decision to pursue surgery as a treatment option should be made in consultation with a multidisciplinary team of healthcare professionals, including oncologists, surgeons, and other specialists. Factors such as the type and stage of cancer, the location and size of the tumor, the patient's overall health, and their treatment preferences all play a pivotal role in determining the appropriateness of surgery.
In summary, our correlation analysis highlights potential factors that may impact patient survival outcomes in cancer treatment. Early-stage diagnosis, appropriate use of surgery, radiation, and chemotherapy, and timely treatment initiation appear to be crucial for better survival rates. However, age, gender, marital status, cancer site, and individual responses to treatments are additional factors that can influence patient outcomes. It is essential to use these insights alongside clinical expertise to develop personalized treatment strategies, considering both positive and negative correlations, to optimize cancer care and improve patient prognosis. Furthermore, recognizing that no certified cancer treatment exists to date emphasizes the need for continuous research and development of innovative therapies to enhance treatment efficacy and ultimately improve patient survival and quality of life.

4.3.2 Advancements in Cancer Treatment Strategies

Our research is centered on a number of hypotheses that arose from our extensive analysis of the outcomes of cancer treatment using machine learning models.

First, we have developed a hypothesis regarding the influence of surgery on patient survival. We predict that patients whose cancer-affected body sections can be completely removed or surgically excised will have higher survival rates than those whose cancerous tissues cannot be completely removed. This hypothesis is predicated on the idea that surgical removal of localized tumors can effectively arrest the progression of cancer and increase patient survival.

In addition, our research led us to investigate the significance of early cancer detection in treatment methods. We hypothesize that early detection of cancer allows for a greater variety of treatment options, and that non-invasive approaches may be effective in the management of cancers in their early stages. However, as cancer advances and becomes more advanced, surgery is likely to become a more prominent and essential treatment option. This hypothesis emphasizes the

significance of early detection as a determining factor for cancer patients' treatment strategies.

Furthermore, From Figure 4.1 in page 32 our analysis revealed the superior performance of deep learning models, especially the Multi-Layer Perceptron (MLP), in predicting the outcomes of cancer treatment. As a result, we hypothesize that deep learning models, which can automatically acquire complex patterns from the data, are more dependable and accurate for cancer-related datasets. The success of hybrid models, such as those employing boosting techniques such as Gradient Boosting and AdaBoost, provides additional evidence for the efficacy of incorporating sophisticated deep learning techniques.

These hypotheses have served as the guiding principles for our research endeavors. They serve as the foundation for our data analysis and interpretation, guiding us to valuable insights that can enhance cancer treatment decision-making and, ultimately, patient outcomes. Through this research, we hope to advance cancer treatment strategies and provide the medical community with valuable knowledge in the fight against cancer.

Chapter 5

Conclusion

Predicting the success rate or effectiveness of cancer treatments is of utmost importance to minimize the risk of mistreatment and optimize patient outcomes. However, there is a notable lack of comprehensive studies exploring the success rates of various treatments for different types of cancer. In this research, we addressed this crucial gap by leveraging machine learning techniques to predict cancer survivability for different treatment modalities.

Through the comparative analysis of multiple prediction models, we gained valuable insights into the relative predictive power of various machine learning techniques. The utilization of a large dataset allowed us to explore the performance of models for different cancer treatment scenarios comprehensively. The results demonstrated the potential of machine learning as a powerful tool for predicting cancer survival and guiding treatment decisions.

Despite the promising findings, our research also highlighted the vast potential for further development in this domain. Additional model building and refinement are essential to continuously improve the accuracy and reliability of predictions. Machine learning, with its ability to analyze vast amounts of complex data, offers a new frontier for predicting cancer survival and holds immense promise for transforming cancer care.

As we move forward, careful evaluation of the applicability and limitations of these models becomes imperative. Understanding the specific scenarios where each model excels and recognizing potential pitfalls will be crucial for successful implementation in clinical settings. Such assessments will enable us to build more robust and accurate prediction models that can be effectively utilized by healthcare professionals to tailor treatments to individual patients.

In conclusion, this research provides a stepping stone toward harnessing the power of machine learning in predicting cancer survival and enhancing the quality of cancer care. The exploration of multiple models and their comparative analysis has laid the foundation for future studies to refine existing models and develop novel approaches. By bridging the gap between machine learning and cancer treatment, we envision a future where predictive models play a pivotal role in delivering personalized and effective cancer therapies, ultimately leading to improved patient outcomes and a brighter outlook for cancer treatment.

5.1 Limitations

Despite the promising accuracy of our machine learning models in predicting the outcomes of cancer treatment, it is essential to recognize that cancer is a highly complex and heterogeneous disease. Due to individual biological factors, genetic mutations, and overall health status, each patient's response to treatment can vary considerably. While our models provide insightful information, they cannot account for all of the complexities that influence treatment response. The complexity of cancer treatment necessitates individualized medical decisions, and the predictions from our models should not serve as the sole premise for treatment plans, but rather as a supportive tool.

The effectiveness of machine learning models is dependent on the availability and quality of training data. We relied on the datasets available to us for our undertaking, which may have been limited in scope or represented specific demographics. The accuracy and generalizability of our models may be affected by data-based biases, such as the underrepresentation of certain patient groups or treatment regimens. Moreover, our models may not incorporate the most recent research findings, as the availability of data on cancer treatment is constantly evolving.

In our endeavor, deep learning models, including the Multilayer Perceptron (MLP), have demonstrated impressive precision. Nevertheless, the inherent complexity of deep learning architectures makes it frequently difficult to interpret the reasons behind their predictions. In the medical community, where comprehending the decision-making process is crucial for establishing confidence in the models and their recommendations, the dearth of interpretability could raise concerns. Continuing research focuses on the interpretability of deep learning models in healthcare.

We develop and evaluate our models based on the data available during the project's duration. The efficacy of machine learning models on new, unobserved data is susceptible to changes in data distribution or patient populations. As medical knowledge and treatment protocols evolve, periodic updates to our predictive models may be necessary to ensure their continued accuracy and relevance.

Regardless of advances in medical research and technology, it is essential to recognize that there is currently no universal remedy for cancer. Different cancer types, stages, and patient-specific factors can influence the efficacy of various cancer treatments. Our models' predictions are based on historical data and do not imply that a certified cancer treatment exists for all cases. When making treatment decisions, medical professionals should exercise caution and consider a holistic approach, taking into consideration clinical expertise, patient preferences, and the most recent medical evidence.

In summary, in spite of the fact that our machine learning models have demonstrated promising accuracy in predicting the outcomes of cancer treatment, they have inherent limitations. Recognizing the complexity of cancer, data limitations, interpretability challenges of deep learning models, generalizability concerns, and the lack of a certified universal cancer treatment is essential for the responsible and informed application of our models in actual clinical settings. As we continue to advance the field of machine learning in healthcare, we must adopt a circumspect and collaborative approach to cancer treatment prediction, incorporating the expertise of medical professionals and embracing the dynamic nature of cancer research.

5.2 Future Directions

Future research in this field would benefit from a greater focus on the fine-tuning of hyperparameters for machine learning models. Advanced optimization techniques, such as Bayesian optimization or genetic algorithms, can be used to find the optimal hyperparameter combination for each model. This strategy would not only improve the efficacy of the models but also facilitate more efficient convergence during the training process.

Adding domain-specific knowledge to the models can also be beneficial. Using domain-specific ontologies or medical knowledge graphs, for instance, can aid in encoding domain expertise and providing contextual data to models. This integration could improve the interpretability of the models and increase the clarity of the predictions for healthcare professionals.

In addition, the investigation of transfer learning techniques may yield fruitful avenues for future research. For cancer treatment prediction tasks, pre-trained models on large-scale healthcare or biomedical datasets could be adapted and refined. Transfer learning has demonstrated extraordinary success in a variety of natural language processing and computer vision tasks, and it could be used to leverage previously learned features that capture pertinent information for cancer treatment prediction.

Through open-sourcing the code and making the dataset publicly available, future researchers can find support and build upon this work. Sharing the trained models and codebase would allow other researchers to reproduce the results, compare their methodologies, and enhance the performance of the models. Collaboration with research communities, such as biomedical informatics and bioinformatics, could provide valuable insights and perspectives from various domains, thereby fostering interdisciplinary advances in cancer treatment prediction.

Researchers can engage in active discussions with healthcare professionals to better comprehend the clinical context and domain-specific challenges in order to produce more progressive work. Oncologists, radiologists, and other specialists can contribute to the development of more tailored and clinically pertinent prediction models by providing feedback. In addition, conducting prospective validation studies and clinical trials would provide validation of the model's efficacy in the real world, bridging the divide between research and clinical practice.

Lastly, it is crucial to address the ethical considerations surrounding the deployment of these models. It is essential to conduct analyses of fairness and bias to identify potential disparities in model predictions based on demographic or clinical factors. Implementing interpretability techniques to provide insights into model decision-making and yielding human-readable explanations for predictions would contribute to the development of trust and acceptance among healthcare professionals and patients.

Future research in this field can be advanced by investigating hyperparameter optimization, integrating domain-specific knowledge, and employing transfer learning techniques. Open-sourcing the code and dataset, collaborating with healthcare professionals, and undertaking prospective validation studies can advance this research and lead to more accurate and ethically sound cancer treatment prediction models for better patient care.

Bibliography

[1] R. E. Wright, "Logistic regression.," 1995.

[2] J. Fetting, P. Anderson, H. Ball, *et al.*, "Outcomes of cancer treatment for technology assessment and cancer treatment guidelines," *Journal of Clinical Oncology*, vol. 14, no. 2, pp. 671–679, 1996.

[3] A. Famili, W.-M. Shen, R. Weber, and E. Simoudis, "Data preprocessing and intelligent data analysis," *Intelligent data analysis*, vol. 1, no. 1, pp. 3–23, 1997.

[4] Y. Freund and L. Mason, "The alternating decision tree learning algorithm," in *icml*, vol. 99, 1999, pp. 124–133.

[5] A. Pinkus, "Approximation theory of the mlp model in neural networks," *Acta numerica*, vol. 8, pp. 143–195, 1999.

[6] S. Kwek, "Iboost: Boosting using an i nstance-based exponential weighting scheme," pp. 245–257, 2002. DOI: 10.1007/3-540-36755-1_21.

[7] S. Vishwanathan and M. N. Murty, "Ssvm: A simple svm algorithm," in *Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN'02 (Cat. No. 02CH37290)*, IEEE, vol. 3, 2002, pp. 2393–2398.

[8] C. Stein, "Modifiable risk factors for cancer," *British Journal of Cancer*, vol. 90, pp. 299–303, 2004. DOI: 10.1038/sj.bjc.6601509.

[9] J. W. Knopf, "Doing a literature review," *PS: Political Science & Politics*, vol. 39, no. 1, pp. 127–132, 2006.

[10] H. Ming, "An improved decision tree classification algorithm based on id3 and the application in score analysis," *2009 Chinese Control and Decision Conference*, pp. 1876–1879, 2009. DOI: 10.1109/CCDC.2009.5192865.

[11] D. Robinson, "An analysis of temporal and generational trends in the incidence of anal and other hpv-related cancers in southeast england," *British Journal of Cancer*, vol. 100, pp. 527–531, 2009. DOI: 10.1038/sj.bjc.6604871.

[12] N. Viney, "Assessing the impact of land use change on hydrology by ensemble modelling(luchem) ii: Ensemble combinations and predictions," *Advances in Water Resources*, vol. 32, pp. 147–158, 2009. DOI: 10.1016/J.ADVWATRES.2008.05.006.

[13] M. Poorkiani, "The effect of rehabilitation on quality of life in female breast cancer survivors in iran," *Indian Journal of Medical and Paediatric Oncology : Official Journal of Indian Society of Medical Paediatric Oncology*, vol. 31, pp. 105–109, 2010. DOI: 10.4103/0971-5851.76190.

[14] *Understanding statistics used to guide prognosis and evaluate treatment*, en, https://www.cancer.net/navigating-cancer-care/cancer-basics/understanding-statistics-used-guide-prognosis-and-evaluate-treatment, Accessed: 2023-7-30, Apr. 2010.

[15] A. Alama, "Targeting cancer-initiating cell drug-resistance: A roadmap to a new-generation of cancer therapies?" *Drug discovery today*, vol. 17 9-10, pp. 435–42, 2012. DOI: 10.1016/j.drudis.2011.02.005.

[16] S. A. Hussain and R. Sullivan, "Cancer control in bangladesh," *Japanese journal of clinical oncology*, vol. 43, no. 12, pp. 1159–1169, 2013.

[17] S. M. A. Hussain, "Comprehensive update on cancer scenario of bangladesh," *South Asian journal of cancer*, vol. 2, no. 04, pp. 279–284, 2013.

[18] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," *Frontiers in neurorobotics*, vol. 7, p. 21, 2013.

[19] R. E. Schapire, "Explaining adaboost," in *Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik*, Springer, 2013, pp. 37–52.

[20] P. J. M. Ali, R. H. Faraj, E. Koya, P. J. M. Ali, and R. H. Faraj, "Data normalization and standardization: A technical report," *Mach Learn Tech Rep*, vol. 1, no. 1, pp. 1–6, 2014.

[21] M. S. Hossain, "Diagnosed hematological malignancies in bangladesh - a retrospective analysis of over 5000 cases from 10 specialized hospitals," *BMC Cancer*, vol. 14, pp. 438–438, 2014. DOI: 10.1186/1471-2407-14-438.

[22] G. Zhang, "Decreased expression of sirt6 promotes tumor cell growth correlates closely with poor prognosis of ovarian cancer.," *European journal of gynaecological oncology*, vol. 36 6, pp. 629–32, 2014. DOI: 10.7727/WIMJ.2014.057.

[23] G. Pennock, "The evolving role of immune checkpoint inhibitors in cancer treatment.," *The oncologist*, vol. 20 7, pp. 812–22, 2015. DOI: 10.1634/theoncologist.2014-0422.

[24] G. Biau and E. Scornet, "A random forest guided tour," *Test*, vol. 25, pp. 197–227, 2016.

[25] K. Potdar, T. S. Pardawala, and C. D. Pai, "A comparative study of categorical variable encoding techniques for neural network classifiers," *International journal of computer applications*, vol. 175, no. 4, pp. 7–9, 2017.

[26] K.-H. Yu, A. L. Beam, and I. S. Kohane, "Artificial intelligence in healthcare," *Nature biomedical engineering*, vol. 2, no. 10, pp. 719–731, 2018.

[27] E. Alawadhi, "Cancer survival by stage at diagnosis in kuwait: A population-based study," *Journal of Oncology*, vol. 2019, 2019. DOI: 10.1155/2019/8463195.

[28] M. D. Ganggayah, N. A. Taib, Y. C. Har, P. Lio, and S. K. Dhillon, "Predicting factors for survival of breast cancer patients using machine learning techniques," *BMC medical informatics and decision making*, vol. 19, pp. 1–17, 2019.

[29] A. Guryanov, "Histogram-based algorithm for building gradient boosting ensembles of piecewise linear decision trees," in *Analysis of Images, Social Networks and Texts: 8th International Conference, AIST 2019, Kazan, Russia, July 17–19, 2019, Revised Selected Papers 8*, Springer, 2019, pp. 39–50.

[30] H. Kamel, D. Abdulah, and J. M. Al-Tuwaijari, "Cancer classification using gaussian naive bayes algorithm," in *2019 international engineering conference (IEC)*, IEEE, 2019, pp. 165–170.

[31] J. Nissen, R. Donatello, and B. Van Dusen, "Missing data and bias in physics education research: A case for using multiple imputation," *Physical Review Physics Education Research*, vol. 15, no. 2, p. 020 106, 2019.

[32] T. Prasanth, "Effective big data retrieval using deep learning modified neural networks," *Mobile Networks and Applications*, vol. 24, pp. 282–294, 2019. DOI: 10.1007/S11036-018-1204-Y.

[33] K. Rendle, "Cancer symptom recognition and anticipated delays in seeking care among u.s. adults.," *American journal of preventive medicine*, vol. 57 1, e1–e9, 2019. DOI: 10.1016/j.amepre.2019.02.021.

[34] A. Shrestha and A. Mahmood, "Review of deep learning algorithms and architectures," *IEEE access*, vol. 7, pp. 53 040–53 065, 2019.

[35] J. A. M. Sidey-Gibbons, "Machine learning in medicine: A practical introduction," *BMC Medical Research Methodology*, vol. 19, 2019. DOI: 10.1186/s12874-019-0681-4.

[36] S. Huang, J. Yang, S. Fong, and Q. Zhao, "Artificial intelligence in cancer diagnosis and prognosis: Opportunities and challenges," *Cancer letters*, vol. 471, pp. 61–71, 2020.

[37] B. W. Papenberg, J. L. Allen, S. M. Markwell, *et al.*, "Disparate survival of late-stage male oropharyngeal cancer in appalachia," *Scientific Reports*, vol. 10, no. 1, p. 11 612, 2020.

[38] A. Robey, "Model-based robust deep learning," *ArXiv*, vol. abs/2005.10247, 2020.

[39] A. Rácz, D. Bajusz, and K. Héberger, "Effect of dataset size and train/test split ratios in qsar/qspr multiclass classification," *Molecules*, vol. 26, no. 4, p. 1111, 2021.

[40] R. Rafique, S. R. Islam, and J. U. Kazi, "Machine learning in the prediction of cancer therapy," *Computational and Structural Biotechnology Journal*, vol. 19, pp. 4003–4017, 2021, ISSN: 2001-0370. DOI: https://doi.org/10.1016/j.csbj.2021.07.003. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2001037021002932.

[41] A. N. Alzubaidi, S. Sekoulopoulos, J. Pham, V. Walter, J. G. Fuletra, and J. D. Raman, "Incidence and distribution of new renal cell carcinoma cases: 27-year trends from a statewide cancer registry," *Journal of Kidney Cancer and VHL*, vol. 9, no. 2, p. 7, 2022.

[42] S. A. M. MOHAMMED, "The investigation of anticancer effect of scorpion carapace," Ph.D. dissertation, 2022.

[43] A. A. Moskalewicz, "Projecting the future prevalence of childhood cancer in ontario using microsimulation modeling," Ph.D. dissertation, University of Toronto (Canada), 2022.

[44] *Cancer*, en, https://www.who.int/news-room/fact-sheets/detail/cancer, Accessed: 2023-7-29.

[45] *Cancer today*, en, https://gco.iarc.fr/today/home, Accessed: 2023-7-28.

[46] *Childhood cancer*, en, https://www.who.int/news-room/fact-sheets/detail/cancer-in-children, Accessed: 2023-7-28.

[47] The International Agency for Research on Cancer (IARC), *Global cancer observatory*, en, https://gco.iarc.fr/, Accessed: 2023-7-28.