M.Tech Dissertation Phase-II (COC7912) Report on

MULTIMODAL DATA ANALYTICS FOR PREDICTING THE SURVIVAL OF CANCER PATIENTS

SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE AWARD OF THE DEGREE OF

Master of Technology

IN

COMPUTER SCIENCE & ENGINEERING

BY

HASAN SHAIKH
21COSM111

Under the Guidance of
Prof. Rashid Ali

Department of Computer Engineering
Zakir Husain College of Engineering & Technology
Aligarh Muslim University
Aligarh (India)-202002, India

2022-2023

Dated:

Declaration

The work presented in the Dissertation Phase – II (COC7912) titled **"Multimodal data analytics for predicting the survival of Cancer Patients**" submitted to the Department of Computer Engineering, Zakir Husain College of Engineering and Technology, Aligarh Muslim University, Aligarh, for the award of the degree of Master of Technology in Computer Science & Engineering, during the **session 2022-23**, is my original work. I have neither plagiarized nor submitted the same work for the award of any other degree.

Date:
Place: Aligarh

Hasan Shaikh
21COSM111

Dated:

Certificate

On the basis of declaration given by candidate**,** this is to certify that the Dissertation titled **“Multimodal data analytics for predicting the survival of cancer patients”,** being submitted by “**Hasan Shaikh**”, in partial fulfillment of the requirements for the award of the degree of **Master of Technology** in **Computer Science & Engineering**, during the **session 2022-23**, in the Department of Computer Engineering, Zakir Husain College of Engineering and Technology, Aligarh Muslim University Aligarh, is a record of candidate’s work carried out by his under my supervision and guidance.

Prof. Rashid Ali
Professor,
Department of Computer Engineering,
ZHCET, AMU, Aligarh

Acknowledgement

In the name of Allah, the Most Gracious, the Most Merciful, I would like to start my acknowledgment. I am profoundly grateful for the blessings and wisdom I have received from Allah (SWT) that provided me with the strength and patience to complete this Dissertation.

My deepest gratitude goes to my supervisor, Prof. Rashid Ali, whose guidance, support, and insight were invaluable throughout this process. His unfailing enthusiasm and profound knowledge have been a driving force behind this work.

To my father, who made the ultimate sacrifice to provide me with the life and opportunities I have today, I dedicate this work. My mother, whose unwavering love and visionary outlook have positioned me where I am today, deserves my sincerest thanks. Her support has been my pillar of strength throughout this journey. I wish to express my deep love and gratitude to my sister and brother, who have been a constant source of support and motivation. Their unfailing faith in me and regular mentoring have been immensely helpful.

My heartfelt appreciation goes to Ehtesham Sana and Mahtab Alam for his invaluable intellectual discussions, honest suggestions, and rigorous reviews that have shaped this Dissertation. Finally, my sincere appreciation goes to everyone else who, directly or indirectly, contributed to the completion of this Dissertation. It has been a long journey, but every step of the way has been worth it. May Allah's blessings be upon you all.

Hasan Shaikh
21COSM111

TABLE OF CONTENT

Acronyms

AI	Artificial Intelligence
ML	Machine Learning
DL	Deep Learning
CNN	Convolution Neural Network
GRL	Graph Representation Learning
MRL	Multimodal Representation Learning
EDA	Exploratory Data Analysis
CNA	Copy Number Alteration
MIMIC-III	Medical Information Mart for Intensive Care III
TCGA-BRCA	The Cancer Genome Atlas Breast Invasive Carcinoma Collection
SEER	The Surveillance, Epidemiology, and End Results
METABRIC	Molecular Taxonomy of Breast Cancer International Consortium

List of Figures

List of Tables

Abstract

Cancer Survival prediction, traditionally focused on time-to-event data such as the interval from cancer diagnosis to the eventual outcome, has been a foundational element of oncological research. In the past, the domain has been directed by classical machine learning techniques in particular; techniques like logistic regression and the support vector machines model have been utilized to predict the likelihood of such events based on available data. However, these traditional methods have challenges with handling high-dimensional or non-linear patterns in complex datasets.

This dissertation moves beyond these classical machine learning techniques, exploring the transformative potential of Deep Learning for cancer survival predictions. The research's primary goal is an exploration of multimodal data types and their significance in predicting cancer survival prediction. Two main objectives guide this journey: the first revolves around a rigorous exploratory analysis of both unimodal and multimodal datasets. In this phase, we aim to understand the complex relationships among different data modalities, gaining insights into their advantages and potential downsides. Simultaneously, we evaluate traditional Machine Learning methods, highlighting the potential challenges they may encounter with multimodal datasets.

The second objective shifts the focus to the utilization of advanced Deep Learning techniques, specifically the Gated Attention Convolution Neural Network model, to enhance prediction accuracy. Recognizing the strength of Deep Learning in processing large and complex datasets, this research harnesses its capabilities for the multimodal data derived from the well-known METABRIC dataset, as a well-known multimodal dataset of breast cancer patients. We've built our model to handle different types of data separately: one part for clinical data, another for gene expression, and a third for copy number changes. This way, we can really focus on the unique details of each type of data for better results and leading to a more holistic and comprehensive analysis.

The main thing of this research is a comparative analysis where the proposed Multimodal Gated Attention CNN model is evaluated against its unimodal models and other state-of-the-art multimodal prediction models. The results showed that our multimodal gated attention CNN model did really well as compared to the others.

In conclusion, this dissertation signifies a paradigm shift from classical machine learning techniques to a more detailed, Deep Learning-focused approach. Through detailed exploration, analysis, and experimentation, this research aspired to set a new standard for more accurate, detailed, and comprehensive predictions in the field of cancer survival, potentially shaping future oncological research and clinical practices.

CHAPTER 1: INTRODUCTION

This chapter provides the background that informed the selection of this dissertation topic. It outlines the objectives, scope within its research domain, and the organization of the report. The subsequent chapters detail the various stages undertaken to complete this dissertation.

1.1 Cancer Survival Prediction

Survival prediction is fundamentally the study of time until certain events happen. When considering cancer survival, it typically refers to the duration between the diagnosis of the disease and the subsequent death attributable to it. In an ideal scenario, every patient involved in survival studies would be monitored continuously until the event under consideration, i.e., death in the case of cancer survival, is observed. However, the reality of clinical practices often diverges from this ideal. Patients might exit clinical monitoring before the event has happened, or in some situations, the event may not occur at all if the patient dies from a different cause. When the anticipated event remains unobserved, the last recorded time point of patient contact is labeled as the censoring time [1].

Historically, survival prediction was primarily rooted in statistical techniques. The foundational principle of these techniques was to create models that could predict the likelihood of an event occurring at a specific time, given the available data [2]. One of the most conventional methods employed in these analyses is the Kaplan-Meier estimator [3]. This non-parametric statistic is used to estimate the survival function from lifetime data. Another widely recognized method is the Cox Proportional-Hazards model, which examines the relationship between the survival of a patient and several explanatory variables [4].

However, while these statistical methods have been instrumental in shaping early research in survival analysis, they come with inherent limitations. For one, they often assume that the underlying risks associated with the survival data are proportional over time, which might not always be the case. Furthermore, traditional statistical methods can sometimes struggle when faced with high-dimensional data or when trying to capture non-linear patterns inherent in complex datasets. The requirement of manual feature engineering in these models also posed significant challenges, as it is not only labor-intensive but could also miss out on nuanced patterns not immediately evident to human researchers [5].

With the technological advancements of the 21st century, the field of Machine Learning, and more specifically, Deep Learning, began to make its mark on cancer survival prediction. Deep Learning, a subset of Machine Learning, employs artificial neural networks to automatically identify patterns and structures in large datasets [6]. These neural networks, inspired by the neural structures of the human brain, can process vast amounts of data, recognize complex non-linear relationships within this data, and do so without the necessity for manual feature engineering [7].

The application of Deep Learning in survival prediction offers a multitude of advantages. Firstly, these models are adept at handling the high-dimensional data commonly found in medical research. With the rise of genomics and personalized medicine, the data associated with each patient has grown both in volume and complexity. Deep Learning models, with their capacity for automatic feature extraction, can seamlessly navigate this data landscape [8].

Secondly, Deep Learning models are inherently flexible. Traditional statistical models often come with predefined assumptions about the data they handle. In contrast, neural networks in Deep Learning can adapt their structure based on the data, ensuring a more tailored and accurate prediction. Finally, the integration of Deep Learning in survival prediction has facilitated the processing of varied data types. Whether it's structured data from clinical trials or unstructured data like medical images, Deep Learning models can integrate and interpret them with unprecedented accuracy [9].

While the potential of Deep Learning in revolutionizing cancer survival prediction is undeniable, it's essential to remember that the field is still evolving. Continuous research is vital to refine these models further, understand their limitations, and ensure their predictions are both accurate and interpretable for clinical applications.

In conclusion, the trajectory of cancer survival prediction has witnessed a significant transformation from traditional statistical techniques to the contemporary landscape dominated by Deep Learning. As the field continues to evolve, it holds the promise of not only more accurate predictions but also a deeper understanding of cancer and its myriad complexities.

1.2 Multimodal Data Analytics

The digital age, often referred to as the era of big data, has brought about an unprecedented accumulation of data. This vast and rapidly growing repository of information is often sourced from diverse origins, each contributing unique formats and types of data. Such diverse data sets, when viewed collectively, are termed as "multimodal data". This data not only offers multiple perspectives of the observed processes or entities but often provides complementary insights, enriching the overall understanding [10].

Multimodal data analysis is not just about dealing with data volume but also its variety. Given the heterogeneous nature of this data, with each modality offering a distinct perspective, the challenge lies in harnessing the cumulative insights they offer. While each data type has its own unique properties, they often complement and enrich each other. This realization has led researchers to develop sophisticated methods to cohesively analyze these diverse data sources, aiming to derive enhanced insights that wouldn't be possible from any single data source [11].

Multimodal data analysis is not a mere academic exercise. Its importance becomes particularly pronounced in fields that require a comprehensive and holistic view of complex systems. One such field is oncology.

- **Holistic Understanding:** In cancer research and treatment, understanding the disease from multiple perspectives is not just beneficial but essential. A single data type might provide insights into one aspect of the disease, but cancer's complexity necessitates a more multifaceted approach.
- **Improved Prognosis:** Combining data types can lead to more accurate prognosis, helping medical professionals devise better treatment strategies tailored to individual patients.

In the present research, the focus lies on three specific data types: clinical data, gene expression, and copy number alteration.

1. Clinical Features

Clinical data stands as a foundational pillar in understanding a patient's health journey. This data type extends beyond basic metrics like age and gender. It delves deep into the intricacies of a specific diagnosis, detailing the type and stage of the disease. Treatment records further enrich this dataset, surgical interventions undertaken, and the outcomes of each therapeutic strategy. Regular health check-ups and their findings are meticulously recorded, offering a timeline of the disease's progression or regression. Furthermore, any complications or side effects experienced during treatments are documented [12]. The importance of clinical data cannot be understated. It's not merely a record but an essential tool that enables medical professionals to tailor treatments to individual patients, ensuring optimized outcomes. It also aids in predicting potential risks or complications, facilitating preemptive interventions. Moreover, by analyzing broader trends and patterns, one can gain insights into disease progression and response to various treatments.

2. Gene Expression Features

Every human cell is a reservoir of genetic information, but not all genes manifest their presence actively at all times. Gene expression encapsulates the process where specific genes are activated to produce functional products, predominantly proteins. Depending on a plethora of factors, genes can exhibit varying levels of activity. The patterns of gene expression are not mere biological phenomena; they're gateways to understanding the very essence of life processes [13]. Analyzing these patterns can aid in identifying potential genetic anomalies or susceptibilities. The level at which a gene is expressed can offer a glimpse into its current activity. Anomalies in these levels can sometimes provide clues about various health conditions, including predispositions to certain diseases.

Specifically, the features of gene expression are:

- **Under-expression (-1):** This signifies a reduced gene activity. Such a decrease could potentially flag issues like the absence of certain proteins or their malfunctioning, both of which can have significant health implications.
- **Over-expression (1):** On the opposite spectrum, an elevated gene activity might indicate overactive protein production. Such heightened activity could be a sign of certain diseases or conditions.
- **Baseline (0):** This represents the standard or normal gene activity, indicative of a balanced and typical protein synthesis.

3. Copy Number Alteration (CNA) Features

Our genetic material isn't a static entity. Over time, and sometimes due to specific triggers, the number of copies of certain genes in our DNA can change. These variations, known as Copy Number Alterations, can influence gene activity and, consequently, the amount of protein a gene produces. Such changes can have cascading effects on health, making the tracking and understanding of CNAs paramount in medical research [14].

The specific features detailing these alterations are:

- **Homozygous deletion (-2):** This represents the complete absence of a gene, which likely results in a total loss of its associated function. Without the gene, the specific protein it codes for might be missing, potentially leading to health issues.
- **Hemizygous deletion (-1):** Here, only one of the two copies of the gene is absent. This could potentially halve the gene's activity, leading to a reduced function but not a total loss.
- **Neutral/no change (0):** This indicates the presence of the expected two copies of a gene, suggesting no significant alterations in its activity.
- **Gain (1):** Representing the presence of extra copies of the gene, this could lead to increased gene activity and, potentially, an overproduction of the associated protein.
- **High-level amplification (2):** This denotes the presence of many extra copies of the gene. Such a significant surge in gene copies could drastically amplify its activity.

The convergence of varied data types offers a promising path in the domain of cancer research. While each data modality offers its unique insights, their combined analysis promises a richer, more comprehensive understanding. As the research frontier advances, the potential of multimodal data analysis in revealing the intricate facets of cancer, and possibly other diseases, remains a domain of significant interest and potential.

1.3 Data Analytics for Cancer Patient Survival Prediction

Data analytics, at its core, is the systematic computational analysis of data or statistics. It involves the process of examining large datasets to uncover hidden patterns, unknown correlations, and other useful information. In the realm of oncology, data analytics plays a pivotal role. It enables healthcare professionals to sift through vast amounts of patient data—ranging from demographic information to detailed genomic profiles—to extract meaningful insights. For cancer patient survival prediction, this means identifying risk factors, understanding disease progression, and tailoring treatment plans based on predictive models. By leveraging data analytics, oncologists can move beyond one-size-fits-all treatments to more personalized approaches, significantly impacting patient survival rates and quality of life.

Machine learning, a subset of artificial intelligence, uses algorithms to parse data, learn from it, and then make a determination or prediction about something in the world. In cancer survival prediction, machine learning techniques play a transformative role. Algorithms like logistic regression can evaluate the likelihood of certain events (like recurrence or remission), while decision trees and random forests provide visual and intuitive ways to classify patients based on various attributes. These techniques help in identifying patterns in patient data that might not be apparent through traditional statistical methods. For instance, a machine learning model might find that a combination of genetic markers and lifestyle factors can accurately predict survival rates for specific cancer types, thereby aiding in more targeted and effective treatment strategies.

Deep learning, an advanced form of machine learning, involves neural networks with multiple layers that can learn and make intelligent decisions on their own. This approach is particularly beneficial in handling complex patterns and large datasets, common in medical data analysis. Deep learning techniques, such as convolutional neural networks (CNNs), are instrumental in analyzing medical images like MRIs or CT scans to detect tumors and assess their progression. Recurrent neural networks (RNNs), another deep learning technique, are adept at processing sequential data, making them suitable for analyzing patient treatment histories or time-series data from monitoring devices. In cancer patient survival prediction, these techniques enable the processing of more nuanced and intricate data patterns, leading to highly accurate predictions. For example, a deep learning model can analyze a patient's entire medical history, along with genetic data, to predict how they will respond to various treatment plans, greatly enhancing the personalization and effectiveness of cancer care.

1.4 Objective of the Study

The primary objective of our research is to deeply examine various types of data and understand their significance in predicting cancer survival rates. We've divided our study into two primary objectives to comprehensively address the broader goal.

Objective 1: Exploratory Analysis of Multimodal Data used for Cancer Survival Prediction

Our first objective focuses on a deep exploration of multimodal data, especially focusing its impact on the performance of cancer survival prediction models. We analyzed both unimodal and multimodal datasets, attempting to understand the complex relationships between different modalities. Through this, we gained insights into the strengths and possible limitations of each data modality. Additionally, we assessed traditional Machine Learning (ML) methods, highlighting certain challenges they might face when dealing with multimodal data. This foundational analysis is crucial as it sets the stage for our subsequent objective.

Objective 2: Use of Deep Learning for Multimodal Data Analysis in Cancer Survival Prediction

Following our initial exploration, the second objective moves towards leveraging advanced deep learning methods, to achieve better prediction results. Recognizing deep learning's capability to process large and intricate data, we sought to harness its potential for our multimodal data. A key element of our approach is the use of the "Gated Attention Convolution Neural Network model", a notable deep learning method adept at managing multimodal data. Instead of a single underlying model, we constructed individual models for each data type, carefully pinpointing unique aspects from each. By merging these aspects, we developed a holistic model. Our hope is that this integrated model, informed by insights from all modalities, will yield more accurate and detailed predictions about cancer survival.

1.5 Scope and Significance of the Study

A. Scope of the Study

Our research is focused on harnessing the capabilities of multimodal data for cancer survival predictions. The study is structured around two central objectives. The first objective delves into a comprehensive analysis of both unimodal and multimodal datasets, aiming to elucidate the relationships inherent within each modality. Subsequently, we employ the "Gated Attention CNN" to navigate and analyze this data, recognizing its potential in handling intricate data landscapes. Instead of a monolithic model, we develop specific models tailored for each data type, including clinical data, gene expression, and copy number alteration. These models are then synthesized to create a unified framework that provides an enriched perspective.

Beyond the computational aspects, the study also focuses on the practical implications of generating accurate survival predictions. Enhanced prediction accuracy can lead to optimized treatments, better resource allocation, and a streamlined healthcare delivery process. Furthermore, accurate predictions can aid clinical decision-making, enabling more informed treatment choices, early identification of high-risk patients, and timely interventions. Additionally, our exploration into biomarkers and molecular mechanisms contributes to a deeper understanding of cancer biology and has the potential to influence the development of targeted therapies. Importantly, the methodologies developed in this study could also be adapted for prognosis predictions in other diseases.

B. Significance of the Study

This research holds considerable importance in the broader context of oncological studies. It seeks to address a critical challenge in medicine: creating individualized cancer survival predictions to enhance patient outcomes. The techniques and methods proposed in this study could redefine how healthcare resources are utilized, emphasizing efficient resource distribution and early, precise interventions. Within the domain of oncology research, this study aims to provide valuable insights and serve as a foundational reference for subsequent research, offering a wealth of knowledge to the broader medical community.

1.6 Methodology Overview

In this study, we've followed a systematic approach to delve into cancer survival prediction. Here's a simple breakdown of our methodology:

Datasets: We used the METABRIC dataset, which is publicly available. This dataset combines information from three areas: clinical data, gene expression, and copy number alteration related to breast cancer.

Exploratory Data Analysis (EDA): At this stage, we checked for any missing data and analyzed both the clinical and genomic data. Our goal was to understand how different factors relate to the overall survival rate of patients.

Modelling: For our prediction model, we propose a deep learning model called "Gated Attention Convolution Neural Network." We first created separate models for each type of data (clinical, gene expression, copy number alteration) and then combined them into one overall model to get a broader understanding. The idea was to see if using all types of data together gives us better results.

Evaluation: Once our models were ready, we tested them using a method called 10-fold cross-validation. We also looked at several performance metrics like accuracy, precision, sensitivity, and the area under the curve to see how well our models performed.

Comparison: Lastly, we compared our combined model with models that use only one type of data. We also compared our model with other popular state-of-the-art models to see where ours stands.

Overall, this step-by-step approach helped us systematically tackle the challenge of predicting cancer survival using multiple types of data.

1.7 Organization of the Research

The introductory chapter has set the foundational context for the research journey ahead. By highlighting the importance of cancer survival prediction and the potential of multimodal data analytics, we've established the basis upon which the subsequent chapters will build. Here's a concise breakdown of the content and focus of each chapter:

Chapter 2 - Literature Review: Diving into the academic backdrop, this chapter systematically examines the existing body of research. It explores both traditional statistical techniques and modern AI techniques used in unimodal and multimodal data scenarios for cancer survival prediction.

Chapter 3 – Design of Cancer Survival Prediction System: At the core of the dissertation, this chapter reveals the methodological framework of the study. It includes an in-depth discussion of the proposed system's architecture optimized for multimodal datasets, an exploration of exploratory data analysis (EDA) on clinical and genomic data, and a detailed look at the specific aspects of the gated attention CNN model, along with its parameters.

Chapter 4 – Implementation of Cancer Survival Prediction System: The chapter begins with a detailed discussion of the tools and techniques, offering an insight into the programming environment, hardware specifications, software details, and the libraries and packages that form the backbone of the research.

Chapter 5 – Experiments & Results: Transitioning from design to implementation, this chapter documents the empirical phase. It highlights the dataset's characteristics, and presents a comprehensive analysis of the results. This includes comparing the proposed model's performance in unimodal versus multimodal contexts and benchmarking it against established state-of-the-art multimodal prediction frameworks.

Chapter 6 – Conclusion & Future Work: Concluding the dissertation, this chapter offers a reflection on the research journey, summarizing the key findings and addressing the inherent limitations. Furthermore, it looks ahead to the future, highlighting potential directions for further exploration and development in the domain.

By progressing through these chapters, readers are guided on a comprehensive journey from understanding the foundational concepts to appreciating the nuanced results and implications of the study.

CHAPTER 2: AN OVERVIEW OF MACHINE LEARNING TECHNIQUES FOR CANCER PATIENTS SURVIVAL PREDICTION

Cancer survival prediction has witnessed significant advancements over time, with the scientific community continuously exploring and innovating methodologies to improve prognosis accuracy. Historically, classical statistical techniques laid the foundation for survival analysis, providing valuable insights based on rigorous mathematical formulations. However, with the onset of computational advancements and the expansion of data, Artificial Intelligence (AI)-based techniques have emerged as a powerful complement, pushing the boundaries of predictive accuracy and presenting new pathways for exploration. This chapter delves into these two predominant approaches, aiming to provide a comprehensive overview of the methodologies and their respective contributions to the field of cancer survival prediction.

The field of cancer survival prediction has long relied on classical statistical techniques to model and analyze survival data. These methods, deeply rooted in rigorous mathematical theory, provide a structured approach to understanding and predicting the time-to-event data, especially when the event in question is the survival or death of a patient [15]. Broadly, these classical statistical techniques can be categorized into three main groups: parametric, non-parametric, and semi-parametric models mentioned in figure 1.

2.1 Artificial Intelligence Techniques for Cancer Survival Prediction

AI techniques offer a compelling approach to cancer survival prediction, bringing a host of advantages to the table. In the intricate landscape of cancer research, where data sources range from clinical records to genomics and medical imaging, AI excels at handling this multidimensional data. Its strength lies in its capacity to integrate and analyze diverse data types, extracting valuable insights from the wealth of information available. AI models can automatically discover relevant features from high-dimensional genomic data and medical images, uncovering intricate patterns and non-linear relationships that may remain hidden through traditional methods. This ability to capture complex dependencies is pivotal for precise and reliable survival predictions [16].

One of AI's notable attributes is its scalability, allowing it to efficiently process and derive insights from large datasets. This scalability facilitates comprehensive studies across extensive patient cohorts, enabling researchers to draw statistically robust conclusions. AI's personalized predictions take into account individual patient characteristics, medical histories, and genomic profiles. This personalized approach opens avenues for tailoring treatment plans to a patient's unique needs, potentially improving treatment outcomes and patient care. Moreover, AI's role in early cancer detection, data integration from multiple sources, task automation, resilience in handling noisy data, and the interpretability of results adds significant value to the field of cancer survival prediction [17].

Ensemble techniques within AI further enhance its utility. By combining predictions from multiple models, these methods improve prediction accuracy and reduce overfitting. This is particularly advantageous in scenarios where a single model may struggle to generalize well across diverse datasets. Ensemble learning leverages the diversity of individual models, harnessing their collective predictive power to yield more robust and accurate survival predictions [18].

In Figure 2.1, we provide an insightful visualization of the diverse landscape of data analytics techniques for cancer survival prediction. This comprehensive classification encompasses classical Machine Learning methods, including classification models, ensemble learning strategies, and advanced deep learning architectures such as Convolutional Neural Networks (CNNs), Graph Representation Learning (GRL), Multimodal Representation Learning (MRL), and Attention Models.

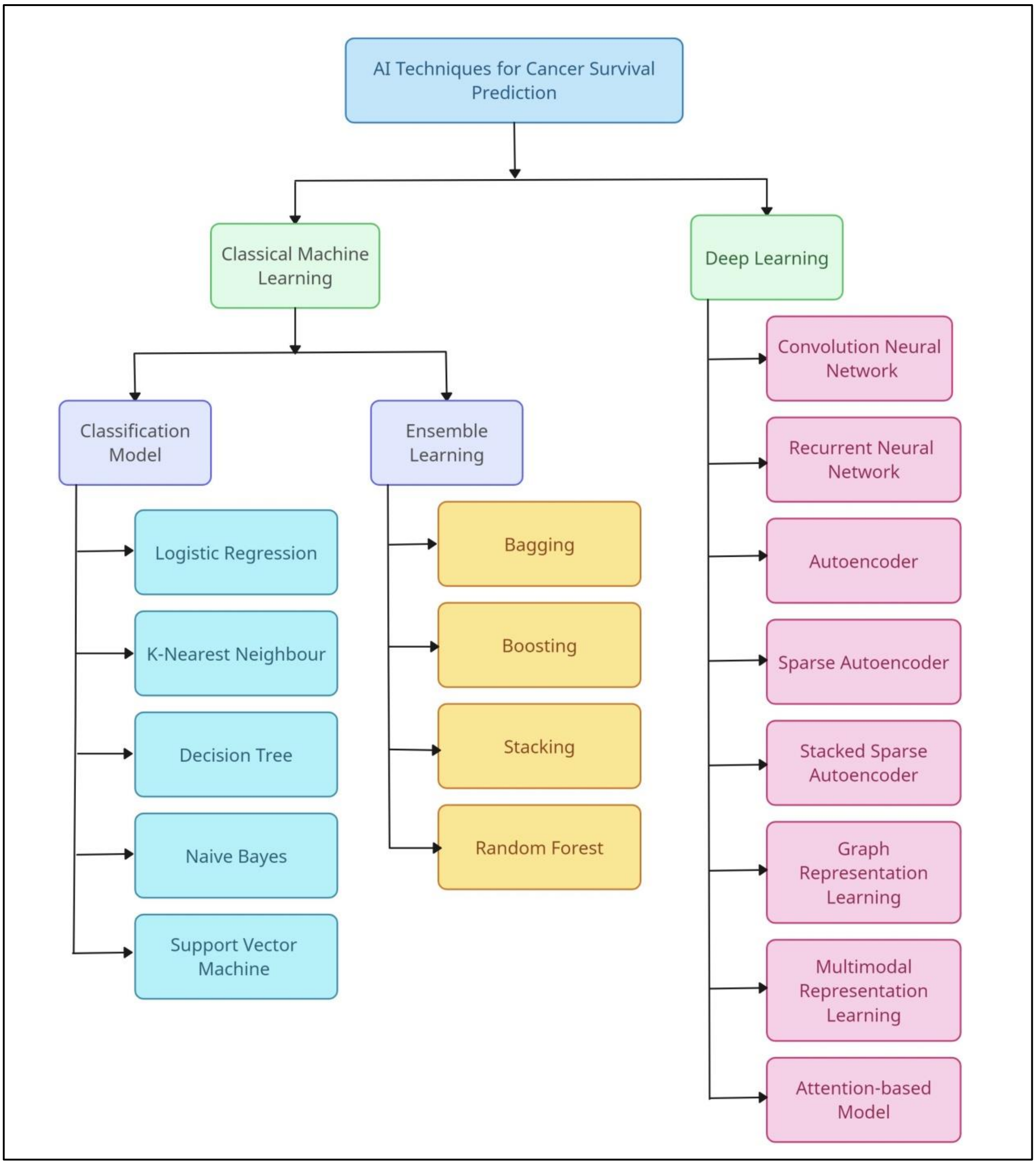


Fig 2.1 Data Analytics for Cancer Survival Prediction

In summary, AI techniques empower cancer researchers and clinicians to navigate the complexities of cancer data effectively. This, in turn, leads to more precise predictions, advancing both cancer research and patient care. AI's ability to seamlessly integrate data from various sources, capture intricate relationships, and provide personalized insights holds the potential to revolutionize cancer survival prediction, ultimately contributing to improved patient outcomes and better-informed clinical decisions.

2.1.1 Classical ML Techniques for Cancer Survival Prediction

In this section, we will explore classical machine learning techniques that have been applied to cancer survival prediction. These include Logistic Regression, K-Nearest Neighbors, Decision Trees, Naive Bayes, and Support Vector Machines. We will discuss their principles, strengths, and weaknesses in the context of cancer prognosis.

1. **Logistic Regression (LR)**

Logistic Regression is a straightforward way to predict outcomes that can be one of two options, like yes or no. When thinking about predicting if a cancer patient will survive, this method looks at different factors and calculates the chances of survival. It gives a result between 0 (meaning no chance of survival) and 1 (meaning sure survival) [19].

2. **K-Nearest Neighbors (KNN)**

KNN is like asking a group of neighbors for advice. For predicting cancer survival, it looks at similar patients and checks their survival outcomes. If most of these 'neighbor' patients survived, then the patient in question is likely to survive too. It's like taking a vote from the nearest similar cases [20].

3. **Decision Tree (DT)**

Decision Trees are like flowcharts that help in making decisions. For predicting cancer survival, they use information about the patient to guide through a series of questions. Each step narrows down the likely outcome. At the end of these questions, the tree gives a prediction about the patient's survival. It's a step-by-step guide based on the patient's details and is helpful in understanding which factors are crucial for survival [21].

4. **Naive Bayes (NB)**

NB is a classification algorithm that leverages Bayes' theorem and probabilistic principles for predicting class probabilities. Despite its "naive" assumption of feature independence, it has shown efficacy in various applications, including cancer survival prediction. In this context, NB calculates the posterior probability of a patient's survival status given the observed features. It is particularly useful when dealing with high-dimensional datasets and offers computational efficiency for large-scale studies [22].

5. Support Vector Machine (SVM)

SVM is a method used to classify data, especially when the data is a bit complicated. For predicting cancer survival, SVM tries to draw the best boundary line (or plane) that separates patients with different outcomes. This way, based on where new patient data falls, we can predict their survival chances. It's especially good when we have lots of data factors to consider [23].

2.1.2 Ensemble Learning Techniques for Cancer Survival Prediction

Ensemble learning techniques are powerful approaches that combine multiple models to improve predictive performance and generalization. These methods combine individual model predictions to produce a more robust and accurate final prediction. In the context of cancer survival prediction, these techniques play a crucial role in enhancing the strengths of diverse models to achieve more accurate and robust results [24]. We will delve into Bagging, Boosting, Stacking, and Random Forest, highlighting how these methods can improve the accuracy and robustness of survival prediction models.

1. Bagging

Bagging is like asking multiple friends for advice and then making a decision based on the majority opinion. In the world of cancer predictions, Bagging uses multiple versions of a model to look at different parts of the data. After every model has made its guess, the final decision is made by either seeing which guess is the most common (for categorizing) or by averaging all the guesses (for number predictions) [25].

2. Boosting

Boosting is an iterative technique that enhances the performance of weak learners to create a powerful predictive model. In the context of cancer survival prediction, Boosting starts by training a weak learner on the entire training dataset. Subsequent weak learners are trained on the instances that were misclassified by previous learners, assigning higher weights to these instances to emphasize their importance. The ultimate prediction is derived by aggregating the weighted predictions from all learners [26].

3. Stacking

Imagine you have several different experts, and each gives you advice. Stacking is like taking all their advice and then asking a super-expert to make the final decision based on that combined advice. For predicting cancer survival, several models first give their predictions. Then, another model, called the meta-model, looks at all these predictions to give a final answer. This way, we get the best bits from all models [27].

4. Random Forest (RF)

A Random Forest is like gathering opinions from a crowd, where everyone has seen only a part of the whole picture. For cancer predictions, it means building lots of decision trees, each looking at a different set of data. After all the trees have made their guesses, the final answer is either the most common guess (for categorizing) or an average of all guesses (for number predictions). This approach is good because it considers many different opinions and can highlight which factors are most important [28].

Table 2.1 offers a detailed summary that captures the variety and development of machine learning techniques in the context of cancer survival prediction. This table methodically outlines significant research in the area, detailing the year of publication, the researchers involved, the variety of cancer types studied, the datasets employed, the specific machine learning methods used, and their key discoveries.

Table 2.1: Overview of Classical ML Techniques in Cancer Survival Prediction

Year of Study	Authors	Cancer Types	Datasets	ML Techniques used and Important Finding
2023	Bozkurt et al. [29]	Breast, Lung, Prostate, Stomach	Medical Information Mart in Intensive Care IV (MIMIC-IV)	• Classification-based approach using multiple classifier, Logistic Regression for feature selection • Using fewer features is efficient
2023	Arya et al. [30]	Breast	The Cancer Genome Atlas Breast Invasive Carcinoma Collection (TCGA-BRCA)	• Principal Component Analysis, Variational Autoencoders, Support Vector Machine, Random Forest • More modalities = More robustness
2023	Zolfaghari et al. [31]	Multiple Cancer Types	Not Specified	• Ensemble Classifiers incorporating deep learning • Review on ensemble methods used in cancer prognosis and diagnosis
2022	Azar et al. [32]	Ovarian	Surveillance, Epidemiology, and End Results Database (SEER)	• K-Nearest Neighbour, Support Vector Machine, Decision Tree, Random Forest, AdaBoost, XGBoost, Sharpley Additive Explanations (SHAP) • Random Forest and XGBoost achieved the best performance for classification and regression respectively

Year of Study	Authors	Cancer Types	Datasets	ML Techniques used and Important Finding
2022	Yan et al. [33]	Gastric, Skin	Surveillance, Epidemiology, and End Results Database (SEER)	• Priori knowledge and stability-based feature selection (PKSFS), two-stage heterogeneous stacked ensemble learning model (BQAXR) • PKSFS performed well in processing high-dimensional datasets. BQAXR outperformed mainstream ML
2022	S.P. et al. [34]	Lung	CT Images	• Logistic Regression-based models within a Quantitative Radiomic Framework • Use of advanced imaging techniques for predicting lung cancer patient survival
2018	Bartholomai et al. [35]	Lung	Surveillance, Epidemiology, and End Results Database (SEER)	• ANOVA for factor selection, Random Forest (for classification and regression), Linear Regression, Gradient Boosted Machines (GBM) • Random Forest performed best for survival times ≤6 and >24 months, while Gradient Boosted Machines performed best for 7–24 months

2.1.3 Deep Learning Techniques for Cancer Survival Prediction

Here, we will explore the application of deep learning techniques in cancer survival prediction. We'll discuss Convolutional Neural Networks (CNNs) for medical imaging analysis, Recurrent Neural Networks (RNNs) for sequential data, Autoencoder for feature learning, Graph Representation Learning (GRL) for complex data relationships, and Multimodal Representation Learning (MRL) for integrating diverse data types, and Attention Models for identifying critical features. We'll highlight how these advanced methods can capture intricate patterns and relationships in cancer data, potentially revolutionizing survival prediction.

1. **Convolutional Neural Network (CNN)**

CNN is a type of deep learning mostly used for analyzing images pictures. For predicting cancer survival, we can use CNNs to study medical images, like MRI or CT scans, to see details that might be hard for the human eye to catch [36]. CNNs work by having many layers that process the image, looking for patterns and important features. This helps doctors get a clearer picture of tumors and predict patient outcomes [37].

2. **Recurrent Neural Network (RNN)**

RNN is a type of deep learning architecture suitable for sequential data analysis, making it applicable to time-series data like patient health records. In cancer survival prediction, RNNs can model the temporal dependencies between clinical events and patient outcomes. The main strength of RNNs lies in their ability to capture long-range dependencies in sequential data, allowing them to consider a patient's entire medical history for survival prediction [38].

3. **Autoencoder**

Autoencoders are tools that simplify complex data. In cancer predictions, they can be used to find the most important information from big datasets, like genomic data. They work by taking in the data, shrinking it down to capture the main points, and then expanding it back out. This process helps in making the data more manageable and can improve how well we predict cancer survival [39].

4. Sparse Autoencoder

Sparse autoencoders are a variant of autoencoder that introduces sparsity constraints to the hidden layer activations. The sparsity constraint encourages only a small subset of the neurons in the hidden layer to be active, resulting in a more concise and interpretable representation of the input data. In cancer survival prediction, sparse autoencoders can be applied to identify critical genomic features or biomarkers associated with patient outcomes [40]**.**

5. Stacked Sparse Autoencoder

Stacked sparse autoencoders combine multiple layers of sparse autoencoders to create a deep architecture. Each layer learns increasingly abstract and higher-level features, leading to a hierarchical representation of the input data. Stacked sparse autoencoders excel at capturing intricate patterns and complex relationships in cancer-related data, enabling accurate and detailed survival predictions [41].

6. Graph Representation Learning (GRL) for Cancer Survival Prediction

The GRL technique has been gaining attention in recent years for its potential in cancer survival prediction. It involves the use of graphs to represent data, with each node representing a data point and the edges between them representing relationships or interactions between those data points [42]. The goal is to learn a low-dimensional representation of the graph that captures its underlying structure and patterns. GRL has several advantages over traditional methods, particularly in cases where the data is complex and heterogeneous, as is often the case with cancer data. It can effectively capture the interactions between different types of data, such as genomic, clinical, and imaging data, and identify hidden relationships that may not be evident through other methods [43].

For example, in a study on computational histopathology, graph convolutional networks (GCNs) were used to analyze digital pathology images [44]. The whole slide image (WSI) was represented as a graph, with each cell or region in the image represented as a node and the edges representing spatial relationships between the nodes. By leveraging the power of GCNs, the study aimed to capture more complex spatial relationships between different regions of the image and incorporate information about the local and global context of each region, potentially improving the accuracy of the classification task and making the method more robust to variations in the size, shape, and position of the regions of interest.

7. Multimodal Representation Learning (MRL) for Cancer Survival Prediction

MRL is a technique that involves integrating multiple types of data, such as genomics, imaging, and clinical data, to improve cancer survival prediction. This approach has become increasingly popular in recent years, allowing for a more comprehensive and holistic view of the patient's condition [45]. The benefit of MRL is that it can leverage the complementary information from different data types to make more accurate predictions and also discover novel relationships between different data types [46]. For example, one study used multimodal graph neural networks (MGNN) to integrate gene expression, copy number alteration and clinical data to predict breast cancer survival. The GNN were able to capture the complex relationships between the different data modalities by constructing a bipartite graph between patient and multimodal data, leading to improved survival prediction accuracy [47]. Overall, we can say that this technique has shown promising results in several studies, indicating that it can significantly improve the accuracy of cancer survival prediction compared to models that only use a single data type.

8. Attention Model for Cancer Survival Prediction

Survival-based attention models are a type of attention model that is used in cancer survival prediction tasks. These models utilize attention mechanisms to highlight relevant features from input data that are most informative for predicting survival outcomes. The attention weights are assigned to each feature based on their contribution to the survival outcome. By using attention mechanisms, survival-based attention models can identify important features that may be missed by traditional models, thus potentially improving the accuracy of cancer survival prediction [48].

Table 2.2 offers a detailed summary that captures the variety and development of deep learning techniques in the context of cancer survival prediction. This table methodically outlines significant research in the area, detailing the year of publication, the researchers involved, the variety of cancer types studied, the datasets employed, the specific deep learning methods used, and their key discoveries.

Table 2.2: Overview of DL Techniques in Cancer Survival Prediction

Year of Study	Authors	Cancer Types	Datasets	DL Techniques used and Important Finding
2023	Wu et al. [49]	Multiple (including Breast, Lung and Brain)	Three cancer datasets obtained from TCGA including LGG, Breast Invasive Carcinoma (BRCA) and LUSC	• Cross-Aligned Multimodal Representation Learning (CAMR) • CAMR effectively reduces modality gaps, generating both modality-invariant and -specific representations for enhanced cancer survival prediction
2023	Fan et al. [50]	Pancancer	TGCA, University of California Santa Cruz Xena (UCSC Xena) (including clinical data, gene expression (mRNA) data, microRNA expression (miRNA) data and gene copy number variation (CNV) data)	• Multimodal integrative DL with unsupervised representation learning • Model performs best using clinical and mRNA modalities, Adding more data modalities risks overfitting, Including the CNV modality reduced prediction performance, potentially due to introducing noise
2022	Arya et al. [51]	Breast	Multi-modal datasets (gene expression profile, copy number alteration, clinical data)	• Stacked-based ensemble model architecture • CNN for feature extraction then stacked-based approach utilizing three modalities of data improves predictive performance especially for imbalanced datasets

Year of Study	Authors	Cancer Types	Datasets	DL Techniques used and Important Finding
2022	Kanwal et al. [52]	Multiple (including Brain, Prostate, Bladder, Colorectal and Breast)	Lower Grade Glioma in the Brain (BRAIN-TCGA), Prostate Cancer Dataset, Bladder Cancer Dataset (BLADDER-TCGA), Metastatic Colorectal Cancer Dataset (MSKCC), and Molecular Taxonomy of Breast Cancer International Consortium (METABRIC)	• Artificial Algae Algorithm (AAA) for feature extraction combined with Double DEEP Q-NETWORK (DDQN), Convolution eXtreme Gradient Boosting (CNN-XGBOOST), and Convolution Support Vector Machine (CNN-SVM) • A novel framework was introduced that integrates DL/ML and RL with AAA for improved cancer prognosis prediction using multimodal data, incorporating early and late fusion techniques
2022	Li et al. [53]	Colorectal	Two international Colorectal Cancer (CRC) datasets: MCO CRC and TCGA COAD-READ	• Distribution-based Multiple-Instance Survival Learning algorithm (DeepDisMISL) • Combining percentile-scored patches with highest and lowest scored ones , Including neighborhood instances around percentiles further boost prediction accuracy

Year of Study	**Authors**	**Cancer Types**	**Datasets**	**DL Techniques used and Important Finding**
2022	Wu et al. [54]	Glioblastoma multiforme (GBM), Ovarian serous cystadenocarcinoma (OV), Breast invasive carcinoma (BRCA)	Gene expression, Copy number variations (CNV), Clinical information for GBM, OV, and BRCA datasets from The Cancer Genome Atlas (TCGA) project via UCSC Xena	• Stacked Autoencoder-based Survival Prediction Neural Network (SAEsurv-net) • SAEsurv-net outperforms single-data-type models and other state-of-the-art methods, Effective handling of multi-omics heterogeneity and dimensionality reduction
2019	Cheerla et al. [55]	Obtained from TCGA (including 20 different cancer types)	Multimodal dataset (including clinical, genomic and Whole Slide Image (WSIs))	• Developed a DL survival model with multimodal representation • Demonstrated efficient use of multimodal data, even with missing modalities, Proposed efficient While Slide Image analysis by sampling key Region of Interests.
2018	Sun et al. [56]	Breast	Multi-modal datasets (gene expression profile, copy number alteration, clinical data)	• Multimodal Deep Neural Network by integrating Multi-dimensional Data (MDNNMD) • MDNNMD integrates multi-dimensional data for better prognosis prediction; outperforms single-dimensional methods

2.2 Summary

This chapter began with an in-depth review of the methods central to cancer survival prediction, highlighting the progression and merging of conventional statistical tools with modern AI-driven strategies. In the past, the field of oncology heavily depended on traditional statistical models for survival analysis. These models, be it parametric with their distinct risk function presumptions, non-parametric with their adaptable stance, or semi-parametric blending both, supplied well-organized and logically based structures to understand the details of survival data. Illustrations, such as diagrams showcasing parametric, non-parametric, and semi-parametric models, added depth to our grasp of these methods.

However, with technological advancements and the surge in data diversity, the chapter shifted focus to the groundbreaking domain of Artificial Intelligence (AI). AI's rise marked a significant change, introducing methods capable of proficiently managing the varied and layered data present in contemporary oncology studies. By effectively merging different data sources and revealing concealed trends, machine learning and deep learning models, as described, demonstrated their potential to reshape cancer survival estimations.

In summary, this chapter highlights a pivotal transition in the world of cancer survival prediction—a shift from the foundational precision of statistical tools to the innovative prowess of AI. The collaborative synergy between these techniques suggests a future where cancer predictions are not just more precise but also deeply tailored, making certain that every patient's individual journey is central to their treatment.

CHAPTER 3: DESIGN OF CANCER SURVIVAL PREDICTION SYSTEM

In this chapter, we outline the design and methodology of our cancer survival prediction system. We introduce the proposed architecture, emphasizing its capability to process multimodal datasets and the significance of understanding the data intricacies. The focus then shifts to a detailed exploratory data analysis of both clinical and genomic datasets. Subsequently, we delve into the Gated Attention CNN model, discussing its design, core principles, and its relevance for our data. The chapter concludes with a breakdown of our architecture's key parameters, ensuring transparency and facilitating future replication. Overall, this chapter establishes the foundational techniques and methods vital for subsequent chapters.

3.1 Cancer Survival Prediction System Using Multimodal Data Analytics

As depicted in Figure 3.1, the proposed architecture of our cancer survival prediction system offers a systematic and structured approach to harnessing the potential of multimodal datasets. The architecture is meticulously designed to ensure a seamless flow from one phase to another, ensuring the coherent integration of both clinical and genomic data.

The initial phase, Data Acquisition, is crucial as it involves the collection and organization of data from the METABRIC dataset. Recognized for its extensive and varied information on breast cancer, this dataset serves as a rich resource, providing both clinical and genomic data that forms the foundation of our analysis.

Following data acquisition, the Exploratory Data Analysis (EDA) phase comes into play. This phase is dedicated to understanding the intricacies of the acquired data, identifying patterns, anomalies, and potential hypotheses. It emphasizes the importance of addressing missing values, understanding the distribution of clinical attributes, and gaining insights into the molecular dynamics of disease progression.

With a clear understanding of the data, we transition to the Modeling phase. Here, the Gated Attention Convolutional Neural Network (CNN) is employed to interpret the data's vast and varied landscape. Recognizing the heterogeneity inherent in the datasets, three distinct models are constructed for clinical data, gene expression, and copy number alteration. This approach ensures that each data type's unique challenges and opportunities are addressed, leading to a more holistic and comprehensive analysis.

The architecture culminates in the Result & Analysis phase. This phase is not just about obtaining outputs from the models but also involves a critical evaluation of these results. A comparative analysis is undertaken, benchmarking the performance of the Gated Attention CNNs against other established classifiers. This provides a relative measure of efficacy and guides future research directions.

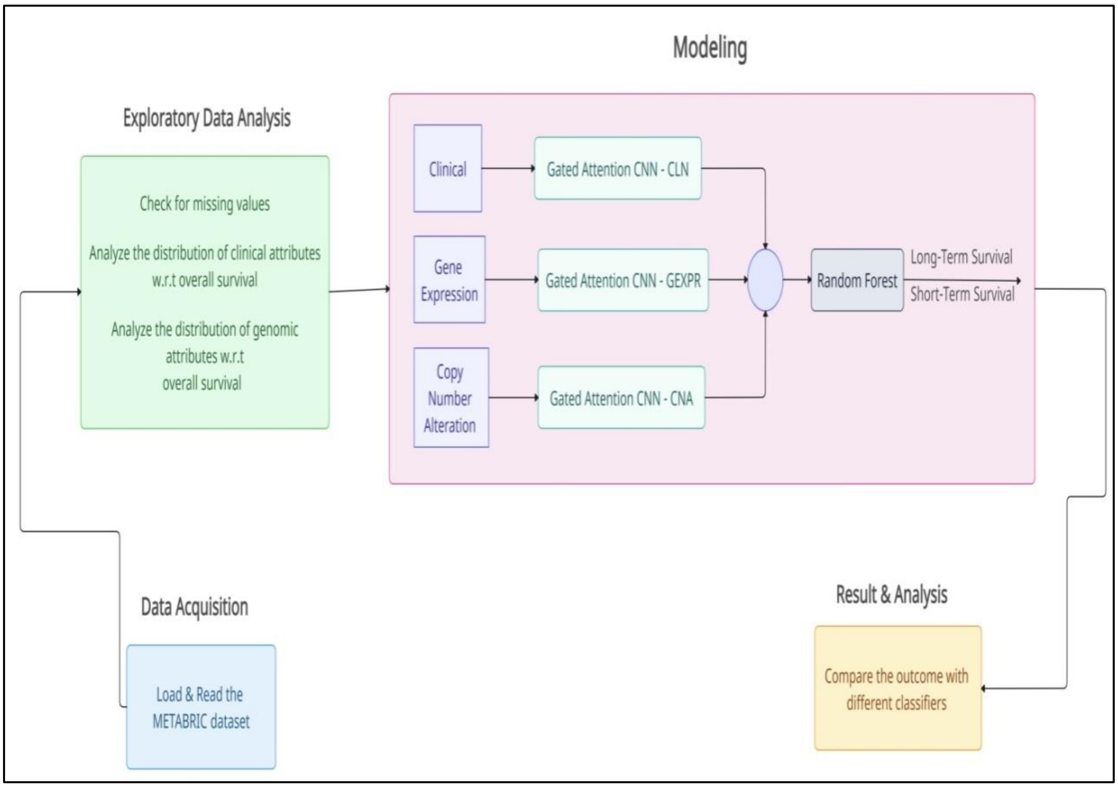

Modeling
Exploratory Data Analysis
Check for missing values
Analyze the distribution of clinical attributes w.r.t overall survival
Analyze the distribution of genomic attributes w.r.t overall survival
Clinical
Gated Attention CNN - CLN
Gene Expression
Gated Attention CNN - GEXPR
Copy Number Alteration
Gated Attention CNN - CNA
Random Forest
Long-Term Survival
Short-Term Survival
Data Acquisition
Load & Read the METABRIC dataset
Result & Analysis
Compare the outcome with different classifiers

Fig 3.1 Architecture for Cancer Survival Prediction using Multimodal Dataset

The process of data acquisition is a foundational step in any scientific investigation. In the realm of biomedical research, data acquisition often refers to the gathering of relevant information, both clinical and genomic, to address specific research questions. The METABRIC dataset, which stands for the Molecular Taxonomy of Breast Cancer International Consortium, is a crucial example in breast cancer research. It offers a comprehensive collection of data that encompasses an abundance of clinical attributes and genomic markers. Historically, this dataset has been pivotal in illuminating the heterogeneous nature of breast cancer, drawing attention to its varied subtypes, each with distinct molecular signatures, clinical trajectories, and therapeutic responses. The inclusion of such a dataset in research not only amplifies the robustness of the study but also ensures that the findings are grounded in a well-characterized and validated source of information. Loading and reading the METABRIC dataset, therefore, is not merely a procedural step, but a conscious choice to leverage a dataset that has been the foundation of numerous transformative insights in breast cancer research.

Exploratory Data Analysis (EDA) stands as an indispensable phase in the data analysis pipeline, acting as the bridge between raw data acquisition and intricate modeling. EDA's primary objective is to understand the nuances of the dataset, detect patterns, identify anomalies, and ascertain potential hypotheses for more advanced analytical tasks. Within the framework of the METABRIC dataset, the initial challenge lies in addressing missing values. In biomedical datasets, missingness is not a mere inconvenience but can signify deeper underlying issues – from procedural lapses in data collection to systematic biases that can skew results. Identifying and wisely addressing these missing values is paramount to ensure the integrity of subsequent analyses. Beyond the mechanics of data cleaning, EDA dives deeper into the heart of the dataset. Analyzing the distribution of clinical attributes, especially with respect to overall survival in the context of breast cancer, can shed light on prognostic markers, hinting at variables that might play a pivotal role in patient outcomes. Similarly, understanding the distribution of genomic attributes offers a window into the molecular foundations of disease progression. In essence, EDA for the METABRIC dataset is a meticulous process of unraveling the complex network of clinical trajectories and genomic landscapes, setting the stage for more focused and hypothesis-driven analyses.

With a clear understanding of the data at hand, the next logical step in the analytical journey is modeling. The choice of the Gated Attention Convolutional Neural Network (CNN) signifies a deliberate move towards leveraging the power of deep learning for this investigation. CNNs, traditionally renowned for their expertise in image analysis, have found increasing applicability in genomics and biomedicine, primarily due to their ability to handle complex patterns in high-dimensional data. The 'gated attention' mechanism adds another layer of sophistication to this model. Attention mechanisms, inspired by human cognitive processes, allow models to dynamically focus on different parts of the input data, lending them an adaptive edge in detecting intricate patterns. In the context of the METABRIC dataset, three distinct models or branches seem to be employed: one for Clinical data (CLN), one for Gene Expression data (GEXPR), and another for Copy Number Alteration data (CNA). This trio approach acknowledges the inherent heterogeneity in the data sources, allowing for specialized models that can capture the nuances of each data type. While clinical data might offer insights into patient demographics, treatment histories, and outcomes, gene expression and copy number alterations shed light on the molecular dynamics driving the disease. By designing separate branches for each data type, the architecture ensures a focused analysis, optimized for the unique challenges and opportunities each dataset presents.

The integration of the Random Forest classifier in our architecture plays a pivotal role in the data analysis process, enhancing the model's predictive capacity. After the detailed processing of clinical, gene expression, and copy number alteration data through individual Gated Attention CNN models, the extracted hidden features captures the intricate complexities of the dataset. This feature, once merged into a single dataset provides a rich, unified dataset that encapsulates the multifaceted nature of the information at hand. The ensemble method employed by Random Forest, leveraging the collective insights of numerous decision trees, works in synergy to uncover hidden patterns that might elude simpler models. Recognized for its resilience and proficiency in managing vast, dimensional data, Random Forest excels in interpreting these complex feature spaces. When fed with the combined feature set, the Random Forest classifier acts as an aggregator of insights, identifying and leveraging the most informative cues from the comprehensive feature pool. By applying this model, we aim to capture the collective predictive power embedded within the multimodal data, effectively synthesizing the diverse signals into a cohesive prediction of cancer survival outcomes.

The importance of using Random Forest in our model cannot be overstated. It offers a methodological advantage by mitigating the risk of overfitting—a common pitfall in machine learning where models may perform well on training data but fail to generalize to new data. The algorithm achieves this by bootstrapping the data and using the aggregate decision-making process across the forest, thus enhancing the model's generalizability. Moreover, Random Forest provides an intrinsic ability to rank the importance of features, granting us invaluable insights into which attributes are most predictive of survival outcomes. This not only informs the medical community of potential biomarkers but also sheds light on the underlying biological processes driving the progression of cancer. By integrating these insights into our model, we aim not only to predict survival outcomes with greater accuracy but also to contribute to the broader understanding of cancer as a complex, multifaceted disease.

Once the models are trained and validated, the analytical process transitions to a phase of introspection and comparison in the results and analysis stage. It's not enough for a model to merely produce outputs; it is imperative for researchers to understand, interpret, and critically evaluate these results. In the context of this investigation, a comparative analysis approach seems to be employed, comparing the performance of the Gated Attention CNNs against other classifiers. This comparative paradigm serves multiple purposes. Firstly, it acts as a benchmarking tool, placing the performance of the proposed models in the context of established algorithms, thereby offering a relative measure of efficacy. Such comparisons can clarify the advantages and potential limitations of the proposed architecture, guiding future refinements. Secondly, in the broader scientific discourse, comparative analyses support the credibility of the study. By demonstrating the superiority, or at least the competitiveness, of the proposed models against established benchmarks, the research stakes its claim in the scientific community, inviting discussions, critiques, and collaborations. In essence, the results and analysis phase is not just an endpoint but a gateway, transitioning the research from an isolated investigation to a dialogue within the global scientific community.

3.2 Exploratory Data Analysis of Clinical Data

Exploratory Data Analysis (EDA) is a fundamental step in the data processing pipeline, especially in biomedical research. EDA involves visualizing, summarizing, and interpreting the information that a dataset contains, all while ensuring that patterns, relationships, anomalies, or any other insightful details are not overlooked. It serves as the foundation upon which further analytical procedures, hypothesis testing, and modeling efforts are built.

The significance of EDA, particularly in the context of clinical data, lies in its ability to provide a holistic view of the data. Clinical datasets, due to their inherent complexity and multifaceted nature, often encapsulate a myriad of variables that can influence patient outcomes. By performing EDA, researchers can gain preliminary insights into these variables, understand their distributions, identify potential correlations or outliers, and, most importantly, formulate pertinent research questions or hypotheses to drive subsequent analyses.

In the scope of this chapter, we embark on a detailed EDA of various clinical attributes, each of which holds potential implications for cancer survival prediction. The attributes under scrutiny include:

1. Age at diagnosis
2. Lymph nodes examined positive
3. Mutation count
4. Nottingham prognostic index
5. Overall survival months
6. Tumor size
7. Tumor size distribution per tumor stage
8. Distribution of histopathological class with respect to survival
9. Attributes showing weak positive correlation with overall survival
10. Attributes showing weak negative correlation with overall survival

Each of these attributes, given their clinical relevance, will be meticulously analyzed to shed light on their respective distributions and relationships with survival outcomes. In the sections that follow, we will dive deep into the patterns and insights gleaned from the EDA of each of these attributes.

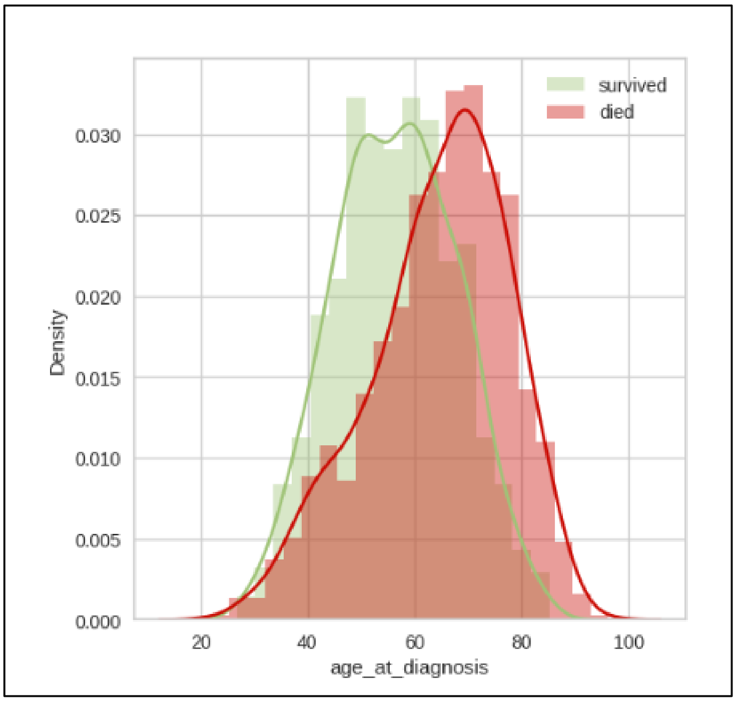


Fig 3.2 EDA performed on age at diagnosis

As depicted in Figure 3.2, the age at diagnosis provides significant insights into the distribution of patient ages at the time of their breast cancer diagnosis. Several patterns emerge from this visualization:

- **Bimodal Distribution for Survivors:** The age distribution for patients who survived showcases a bimodal trend, indicating two prominent peaks. This suggests that there are two age groups among survivors where the incidence of breast cancer diagnosis is comparatively higher.
- **Left-skewed Distribution for Non-Survivors:** For the patients who unfortunately did not survive, the age distribution is left-skewed. This skewness indicates that a majority of the patients who succumbed to the disease were diagnosed at older ages. The left tail, representing younger ages, has fewer patients, suggesting that younger individuals are less likely to face mortality due to the disease, at least within the timeframe captured by the dataset.

These observations, derived from Figure 3.2, underscore the potential influence of age at diagnosis on patient outcomes. The age at which a patient is diagnosed with breast cancer may have implications for their survival trajectory, making it an essential variable for further analytical exploration.

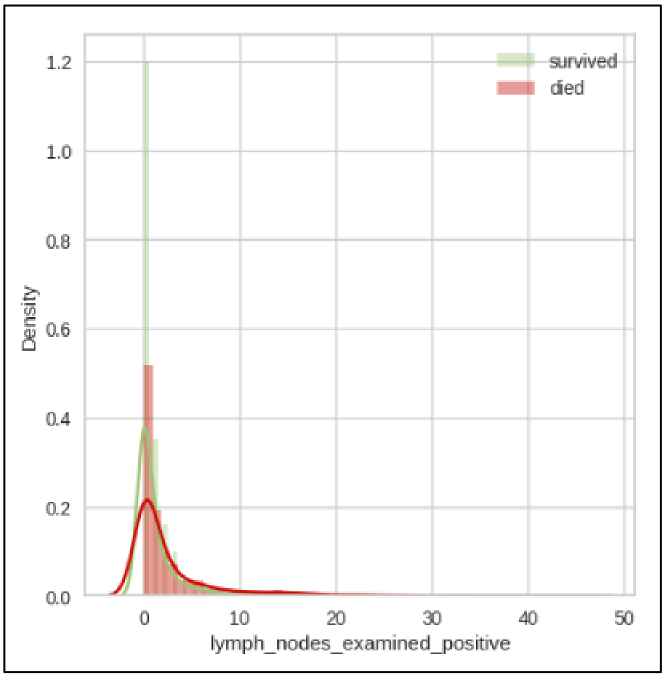


Fig 3.3 EDA performed on lymph nodes examined positive

As illustrated in Figure 3.3, the distribution of the number of positive lymph nodes examined offers critical insights into the breast cancer patients' clinical profiles. Several distinct patterns can be discerned from this visual representation:

- **Right-skewed Distribution for Both Survivors and Non-Survivors:** Both the surviving and non-surviving patient groups exhibit a right-skewed distribution regarding the number of positive lymph nodes examined. This skewness indicates that a majority of patients, irrespective of their survival outcome, have a lower count of positive lymph nodes.
- **Presence of Extreme High Values:** The rightward pull of the distribution suggests the existence of a subset of patients with an exceptionally high number of positive lymph nodes. These extreme values, though fewer in number, can be particularly informative as they might represent more aggressive or advanced stages of the disease.

The number of positive lymph nodes examined is a crucial clinical metric, often used to assess the disease's spread and severity. A higher count typically indicates a more widespread presence of cancer cells, which can have implications for prognosis, treatment choices, and overall patient outcomes. The patterns observed in Figure 4 emphasize the need for a detailed exploration of this variable's relationship with survival rates and its potential incorporation into predictive models.

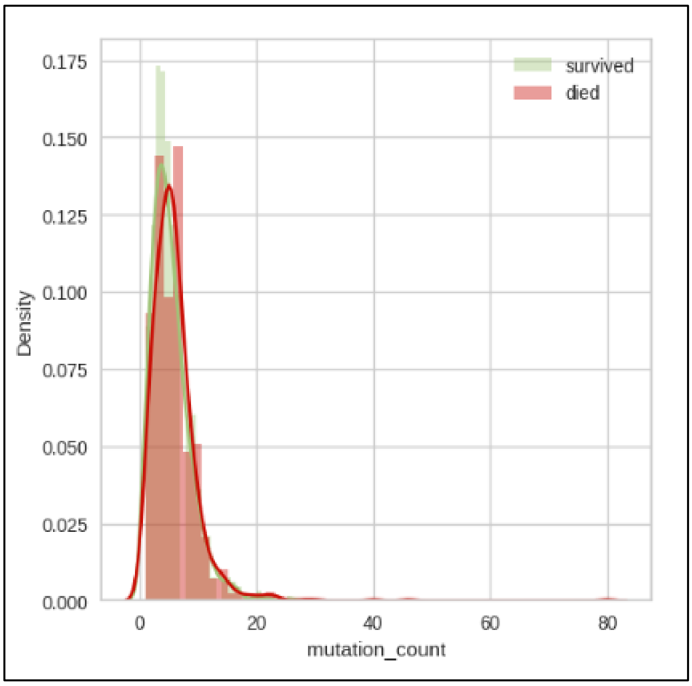


Fig 3.4 EDA performed on mutation count

Figure 3.4 visually captures the distribution of mutation counts among breast cancer patients, providing pivotal insights into the genetic variations observed within the patient cohort.

- **Right-skewed Distribution across both Groups:** The data visualization for both surviving and non-surviving patients distinctly portrays a right-skewed distribution concerning the mutation count. This suggests that a predominant segment of patients, regardless of their eventual survival outcome, presents with a relatively low count of mutations in their genomic profiles.
- **Existence of Patients with High Mutation Counts:** The tail of the distribution extending to the right indicates the presence of patients with exceptionally high mutation counts. Even though they constitute a minority, these patients potentially reflect more aggressive or genetically complex forms of the disease.

Mutation count serves as an essential indicator in oncological research, reflecting the extent of genetic variations within a tumor. A higher mutation count might be linked to increased tumor heterogeneity, potentially influencing treatment responses and prognostic outcomes. The mutation count, therefore, serves as a vital indicator of the genetic landscape of the tumor. The patterns discerned from Figure 5 underline the significance of understanding the mutation profile and its nuanced relationship with patient outcomes, further emphasizing its relevance in predictive modeling and therapeutic strategies.

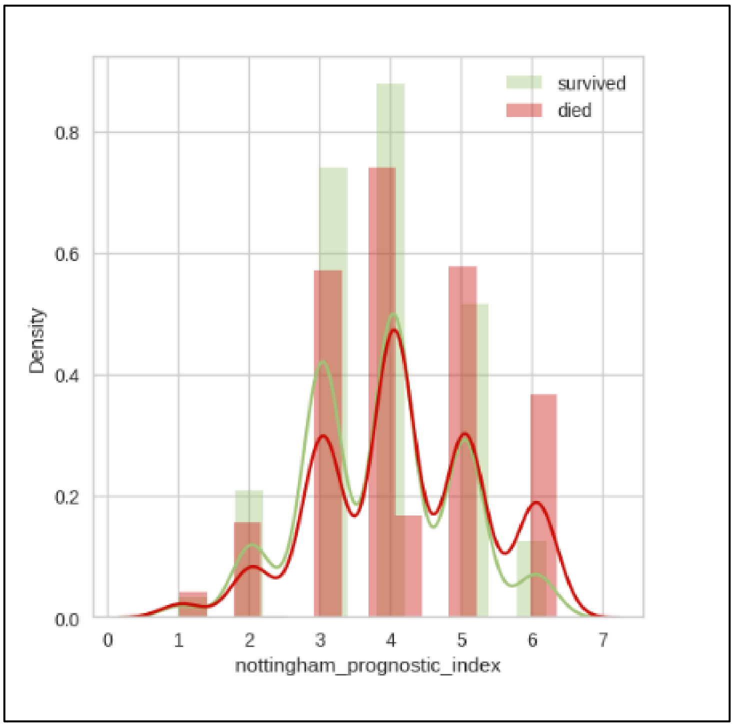


Fig 3.5 EDA performed on Nottingham prognostic index

The visual representation in Figure 3.5 showcases the distribution of the Nottingham Prognostic Index (NPI) across the patient cohort, revealing certain patterns and insights:

- **Distinct Multimodal Distribution:** Both the survivors and non-survivors display a multimodal distribution for NPI values. Multimodal distributions, characterized by multiple peaks or clusters, hint at a stratified nature of the data. In the case of NPI, this might correspond to different risk categories or groups within the patient population.
- **Implications for Risk Stratification:** The presence of multiple modes suggests that patients can be segregated into distinct risk groups based on their NPI values. Each of these groups or clusters may be associated with a unique prognosis, underlining the utility of NPI as a stratification tool in clinical practice.

The Nottingham Prognostic Index is a renowned metric in breast cancer prognosis. It consolidates multiple clinical and pathological parameters to provide a composite score, aiding in patient risk stratification. The distribution observed in Figure 6 reaffirms the stratified nature of NPI and its role in guiding clinical decisions, tailoring treatments, and informing follow-up strategies for patients.

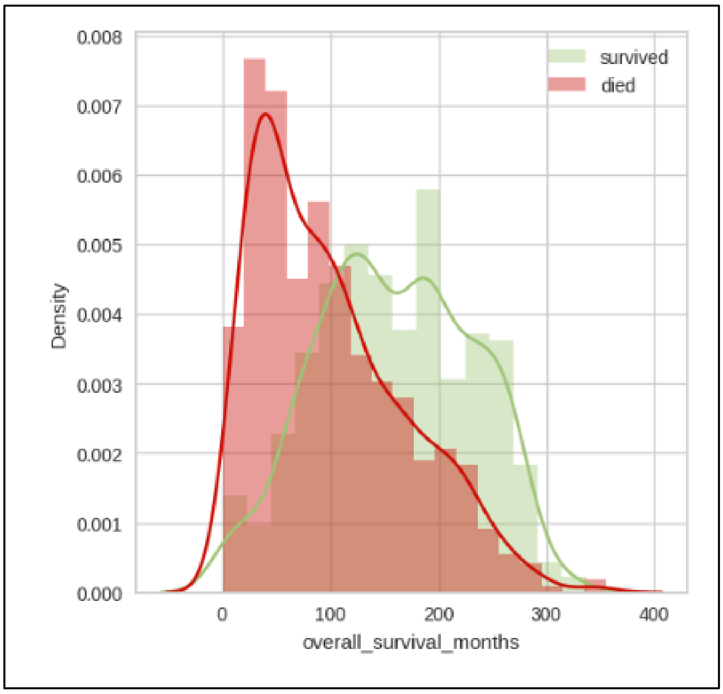


Fig 3.6 EDA performed on overall survival months

Illustrated in Figure 3.6, the distribution of overall survival months for the patient cohort brings forth several noteworthy observations:

- **Bimodal Distribution among Survivors:** The survival month's distribution for patients who survived presents a bimodal nature, marked by two distinct peaks. This suggests two predominant groups within the survivors, possibly corresponding to different durations of survival or related to specific treatment regimens, disease stages, or other clinical factors.
- **Right-Skewed Distribution among Non-Survivors:** For those who unfortunately died to the disease, the distribution leans towards a right-skewed shape. This indicates that a significant portion of the patients who died experienced a shorter duration of overall survival, with fewer patients having longer survival durations.

The overall survival month's metric is a crucial endpoint in oncology research, offering insights into the effectiveness of treatments, progression of the disease, and the overall impact of various clinical and genomic factors on patient outcomes. The patterns discerned from Figure 3.6 underscore the need to delve deeper into the factors influencing these distributions, aiming to improve therapeutic approaches and patient care strategies.

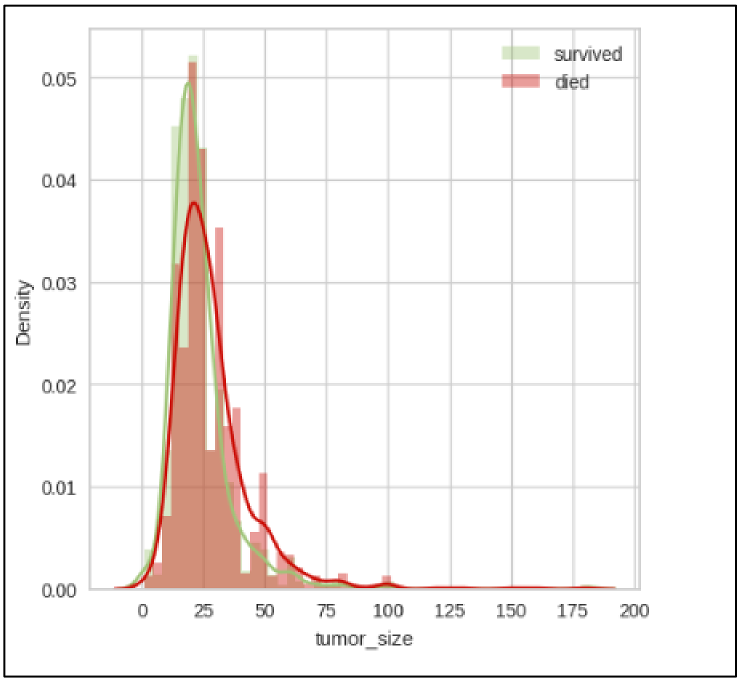


Fig 3.7 EDA performed on tumor size

Figure 3.7 visualizes the distribution of tumor sizes across the patient cohort, and several key observations can be drawn:

- **Right-Skewed Distribution:** Both the patients who survived and those who unfortunately passed away exhibit a right-skewed distribution for tumor sizes. This pattern indicates that a predominant portion of the patients in this study presented with smaller tumor sizes.
- **Implications of Smaller Tumor Sizes:** The prevalence of smaller tumors might be a testament to early detection or the natural progression of the specific type of cancer under study. Smaller tumor sizes are typically more manageable and have better prognostic outcomes, emphasizing the importance of early detection and timely interventions in cancer care.

Tumor size is a critical metric in oncology, often correlating with the stage of the disease, potential for metastasis, and overall prognosis. The insights obtain from Figure 3.7 underline the significance of understanding tumor growth dynamics and its implications for therapeutic strategies and patient outcomes.

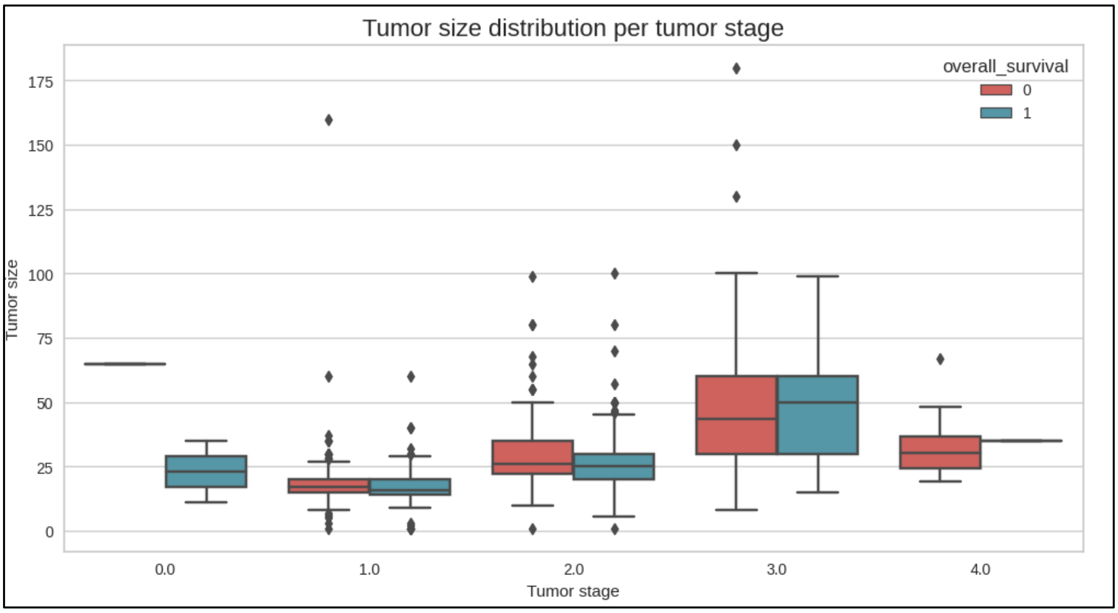


Fig 3.8 EDA performed on tumor size distribution per tumor stage

Figure 3.8 presents a comparative analysis of the tumor size distributions segmented by tumor stages and based on patient survival outcomes. Delving into the details of this visualization, we discern the following patterns:

- **Stage 0.0:**
  - **Died:** Patients in this category exhibit a concentration in tumor size, predominantly clustered around the 60-70 units mark.
  - **Survived:** These patients showcase a more varied tumor size spectrum, with a notable density between the 15 and 35 units range.
- **Stage 1.0:**
  - **Died and Survived:** The tumor size distributions for both groups in this stage are relatively congruent, with the exception of an outlier in the deceased group, where the tumor size exceeds 150 units.
- **Stage 2.0:**
  - **Died and Survived:** While the distributions show some variance, the median tumor sizes for both cohorts appear analogous, indicating consistent tumor progression characteristics within this stage.
- **Stage 3.0:**
  - **Died:** This group showcases a broad spectrum of tumor sizes, spanning from approximately 10 to a pronounced 175 units or more.

    - **Survived:** Patients in this cohort present with tumor sizes chiefly distributed between 20 and 100 units.
- **Stage 4.0:**
    - **Died:** The tumor sizes for this unfortunate group are chiefly found between 20 and 70 units.
    - **Survived:** These patients depict a more homogenized tumor size distribution, predominantly hovering around the 60-70 units range.

The insights derived from Figure 3.8 emphasize the nuanced variations in tumor size, contingent upon the tumor stages, and their corresponding implications on survival outcomes. The differential distributions across stages underline the significance of tumor size as a pivotal metric in oncological assessments. Recognizing these patterns and correlations augments the precision of prognostic evaluations and aids clinicians in tailoring treatment regimens aligned with the unique tumor-stage and size dynamics of individual patients.

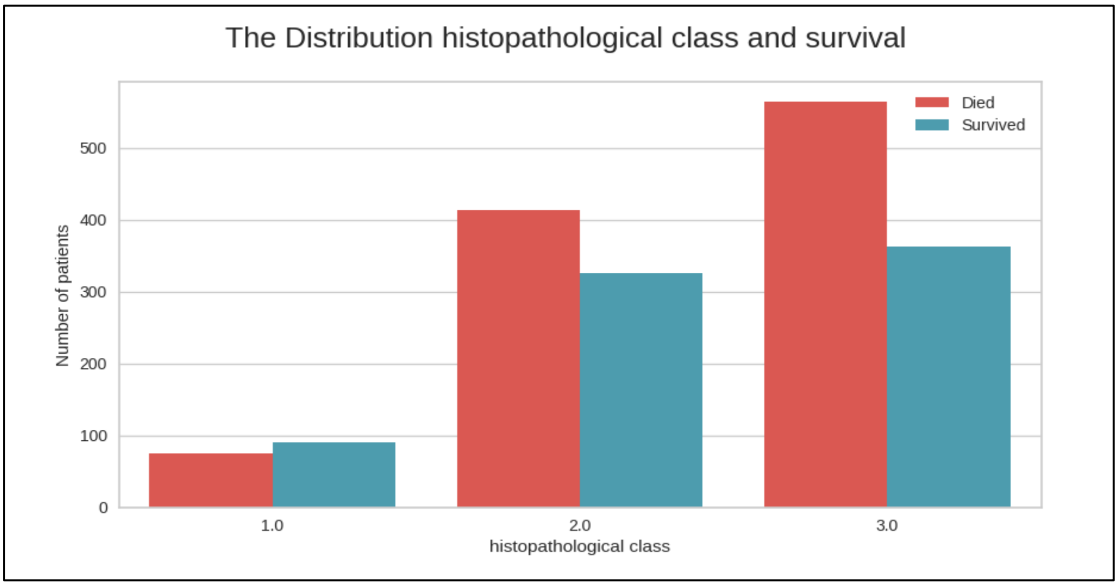


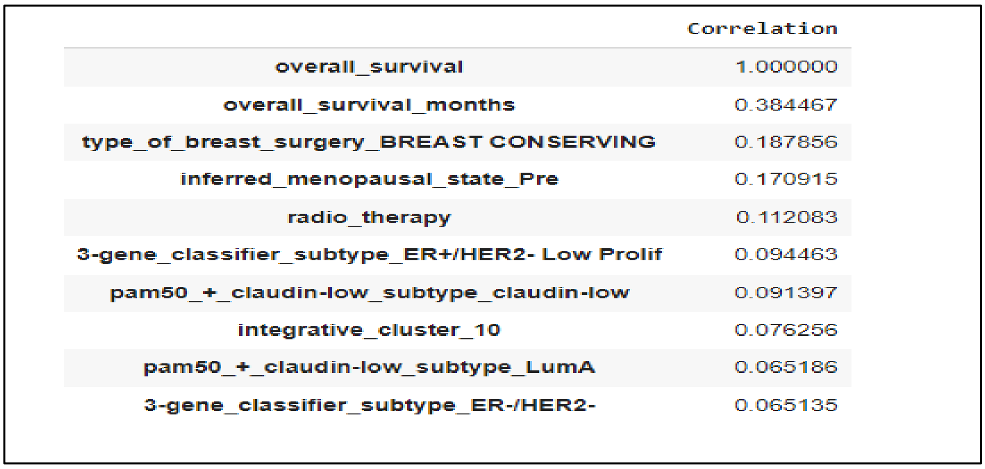
Fig 3.9 EDA performed on distribution of histopathology class and survival

Figure 3.9 visualizes the relationship between different histopathological classes and patient survival outcomes. By examining the chart, the following observations can be made:

- **General Class Distribution and Survival Trend:** Across most histopathological classes, there's a clear trend where the number of patients who passed away exceeds those who survived. This general pattern suggests a possible link between specific histopathological classifications and increased mortality rates, highlighting the importance of histopathological assessments in prognostic evaluations.

- **Class 1.0 Specific Observation:** A notable deviation from the general trend is seen in class 1.0. Here, the survival and mortality counts are nearly equal, indicating an almost similar distribution of patients who survived and those who didn't within this class. This balance contrasts significantly with other classes, where differences between survival and non-survival groups are more pronounced.

The insights from Figure 3.9 emphasize the profound influence of histopathological classifications on patient survival outcomes. While some classes might indicate a more favorable prognosis, others could signify a higher risk of mortality. Recognizing these class-specific patterns is vital for clinicians to formulate informed therapeutic strategies and provide patients with a clearer understanding of their prognosis based on histopathological evaluations.



	Correlation
overall_survival	1.000000
overall_survival_months	0.384467
type_of_breast_surgery_BREAST CONSERVING	0.187856
inferred_menopausal_state_Pre	0.170915
radio_therapy	0.112083
3-gene_classifier_subtype_ER+/HER2- Low Prolif	0.094463
pam50_+_claudin-low_subtype_claudin-low	0.091397
integrative_cluster_10	0.076256
pam50_+_claudin-low_subtype_LumA	0.065186
3-gene_classifier_subtype_ER-/HER2-	0.065135

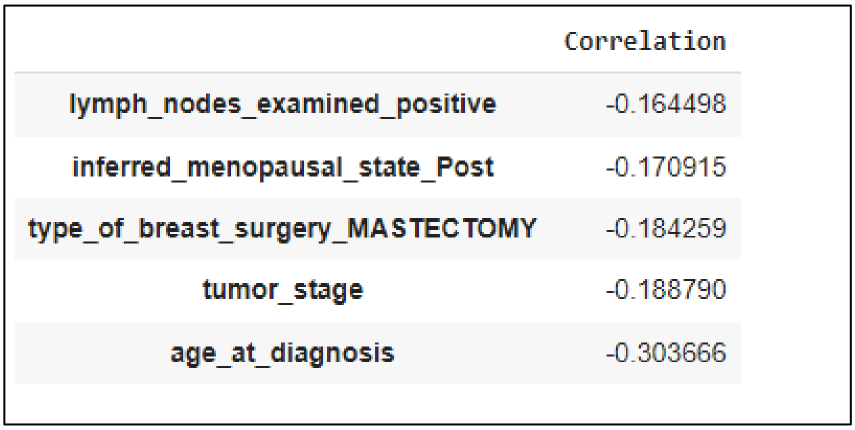
Fig 3.10 Positive correlation with overall survival

As depicted in Figure 3.10, certain attributes display a positive correlation with overall survival. Analyzing the visualization, we can derive the following insights:

- **Attributes Displaying Positive Correlation:** A group of particular attributes emerges as having a favorable relationship with overall survival. The presence of these attributes or their increased levels suggest an enhancement in the likelihood of a patient's survival.

- **Degree of Correlation:** The visualization communicates not just the existence of a correlation, but also its intensity. While the presence of these attributes might suggest a better survival probability, it's pivotal to remember that correlation doesn't necessarily indicate a direct cause-and-effect relationship.
- **Implications in a Clinical Context:** Identifying and understanding attributes that positively correlate with survival can be crucial in clinical scenarios. They could serve as beneficial indicators, pointing towards improved prognostic outcomes. Equipped with knowledge about these attributes, healthcare professionals can devise more effective treatment plans and strategies, ensuring the well-being of patients.

The conclusions drawn from Figure 3.10 emphasize the importance of certain attributes in positively influencing patient survival outcomes. This understanding can be instrumental in framing research directions, therapeutic interventions, and patient management in oncology.



	Correlation
lymph_nodes_examined_positive	-0.164498
inferred_menopausal_state_Post	-0.170915
type_of_breast_surgery_MASTECTOMY	-0.184259
tumor_stage	-0.188790
age_at_diagnosis	-0.303666

Fig 3.11 Negative correlation with overall survival

Figure 3.11 depicts the attributes that manifest a negative correlation with overall survival. Analyzing the visualization, the following observations can be made:

- **Attributes with Negative Correlation:** The visualization outlines certain attributes that display an inverse relationship with overall survival. This suggests that higher values of these attributes might correspond with decreased survival rates.

- **Intensity of Correlation:** The visualization offers intensities, each representing the strength of the negative correlation. More intense zones may suggest a stronger negative correlation, emphasizing the attributes that might notably impact overall survival.
- **Clinical Implications:** Identifying attributes that have a negative correlation with survival can be instrumental in clinical scenarios. These attributes can act as indicators or markers that could hint at unfavorable prognostic outcomes. Through a keen understanding of these attributes, healthcare practitioners can be better equipped to forecast challenges and strategize interventions.

The information derived from Figure 3.11 emphasizes the critical role certain attributes play in potentially influencing decreased patient survival rates. Such understandings are vital for devising proactive clinical strategies and ensuring optimal patient care in oncology.

3.3 Exploratory Data Analysis of Genomic Data

Exploratory Data Analysis (EDA) stands as a cornerstone in the data analysis landscape, especially within the intricate realm of genomics. Genomic data, characterized by its high dimensionality and complexity, encapsulates the genetic information that can be pivotal in understanding disease mechanisms, progression, and potential therapeutic targets. EDA, in this context, involves visualizing, summarizing, and interpreting the genetic variations or mutations present within this data, ensuring that significant patterns, relationships, or anomalies are effectively identified.

The importance of EDA in genomics cannot be overstated. Given the vast amount of genetic data available, EDA acts as a sieve, helping researchers discern the few relevant genetic markers or mutations from the many that might be inconsequential. This is particularly crucial in cancer research, where certain genes might play a role in the onset, progression, or treatment response of the disease. By performing a detailed EDA on genomic data, researchers can gain a preliminary understanding of these critical genes, hypothesize their role in the disease, and strategize subsequent analytical or experimental approaches based on these insights.

In the context of this chapter, our focus is honed on the EDA of specific genes that have previously shown relevance in cancer biology and patient outcomes. The genes under consideration are:

- **PIK3CA:** Mutations in the PIK3CA gene have been associated with various cancers and can influence patient outcomes. The PI3K/AKT/mTOR pathway, in which this gene plays a pivotal role, is a target for many cancer therapies.
- **KIT:** Mutations in KIT are of particular importance in gastrointestinal stromal tumors (GISTs) and some leukemia. The presence or absence of these mutations can influence therapy choices and outcomes.
- **MYC:** Amplification or over expression of the MYC oncogene is associated with a variety of cancers and is often linked to aggressive tumor behavior and poor prognosis.
- **EGFR:** Implicated in tumor progression and holds prognostic implications in specific cancers.
- **TP53:** The TP53 gene, which encodes the p53 protein, is one of the most frequently mutated genes in human cancers. Mutations in TP53 can influence tumor behavior and patient outcomes. Depending on the specific mutation and the cancer type, TP53 mutations can be associated with either better or worse prognosis. One of the most frequently mutated genes in human cancers, playing a key role in tumor suppression.
- **ATM:** Involved in DNA repair mechanisms and has associations with certain cancer types.
- **CDH1:** Known for its role in invasive properties of tumors, especially in lobular breast cancer and gastric cancer.

It's important to understand that while these genes have been associated with cancer outcomes in various studies, their importance in predicting overall survival will depend on the specific dataset, the cancer type, and the context in which they are studied. Additionally, the relationship between gene status (e.g., mutated vs. wild-type) and survival may not always be linear or straightforward.

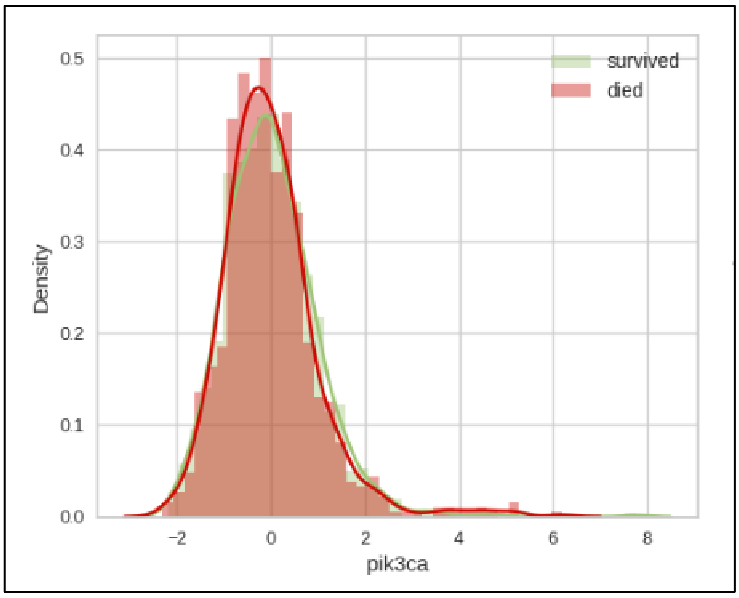


Fig 3.12 Mutation Distribution of the PIK3CA Gene across Patient Samples

The mutation spectrum of the PIK3CA gene, as presented in Figure 3.12, provides a comprehensive overview of the genetic alterations associated with this gene across various samples. Each vertical line depicts a specific mutation site within the gene, and its height indicates the frequency of that particular mutation within the dataset. Several observations emerge from this visualization:

- **Variability in Mutation Sites:** The PIK3CA gene showcases a diverse range of mutation sites, reflecting the genetic heterogeneity associated with this gene.
- **Frequency of Specific Mutations:** Some mutation sites, evident from the taller vertical lines, are more frequent than others. These dominant mutations might hold higher genomic significance, warranting further investigation.
- **Potential Hotspots:** Areas with a dense clustering of vertical lines may represent mutation "hotspots" — specific areas within the gene that are more prone to genetic alterations. Such hotspots can be crucial in understanding the gene's role in disease mechanisms and might serve as potential targets for therapeutic interventions.

Understanding the mutation spectrum of the PIK3CA gene is pivotal, given its established role in various cancers. The visual representation in Figure 3.12 aids in simplifying complex genomic information, offering insights that can guide future research directions and therapeutic strategies.

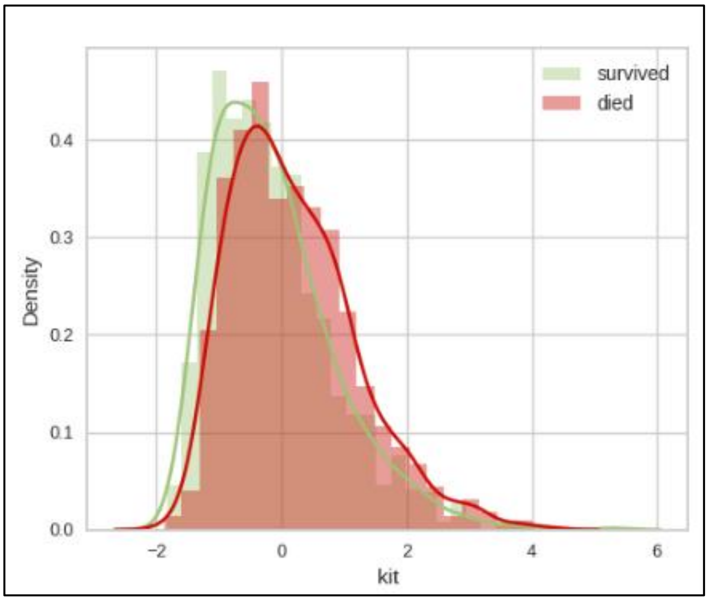


Fig 3.13 Mutation Distribution of the KIT Gene across Patient Samples

The given figure 3.13 illustrates the distribution of the KIT gene across various patient samples. The visualization utilizes bars to denote the frequency or count of patients exhibiting specific levels or statuses of the KIT gene. The horizontal axis perhaps categorizes patients based on certain criteria or groupings, while the vertical axis quantifies the number of patients within each category. Recognizing the prominence of specific categories can be instrumental, as the expression or mutation status of the KIT gene might be linked to disease progression, patient prognosis, and response to treatments.

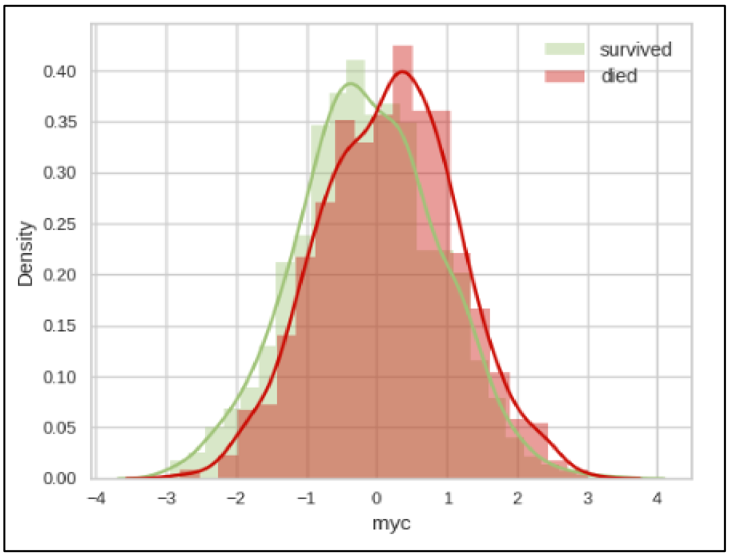


Fig 3.14 Mutation Distribution of the MYC Gene across Patient Samples

This figure 3.14 presents the distribution of the MYC gene across different patient samples. Through the use of bars, the visualization effectively displays the frequency or count of patients showcasing specific levels or statuses of the MYC gene. The horizontal axis likely represents distinct categories or classifications related to the gene, while the vertical axis enumerates the patient count within each category. Observing predominant categories can be pivotal, as the expression or mutation status of the MYC gene can influence various factors such as disease behavior, patient prognosis, and potential therapeutic responses.

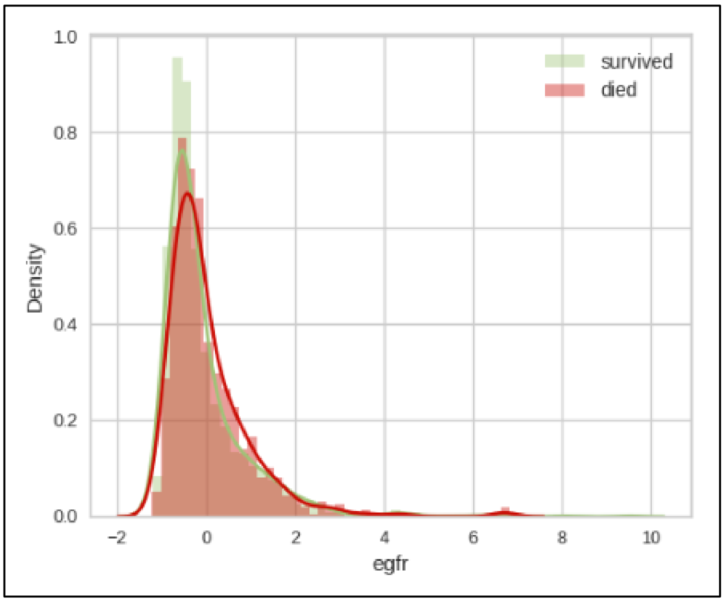


Fig 3.15 Mutation Distribution of the EGFR Gene across Patient Samples

The provided figure 3.15 presents the distribution of the EGFR gene across patient samples, revealing a right-skewed pattern. This suggests that the majority of patients have lower expression levels (or mutation frequencies) of the EGFR gene, while a smaller segment of patients display higher expression levels.

A right-skewed distribution in this context highlights that while lower expression or mutation levels of the EGFR gene are more prevalent in the patient cohort, there are still a noteworthy number of cases with elevated levels. Given that the EGFR gene plays pivotal roles in various cancers, understanding its distribution can provide valuable insights into disease progression, prognosis, and potential therapeutic strategies.

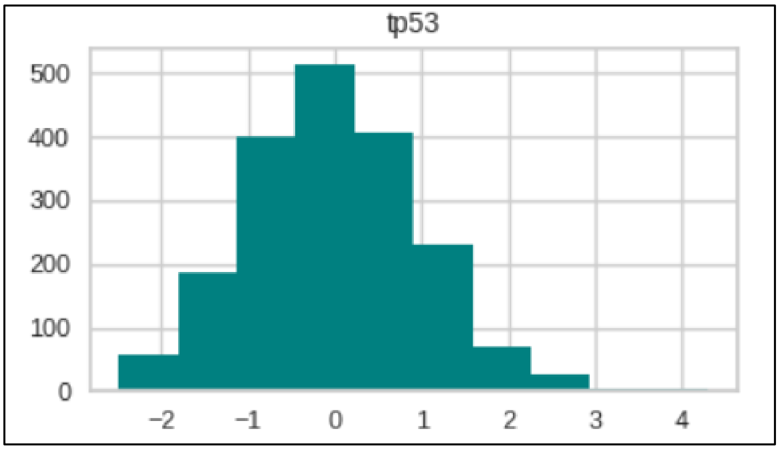


Fig 3.16 Distribution of the TP53 Gene across Patient Samples

The displayed figure 3.16 depicts the distribution of the TP53 gene across patient samples. The pattern shown resembles a right-skewed distribution, indicating that a majority of patients possess lower expression levels (or mutation frequencies) for the TP53 gene, while a fewer number of patients exhibit higher expression levels.

The right-skewed nature in this scenario suggests that although lower expression or mutation levels of the TP53 gene are more common within the patient cohort, there exists a certain subset of patients with pronounced levels. Given the significance of the TP53 gene in tumorigenesis and its designation as a tumor suppressor, understanding its distribution can be instrumental for insights into cancer biology, potential risk factors, and therapeutic considerations.

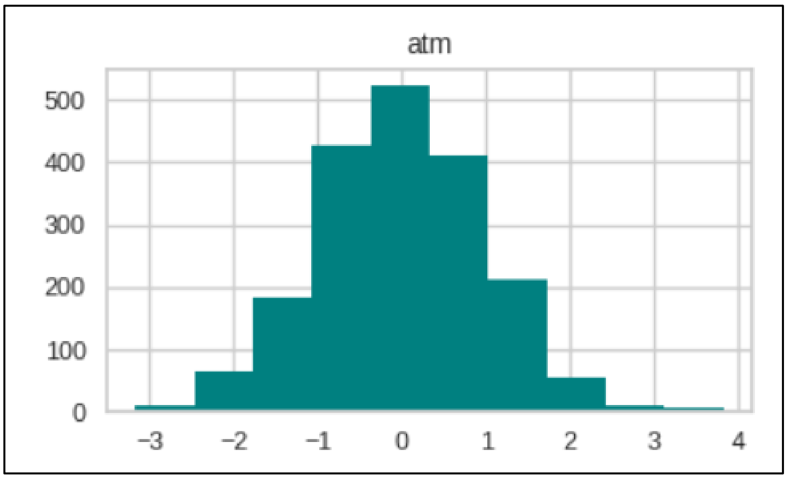


Fig 3.17 Distribution of the ATM Gene across Patient Samples

The provided figure 3.17 portrays the distribution of the ATM gene across various patient samples. Observationally, the data seems to be centered around a particular range, indicating a form of central or near-normal distribution. A large number of patients have ATM gene expression or mutation frequencies that cluster around this central value, with fewer patients deviating significantly to the higher or lower ends.
The ATM gene, known for its pivotal role in DNA damage repair, plays a crucial part in many cellular processes related to cancer. Understanding its central distribution in this dataset might offer insights into its standard behavior in the patient cohort under study, suggesting that drastic deviations from this central range might be of clinical significance.

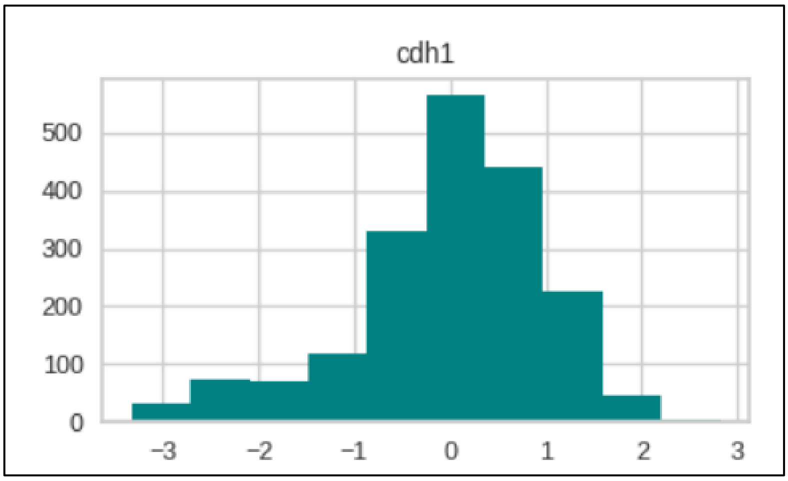


Fig 3.18 Distribution of the CDH1 Gene across Patient Samples

The provided graph 3.18 illustrates the distribution of the CDH1 gene across various patient samples. At first glance, the distribution is noticeably left-skewed, meaning a majority of the patient samples exhibit higher values for the CDH1 gene expression or mutation frequency, while a lesser number of samples have significantly lower values. The CDH1 gene, crucial for cell adhesion processes, has been linked to various cancer types when mutated. Given its left-skewed distribution, it indicates that in the patient cohort under investigation, higher expression or mutation frequencies of the CDH1 gene are more common. Such insights can be vital for understanding the genomic landscape of the patient population and tailoring therapeutic approaches accordingly.

3.4 Architecture of Gated Attention Convolution Neural Network

The realm of biomedical research, especially in the domain of cancer prognosis, has witnessed a paradigmatic shift with the introduction of advanced deep learning models. Among these, the Gated Attention Convolution Neural Network (Gated Attention CNN) has emerged as a potent tool, meticulously designed to cater to the intricacies of multimodal data processing inherent in oncological datasets.

Significance of Gated Attention CNN model:

- **Efficiency with Multimodal Data:** In oncological research, the confluence of data from varied sources is a frequent occurrence, with each modality offering its unique perspective and information content. The Gated Attention CNN is adept at seamlessly integrating such heterogeneous datasets, ensuring that salient features from each modality are coherently amalgamated, thus enhancing the predictive prowess of the model.
- **Dynamic Adaptability:** A hallmark of the Gated Attention mechanism is its ability to adaptively discern which features or data segments warrant emphasis, mirroring the cognitive ability of humans to selectively focus on pertinent stimuli. This dynamic adaptability ensures that the model remains attuned to subtle, yet potentially critical, patterns in the data, enhancing its prognostic accuracy.
- **Robustness and Generalizability:** In the context of medical applications, a model's ability to generalize across diverse patient cohorts is of paramount importance. The architectural nuances of the Gated Attention CNN, particularly its attention mechanism, act as a bulwark against overfitting, ensuring that the model remains robust and generalizable across varied datasets.
- **Relevance in Cancer Survival Prognosis:** Cancer survival prediction is an intricate task, contingent upon a nuanced understanding of a myriad of clinical and genomic variables. The Gated Attention CNN, by virtue of its dynamic attention mechanism, ensures that these intricate interdependencies are not merely acknowledged but are also factored into the prognostic predictions, thus enhancing both their accuracy and clinical relevance.

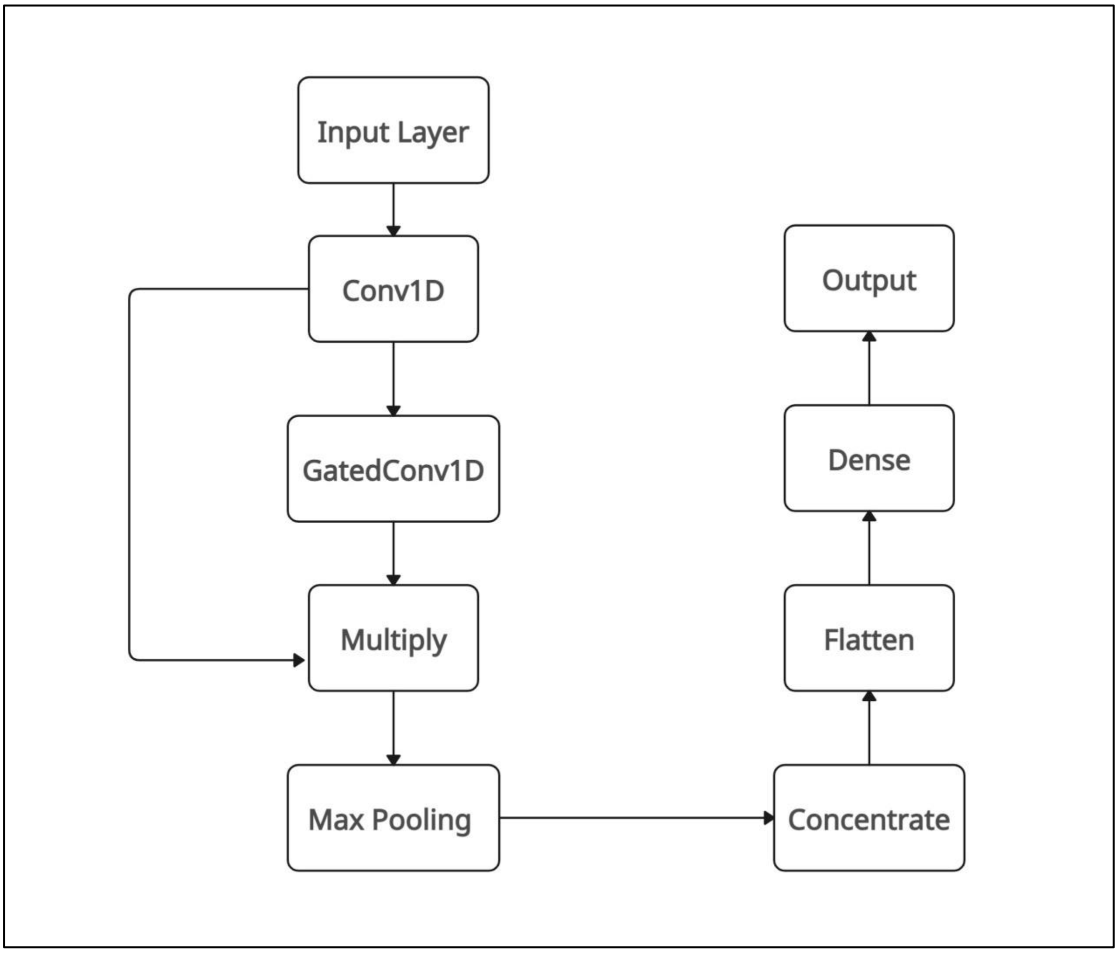


Fig 3.19 Low level Architecture of Gated Attention CNN model

Upon examination of the provided architecture as seen in figure 3.19, the model appears to have multiple layers designed for feature extraction, attention mechanism, and final decision-making.

- **Feature Extraction Layers:** These layers, typically convolutional in nature, are responsible for extracting relevant patterns from the input data. Given the multimodal nature of our dataset, separate branches seem to be employed for each data type (e.g., clinical data, gene expression data, and copy number alteration data).
- **Gated Attention Layers:** Positioned after the feature extraction layers, this mechanism dynamically determines which features to focus on. This is achieved through a gating mechanism that assigns weights to different features based on their relevance, allowing the model to pay "attention" to more pertinent information.

- **Fusion and Decision Layers:** After processing each data modality, the model fuses the information and passes it through additional layers (often fully connected) to make the final prediction about cancer survival.

In essence, the Gated Attention CNN model is a holistic architecture, effectively marrying the prowess of convolutional neural networks with the adaptability of attention mechanisms. When applied to the task of cancer survival prediction, this model holds the potential to unravel the nuanced relationships between various clinical and genomic attributes, setting the stage for predictions that are not only accurate but also clinically interpretable.

3.5 Parameter Details of Proposed Model

A robust architecture is the backbone of any effective deep learning model. However, the granular details, particularly the parameters that govern the behavior and performance of this architecture, are equally critical. This section discusses the parameters employed in the proposed Gated Attention CNN model, providing a comprehensive breakdown to ensure clarity and aid in potential replication efforts.

Table 3.1: Specification of the Convolution Layer

Number of layers	2
Dimensions of the Filter	2, 3
Total filters used	30
Step size	2
Padding	Identical
Function for Activation	Rectified Linear Unit

In table 3.1, the convolutional layer is a foundational component of our Gated Attention CNN model, and understanding the selected parameters is crucial for grasping its operational mechanics.

- **Number of layers (2):** Two horizontal convolution layers are employed to ensure a multi-level feature extraction. This allows the model to capture both low-level and slightly more abstract features from the input data.
- **Dimensions of the Filter (2*1, 3*1):** The use of multiple filter sizes, specifically 2*1 and 3*1, allows the model to scan the input data through different receptive fields. This ensures that both granular and broader patterns within the data are identified and processed.
- **Total filters used (30):** 30 filters in each convolution layer enhance the depth of the feature maps. This depth allows for the detection of a variety of patterns, making the model versatile in its feature recognition capabilities.
- **Step size (2):** A step size of 2 ensures that the filters move in larger increments, leading to a reduction in the spatial dimensions of the output feature maps. This not only reduces computational load but also aids in capturing more generalized features.
- **Padding (Identical):** 'Identical' padding ensures that the spatial dimensions of the output feature maps are consistent with the input, preserving the spatial context and ensuring no information is lost at the borders.
- **Function for Activation (Rectified Linear Unit):** The Rectified Linear Unit activation function introduces non-linearity into the model, enabling it to learn complex relationships in the data. Its popularity stems from its efficiency and ability to mitigate the vanishing gradient problem, which is common in deep neural networks.

By understanding these parameters, one gains insight into the convolutional layer's capability to scan, recognize, and transform the input data into meaningful feature maps that serve as the foundational building blocks for the subsequent layers in the architecture.

Table 3.2: Specification of the Gated Attention Layer

Number of layers	2
Dimensions of the Filter	1, 3
Total filters used	30
Step size	2
Padding	Identical
Function for Activation	Sigmoid

In table 3.2, the Gated Attention Layer plays a central role in enabling our model to dynamically focus on different parts of the input data, enhancing its ability to detect intricate patterns.

- **No. of layers (2):** Much like the convolution layer, the Gated Attention Layer employs two horizontal convolution layers. This design choice ensures that the model can capture attention patterns at multiple levels, allowing for a nuanced understanding of the data.
- **Dimensions of the Filters (1*1, 3*1):** These filter sizes are chosen to capture both local and slightly broader attention patterns. A filter size of 1*1 captures individual data points, while a size of 3*1 encompasses a slightly larger context, ensuring the model can weigh the importance of data points in relation to their neighbors.
- **Total filters used (30):** With 30 filters; the model has the capacity to detect a diverse set of attention patterns, enhancing its adaptability and depth of understanding.
- **Step size (2):** A step of 2 ensures that the attention mechanisms cover the input data comprehensively, without unnecessary overlap, thereby optimizing computational efficiency.
- **Padding (Identical):** By using 'Identical' padding, the output after applying the filters retains the same dimension as the input. This ensures that no data point is left unattended and the spatial hierarchies are preserved.

- **Function for Activation (Sigmoid):** The sigmoid activation function ensures that the attention weights lie between 0 and 1. This is crucial as it allows the model to assign varying degrees of importance to different data points, effectively 'gating' the flow of information based on its relevance.

By meticulously setting these parameters, the Gated Attention Layer is optimized to discern the most salient features in the data, laying the groundwork for the subsequent layers to make informed predictions.

Table 3.3: Specification of the Max Pooling Layer

Pool size	2*2
Step size	1
Padding	Identical

In Table 3.3, the Max Pooling Layer serves as a dimensionality reduction mechanism, ensuring that while the data is condensed, the most informative features are retained.

- **Pool size (2*2):** With a pool size of 2*2, the layer takes the maximum value from a 2x2 patch of the input data. This design choice allows the model to condense data by half, reducing computational demands for subsequent layers without significant loss of information.
- **Step size (1):** A step of 1 ensures overlapping pooling. While this might seem counterintuitive given pooling's objective to reduce dimensionality, a stride of 1 ensures that no crucial information is overlooked. By considering overlapping patches of input data, the model ensures that it captures the most dominant features irrespective of their position.
- **Padding (Identical):** By employing 'Identical' padding, the output post-pooling retains the same spatial dimensions as the input prior to pooling. This choice ensures a consistent data structure, allowing for smoother transitions between layers and easier integration with subsequent architectural components.

The careful configuration of the Max Pooling Layer aids in preserving the architecture's efficiency. By retaining only the most salient features and discarding redundant or less informative data, the model remains computationally efficient without compromising on its ability to discern patterns and make accurate predictions.

Table 3.4: Specification of the Fully Connected Layer

Total count of hidden layers	3
Neurons count in each hidden layers	200, 150, 100
Percentage of neurons ignored	50%
Function for Activation	Hyperbolic Tangent

In Table 3.4, the Fully Connected Layers serve as the integration hubs, where all the extracted features from previous layers are connected, processed, and used for the final decision-making.

- **Total count of hidden layers (3):** The model employs a three-layered dense network, which means that the extracted features undergo three levels of transformation before the final output. This multi-layered approach facilitates intricate pattern recognition, allowing the model to capture non-linear relationships within the data effectively.
- **Neurons count in each hidden layers (200, 150, 100):** The gradual decrease in the number of neurons from 200 to 100 in subsequent layers is strategic. It ensures that while the initial layer can accommodate a broader spectrum of features, the subsequent layers focus on refining and condensing this information, ensuring that only the most pivotal features influence the final decision.
- **Percentage of neurons ignored (50%):** Dropout is a regularization technique employed to prevent overfitting. By randomly setting 50% of the neurons to zero during each training iteration, the model ensures that no single neuron becomes overly specialized. This not only promotes a more generalized model but also enhances its robustness.
- **Function for Activation (Hyperbolic Tangent):** The hyperbolic tangent activation function transforms the weighted sum of its inputs from the previous

layer to values between -1 and 1. It is particularly advantageous for tasks where the output needs to be centered around zero, ensuring a balanced model response.

Incorporating these parameters within the fully connected layers ensures the model's capacity to effectively process, integrate, and transform the learned features into a meaningful output, thereby enhancing its predictive accuracy and generalization capabilities.

Table 3.5: Parameter Details of Other Parameter

Activation mechanism in the output layer	Sigmoid activation
Batch size during training	8
Number of training cycles	50
Function to calculate model loss	Combination of Binary Cross-Entropy and L2 Regularization

In table 3.5, the additional parameters serve to fine-tune the model's learning process, ensuring that it converges to an optimal solution efficiently and effectively.

- **Activation mechanism in the output layer (Sigmoid activation):** The output layer uses the sigmoid activation function, which transforms its input to a value between 0 and 1. This is especially useful for binary classification tasks like ours, where the output represents the probability of a particular class.
- **Batch size during training (8):** Mini-batch gradient descent, with a size of 8, is employed in the model's training. This means that instead of updating the model's weights after each individual data point (stochastic) or after the entire dataset (batch), the model updates its weights after every 8 data points. This strikes a balance between computational efficiency and the ability to escape potential local minima during training.
- **Number of training cycles (50):** The model undergoes 50 complete forward and backward passes of all the training examples. This ensures that the model has ample opportunity to learn the intricate patterns in the data without overfitting.

- **Function to calculate model loss (Combination of Binary Cross-Entropy and L2 Regularization):** The model employs a binary cross-entropy loss function, which is suitable for binary classification tasks. It measures the difference between the actual and the predicted probabilities. Additionally, L2 regularization is incorporated, penalizing large coefficients in the model. This encourages the model to have smaller weights, making it simpler and helping prevent overfitting.

In summary, these additional parameters are meticulously chosen to harmonize with the overarching model architecture, ensuring that the training process is both efficient and effective, leading to a model that is both robust and generalizable.

3.6 Summary

The journey undertaken in this chapter has been both methodical and illuminating. Beginning with a structured approach towards understanding the design choices and intricacies of cancer survival prediction, we delved deep into the nuances of data – both clinical and genomic. The Exploratory Data Analysis (EDA) provided a lens to view the multifaceted characteristics of the data, discerning patterns, anomalies, and potential correlations that could be pivotal in predicting outcomes.

Transitioning from data exploration, the emphasis shifted to the Gated Attention Convolutional Neural Network (CNN) model. The architecture's design, driven by the complexities of multimodal datasets, revealed its potential in harnessing this data for effective survival prediction. The segmented approach – treating clinical, gene expression, and copy number alteration data uniquely – showcased the model's adaptability and precision. The granular details, especially the parameters governing this architecture, were meticulously described to ensure replicability and a deep understanding of the underlying mechanics.

This chapter has not just been about laying down the technical foundations; it's also about painting a clear picture of the methodological precision required for such a task. The insights garnered from the EDA, combined with the robustness of the Gated Attention CNN, offer a promising avenue for future research in cancer survival prediction. As we transition to the subsequent chapters, the groundwork laid here ensures that the analytical and evaluative processes are built on a foundation of clarity, depth, and scientific rigor.

CHAPTER 4: IMPLEMENTATION OF CANCER SURVIVAL PREDICTION SYSTEM

This chapter outlines the concrete steps taken to translate the theoretical foundations of cancer survival prediction into a practical, executable system. The chapter begins with a detailed discussion of the tools and techniques, offering an insight into the programming environment, hardware specifications, software details, and the libraries and packages that form the backbone of the research. Here, the reader is introduced to the intricacies of the Python programming language and its associated tools, which have been leveraged to handle the extensive data processing and complex model training integral to this study.

Subsequent sections provide a granular view of the implementation details for various modules of the cancer survival prediction system. Through an examination of source code snippets and the corresponding outputs, the chapter aims to demystify the inner workings of the system, offering clarity on how each component contributes to the overall goal of accurate survival prediction. From preprocessing steps across different types of data within the METABRIC dataset to the construction and operation of the Gated Attention CNN models, and the evaluation metrics used to validate their performance, each piece of the puzzle is laid out for examination.

The chapter promises to culminate in a comprehensive discussion on the role of the Random Forest algorithm—an ensemble learning method known for its robustness and accuracy. By synthesizing insights across multiple data modalities, the Random Forest model is poised to provide a nuanced assessment of the predictive power of the combined feature sets, solidifying the research's contribution to the field of oncological data science.

In essence, this chapter not only serves as a technical repository of the research's computational aspects but also as a testament to the methodical approach adopted to ensure that the findings are replicable, robust, and grounded in the latest advancements in machine learning and data analytics. The transparency and detail provided herein underscore the reliability of the research and the potential of the developed system to influence the landscape of cancer prognosis significantly.

4.1 Tools & Techniques

In the pursuit of an advanced and meticulous analysis of cancer survival prediction using the METABRIC dataset, a range of tools and techniques were employed. These tools not only facilitated data processing and model training but also ensured that the analysis was grounded in state-of-the-art methodologies. This section elucidates the technological environment and the software libraries that underpinned our research.

4.1.1 Programming Language & Environment

- **Python:** Our choice of programming language was Python, renowned for its versatility and wide applicability across data analysis, machine learning, and scientific computing realms. The syntax simplicity coupled with its vast library ecosystem makes Python a preferred language for data-driven research.
- **Version:** We used Python version 3.11.3, an updated version equipped with optimizations and new features, which augmented the efficiency of our coding endeavors.
- **Platform:** The experiments were conducted on Google Colab, a cloud-based extension of Jupyter Notebooks. It's a platform tailored for interactive Python execution with the added advantage of GPU access, seamless sharing, and a hassle-free setup.
- **Specifications:** Google Colab provided an Intel Xeon CPU with 2 vCPUs and 13GB RAM, facilitating robust data processing and facilitating extensive model training sessions.

4.1.2 Hardware Specifications

- **Processor:** The computational heavy lifting was handled by the Intel(R) Core(TM) i5-8265U CPU. With a speed ranging from 1.60GHz to 1.80 GHz, this processor is adept at ensuring swift and consistent performance, which is pivotal when dealing with intensive data analysis and machine learning model training.
- **RAM:** The system was equipped with 12.0 GB RAM, of which 11.9 GB was usable. This sizable RAM was instrumental in multitasking and managing large datasets seamlessly.

4.1.3 Software Details

- **TensorFlow:** TensorFlow, an open-source library by Google, was central to our machine learning endeavors. We employed version 2.8.0, which encapsulates a plethora of features optimized for efficient model training and deployment.
- **Operating System:** The research was conducted on a system running Windows 10 Home Single Language edition.
- **Version & Build:** Specifically, the system was on version 22H2 with an OS build of 19045.3448. This detailed versioning provides a snapshot of the specific updates and features available during our research.

4.1.4 Libraries & Packages

The analysis heavily relied on several Python libraries to streamline the process. Here's a brief overview:

- **Keras:** In the complex domain of cancer survival prediction, the development, design, and training of neural network models are central to the analytical framework. Keras, renowned for its high-level neural networks API, facilitated an efficient and streamlined model creation process. It provided a user-friendly interface, allowing for iterative prototyping of different architectures and configurations. Given its seamless integration with TensorFlow, Keras ensured that the underlying computational mechanics were optimized and efficient. Specifically, in our research, Keras was instrumental in the meticulous design of the Gated Attention CNN model. This ensured the model was not just theoretically robust, but practically efficient, harnessing the full potential of the underlying neural network architecture.
- **Sklearn:** The intricate task of cancer survival prediction necessitates rigorous data preprocessing, model validation, and evaluation. Sklearn, with its comprehensive suite of tools, was indispensable in these facets of our research. The library's utilities facilitated data normalization, transforming the METABRIC dataset into a format primed for model training. The stratified K-fold cross-validation offered by Sklearn ensured a robust validation of our model, illuminating its potential for generalizability across diverse datasets. Further, Sklearn's evaluation metrics

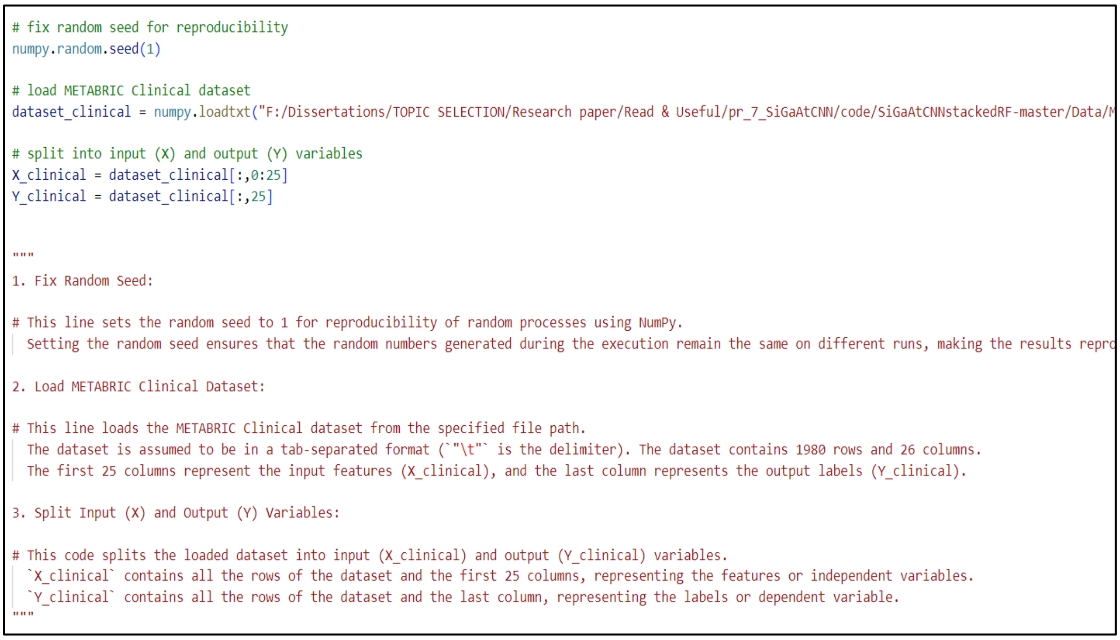
provided a nuanced assessment of our model's performance, aligning our computational predictions with clinical relevance.

- **TensorFlow:** While Keras offers a high-level, intuitive interface, our research occasionally demanded more granular control and bespoke functionalities. TensorFlow, with its vast computational capabilities, was our tool of choice in these scenarios. As a cutting-edge machine learning framework, TensorFlow enabled intricate model customizations, ensuring our Gated Attention CNN model was both trained and optimized with precision. Leveraging the capabilities of TensorFlow version 2.8.0, we ensured that our model's training phase was computationally efficient and that the resultant predictions were rooted in reliability.
- **Matplotlib:** In the world of data science and, by extension, cancer survival prediction, the role of data visualization cannot be overstated. Matplotlib, a versatile visualization library, was invaluable in providing a comprehensive visual perspective on our data and results. From elucidating the intricate distributions of the genomic data, such as gene expression profiles, to portraying the performance metrics of our model, Matplotlib offered a canvas to represent our insights visually. Such graphical representations, especially in a domain as intricate as cancer survival prediction, facilitated a more intuitive comprehension of the data's underlying patterns, enriching both our model design and validation processes.

In conclusion, the ensemble of these tools and libraries crafted a robust and state-of-the-art environment. This ensured the research was not just theoretically sound but was also underpinned by the latest advancements in computational tools and machine learning frameworks.

4.2 Implementation Details of Cancer Survival Prediction Modules

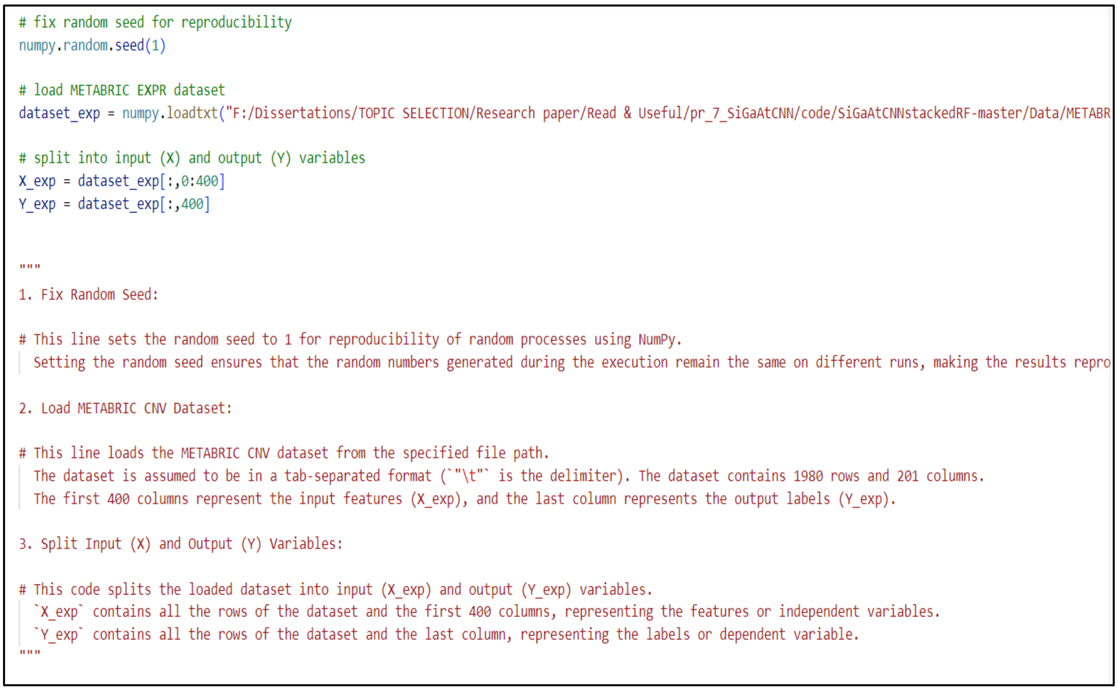
In this section, we delve into the practical aspects of our cancer survival prediction system. By examining snippets of source code and their corresponding outputs, we offer a transparent view into the inner workings of our system. This includes preprocessing steps for various data types within the METABRIC dataset, the architecture and operation of the Gated Attention CNN models, and the utilization of evaluation metrics. We conclude with a discussion on the ensemble application of Random Forest, which synthesizes the insights drawn from multiple data modalities.



```
# fix random seed for reproducibility
numpy.random.seed(1)

# load METABRIC Clinical dataset
dataset_clinical = numpy.loadtxt("F:/Dissertations/TOPIC SELECTION/Research paper/Read & Useful/pr_7_SiGaAtCNN/code/SiGaAtCNNstackedRF-master/Data/M

# split into input (X) and output (Y) variables
X_clinical = dataset_clinical[:,0:25]
Y_clinical = dataset_clinical[:,25]


"""
1. Fix Random Seed:

# This line sets the random seed to 1 for reproducibility of random processes using NumPy.
  Setting the random seed ensures that the random numbers generated during the execution remain the same on different runs, making the results repro

2. Load METABRIC Clinical Dataset:

# This line loads the METABRIC Clinical dataset from the specified file path.
  The dataset is assumed to be in a tab-separated format (`"\t"` is the delimiter). The dataset contains 1980 rows and 26 columns.
  The first 25 columns represent the input features (X_clinical), and the last column represents the output labels (Y_clinical).

3. Split Input (X) and Output (Y) Variables:

# This code splits the loaded dataset into input (X_clinical) and output (Y_clinical) variables.
  `X_clinical` contains all the rows of the dataset and the first 25 columns, representing the features or independent variables.
  `Y_clinical` contains all the rows of the dataset and the last column, representing the labels or dependent variable.
"""
```

Fig 4.1: Snapshot of the code for Initialization & Preprocessing of the METABRIC Clinical Dataset

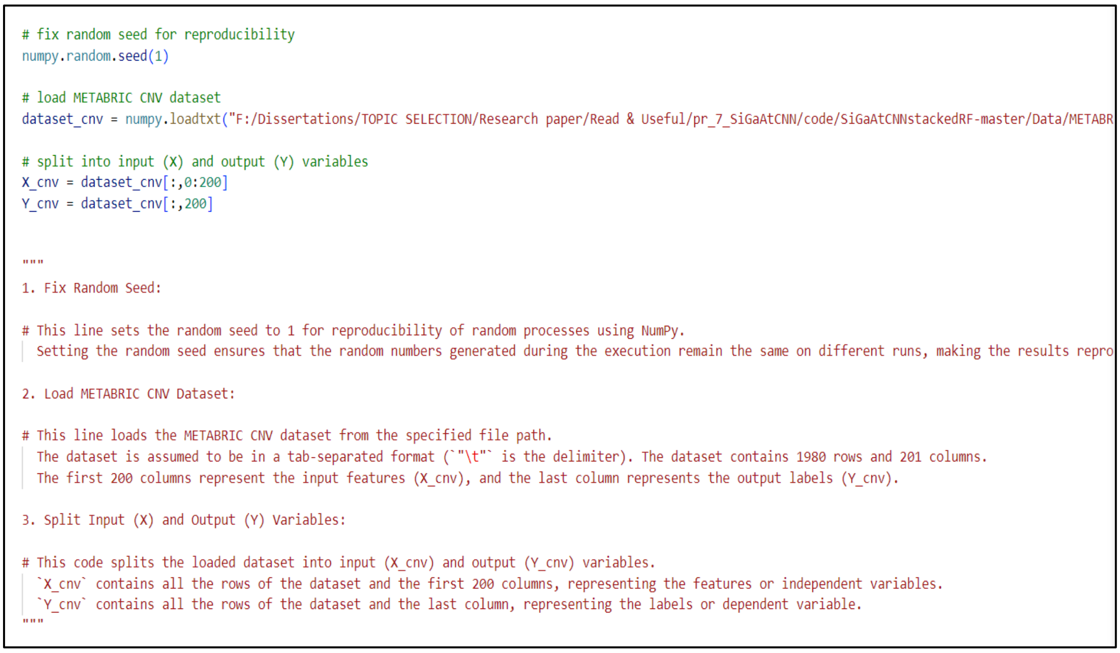
Figure 4.1 provides a detailed illustration of the initial steps in processing the METABRIC clinical dataset, an essential part of the cancer survival prediction system. This figure outlines the code that establishes a reproducible environment by fixing the random seed, ensuring that the stochastic elements of our analysis can be consistently recreated for verification and comparison. It then details the process of loading the dataset from a specified location, highlighting the use of a tab-separated format for the inclusion of 1980 instances each characterized by 26 distinct attributes. The subsequent step in the figure delineates the division of the dataset into independent variables (X_clinical), encompassing a range of clinical features, and the dependent variable (Y_clinical), representing the outcome labels. This segmentation is critical as it sets the foundation for the model to learn the underlying patterns associated with patient survival.



```
# fix random seed for reproducibility
numpy.random.seed(1)

# load METABRIC EXPR dataset
dataset_exp = numpy.loadtxt("F:/Dissertations/TOPIC SELECTION/Research paper/Read & Useful/pr_7_SiGaAtCNN/code/SiGaAtCNNstackedRF-master/Data/METABR

# split into input (X) and output (Y) variables
X_exp = dataset_exp[:,0:400]
Y_exp = dataset_exp[:,400]


"""
1. Fix Random Seed:

# This line sets the random seed to 1 for reproducibility of random processes using NumPy.
  Setting the random seed ensures that the random numbers generated during the execution remain the same on different runs, making the results repro

2. Load METABRIC CNV Dataset:

# This line loads the METABRIC CNV dataset from the specified file path.
  The dataset is assumed to be in a tab-separated format (`"\t"` is the delimiter). The dataset contains 1980 rows and 201 columns.
  The first 400 columns represent the input features (X_exp), and the last column represents the output labels (Y_exp).

3. Split Input (X) and Output (Y) Variables:

# This code splits the loaded dataset into input (X_exp) and output (Y_exp) variables.
  `X_exp` contains all the rows of the dataset and the first 400 columns, representing the features or independent variables.
  `Y_exp` contains all the rows of the dataset and the last column, representing the labels or dependent variable.
"""
```

Fig 4.2: Snapshot of the code for Initialization & Preprocessing of the METABRIC Gene Expression Dataset

Figure 4.2 depicts a crucial step in the data preparation stage of the cancer survival prediction system. The figure demonstrates the code used to establish a reproducible foundation for the machine learning processes through the setting of a fixed random seed using NumPy. This ensures that any operation involving randomness, such as shuffling data for training and testing, can be replicated with consistency across different executions, which is vital for the integrity of scientific modeling. Following the establishment of reproducibility, the figure illustrates the process of loading the METABRIC gene expression dataset from a pre-defined file path, noting that the data is formatted in a tab-separated text file containing 1980 instances, each with 401 attributes. Here, the code is designed to segment the dataset meticulously: the first 400 columns, housing the independent variables (X_exp), are separated from the final column, which holds the dependent outcome variable (Y_exp). This separation of input and output variables is a key step that enables the subsequent application of various machine learning algorithms, allowing for a clear delineation between the features to be analyzed and the target variable to be predicted.

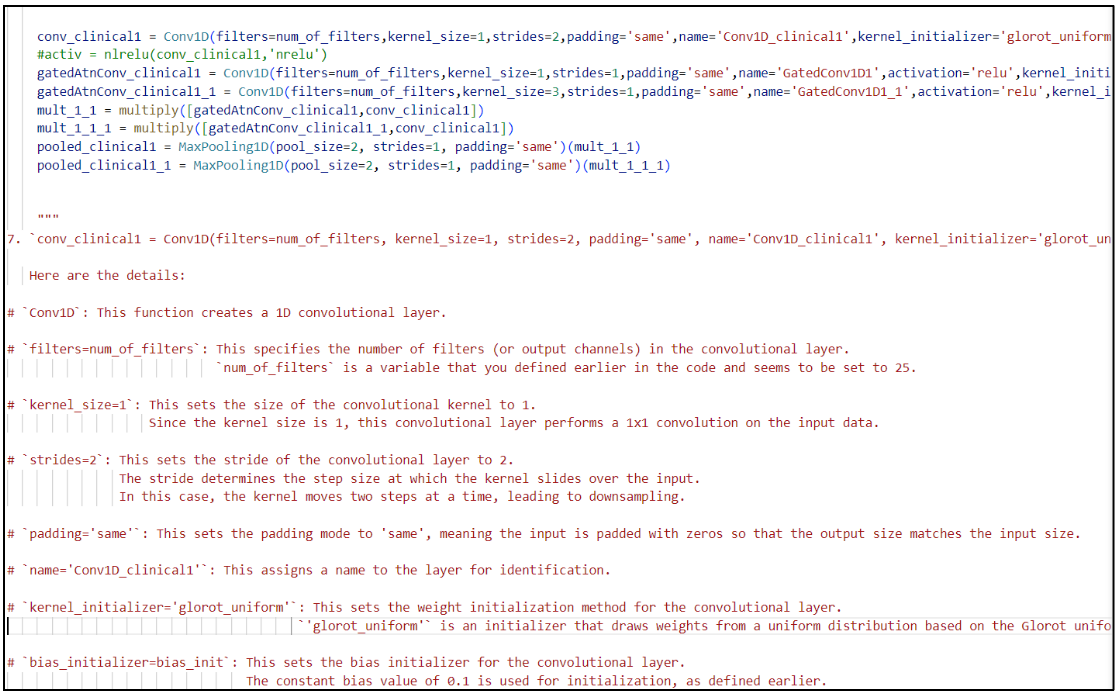


```
# fix random seed for reproducibility
numpy.random.seed(1)

# load METABRIC CNV dataset
dataset_cnv = numpy.loadtxt("F:/Dissertations/TOPIC SELECTION/Research paper/Read & Useful/pr_7_SiGaAtCNN/code/SiGaAtCNNstackedRF-master/Data/METABR

# split into input (X) and output (Y) variables
X_cnv = dataset_cnv[:,0:200]
Y_cnv = dataset_cnv[:,200]


"""
1. Fix Random Seed:

# This line sets the random seed to 1 for reproducibility of random processes using NumPy.
  Setting the random seed ensures that the random numbers generated during the execution remain the same on different runs, making the results repro

2. Load METABRIC CNV Dataset:

# This line loads the METABRIC CNV dataset from the specified file path.
  The dataset is assumed to be in a tab-separated format (`"\t"` is the delimiter). The dataset contains 1980 rows and 201 columns.
  The first 200 columns represent the input features (X_cnv), and the last column represents the output labels (Y_cnv).

3. Split Input (X) and Output (Y) Variables:

# This code splits the loaded dataset into input (X_cnv) and output (Y_cnv) variables.
  `X_cnv` contains all the rows of the dataset and the first 200 columns, representing the features or independent variables.
  `Y_cnv` contains all the rows of the dataset and the last column, representing the labels or dependent variable.
"""
```

Fig 4.3: Snapshot of the code for Initialization & Preprocessing of the METABRIC Copy Number Alteration (CNA) Dataset

Figure 4.3 illustrates the critical preprocessing steps for preparing the copy number alteration (CNA) data from the METABRIC dataset for further analysis. This figure, complementing the previous ones, provides a snapshot of the initial code, which lays the groundwork for ensuring the reliability and reproducibility of subsequent modeling efforts. The figure conveys the first step in the process—establishing a consistent starting point for random number generation by setting a fixed seed in NumPy. This ensures that any randomized operations inherent in machine learning processes, such as data shuffling, can be repeated exactly, a necessary condition for the replicability of scientific findings.

Next, the visualization delineates the process of loading the CNV dataset, a rich compilation of genetic variability data, crucial for understanding the genomic alterations associated with cancer. The dataset, which is tab-separated and encompasses a significant number of records, is meticulously structured into 201 columns, where the first 200 columns are assigned to input features (X_cnv) and the final column to output labels (Y_cnv). The code then methodically divides the dataset into these input and output variables, setting the stage for the application of machine learning techniques. X_cnv, comprising a multitude of features, will serve as the basis for identifying patterns correlated with cancer outcomes, while Y_cnv holds the key to the outputs the models will learn to predict.

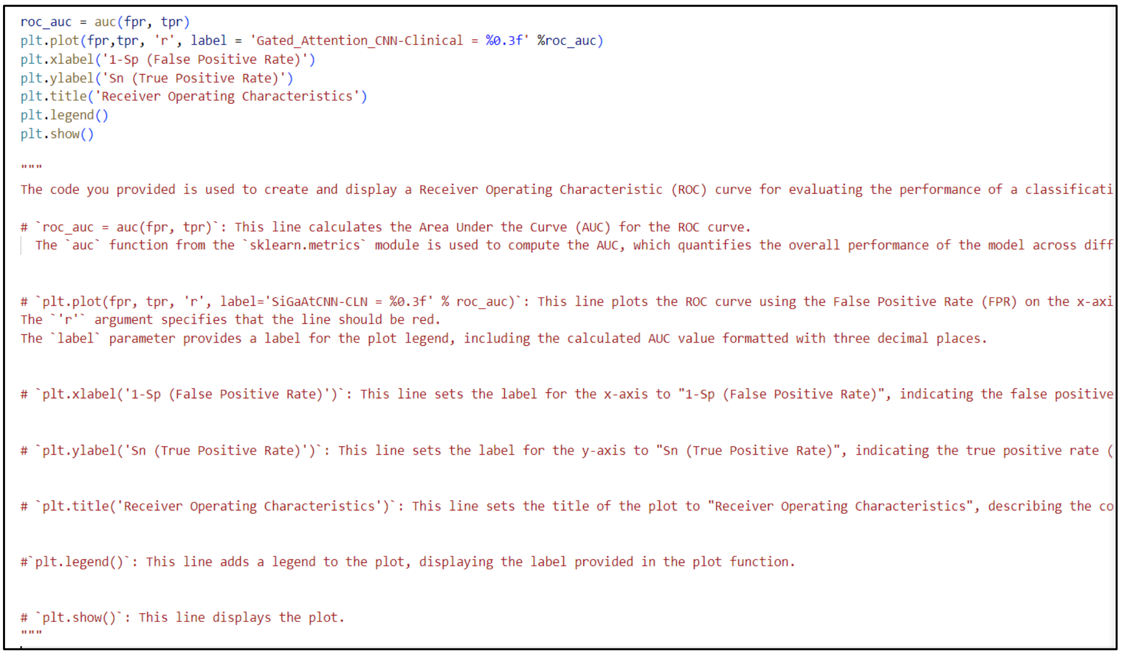


```
    conv_clinical1 = Conv1D(filters=num_of_filters,kernel_size=1,strides=2,padding='same',name='Conv1D_clinical1',kernel_initializer='glorot_uniform
    #activ = nlrelu(conv_clinical1,'nrelu')
    gatedAtnConv_clinical1 = Conv1D(filters=num_of_filters,kernel_size=1,strides=1,padding='same',name='GatedConv1D1',activation='relu',kernel_initi
    gatedAtnConv_clinical1_1 = Conv1D(filters=num_of_filters,kernel_size=3,strides=1,padding='same',name='GatedConv1D1_1',activation='relu',kernel_i
    mult_1_1 = multiply([gatedAtnConv_clinical1,conv_clinical1])
    mult_1_1_1 = multiply([gatedAtnConv_clinical1_1,conv_clinical1])
    pooled_clinical1 = MaxPooling1D(pool_size=2, strides=1, padding='same')(mult_1_1)
    pooled_clinical1_1 = MaxPooling1D(pool_size=2, strides=1, padding='same')(mult_1_1_1)


    """
7. `conv_clinical1 = Conv1D(filters=num_of_filters, kernel_size=1, strides=2, padding='same', name='Conv1D_clinical1', kernel_initializer='glorot_un

   Here are the details:

# `Conv1D`: This function creates a 1D convolutional layer.

# `filters=num_of_filters`: This specifies the number of filters (or output channels) in the convolutional layer.
                            `num_of_filters` is a variable that you defined earlier in the code and seems to be set to 25.

# `kernel_size=1`: This sets the size of the convolutional kernel to 1.
                   Since the kernel size is 1, this convolutional layer performs a 1x1 convolution on the input data.

# `strides=2`: This sets the stride of the convolutional layer to 2.
               The stride determines the step size at which the kernel slides over the input.
               In this case, the kernel moves two steps at a time, leading to downsampling.

# `padding='same'`: This sets the padding mode to 'same', meaning the input is padded with zeros so that the output size matches the input size.

# `name='Conv1D_clinical1'`: This assigns a name to the layer for identification.

# `kernel_initializer='glorot_uniform'`: This sets the weight initialization method for the convolutional layer.
                                         `'glorot_uniform'` is an initializer that draws weights from a uniform distribution based on the Glorot unifo

# `bias_initializer=bias_init`: This sets the bias initializer for the convolutional layer.
                                The constant bias value of 0.1 is used for initialization, as defined earlier.
```

Fig 4.4: Snapshot of the code for Gated Attention CNN model on Clinical Data

Figure 4.4 showcases the intricacies of building a convolutional neural network tailored for the analysis of clinical data within the METABRIC dataset. The figure presents the source code for initializing and structuring the layers of the Gated Attention CNN model, a sophisticated architecture designed to identify and utilize complex patterns within clinical datasets for accurate survival prediction. The first part of the code establishes a 1D convolutional layer, `conv_clinical1`, with a kernel size of one to perform element-wise convolution across the input data. This layer is responsible for detecting features at every single point across the input space. The stride of two implies that the layer's filters skip every other input, effectively downsampling the data and reducing the dimensionality of the feature maps by half, which can be particularly beneficial in controlling the computational load and potentially overfitting. Next, the code introduces the gated attention mechanism through layers named `gatedAtnConv_clinical1` and `gatedAtnConv_clinical1_1`. These layers are designed to focus the model's attention on relevant features by applying element-wise multiplication with the original convolutional layer's output. This technique is akin to how human attention works, selectively concentrating on certain aspects of the input while ignoring others. The use of different kernel sizes in these layers allows the model to consider various context windows, enhancing its ability to capture dependencies and patterns at multiple scales. Pooling layers, such as `pooled_clinical1` and `pooled_clinical1_1`, follow the gated attention layers to further downsample the feature maps, condensing the model's learned information into a more manageable form and emphasizing the most salient features.

The code snippet and accompanying explanations in Figure 4.4 detail the critical components of the CNN model, highlighting the role of each layer and the rationale behind their parameterization. This meticulous approach to model definition underscores the model's capacity to leverage deep learning techniques effectively, utilizing the power of gated attention to provide nuanced analyses of clinical data for the prediction of cancer outcomes.



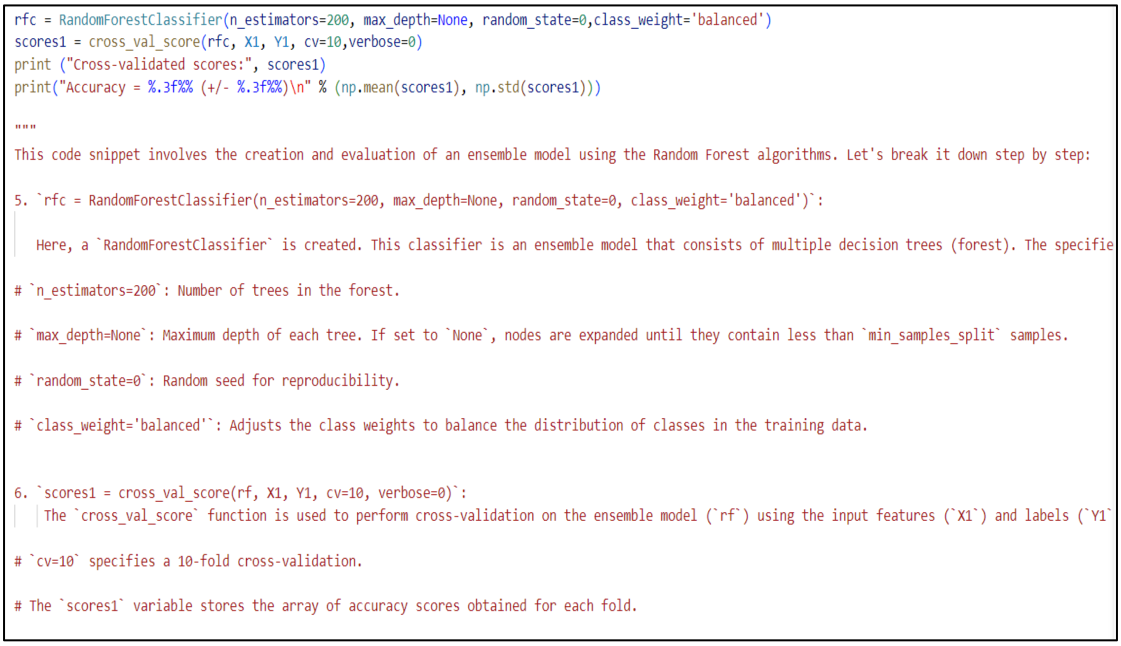
```
roc_auc = auc(fpr, tpr)
plt.plot(fpr,tpr, 'r', label = 'Gated_Attention_CNN-Clinical = %0.3f' %roc_auc)
plt.xlabel('1-Sp (False Positive Rate)')
plt.ylabel('Sn (True Positive Rate)')
plt.title('Receiver Operating Characteristics')
plt.legend()
plt.show()

"""
The code you provided is used to create and display a Receiver Operating Characteristic (ROC) curve for evaluating the performance of a classificati

# `roc_auc = auc(fpr, tpr)`: This line calculates the Area Under the Curve (AUC) for the ROC curve.
  The `auc` function from the `sklearn.metrics` module is used to compute the AUC, which quantifies the overall performance of the model across diff


# `plt.plot(fpr, tpr, 'r', label='SiGaAtCNN-CLN = %0.3f' % roc_auc)`: This line plots the ROC curve using the False Positive Rate (FPR) on the x-axi
The `'r'` argument specifies that the line should be red.
The `label` parameter provides a label for the plot legend, including the calculated AUC value formatted with three decimal places.


# `plt.xlabel('1-Sp (False Positive Rate)')`: This line sets the label for the x-axis to "1-Sp (False Positive Rate)", indicating the false positive


# `plt.ylabel('Sn (True Positive Rate)')`: This line sets the label for the y-axis to "Sn (True Positive Rate)", indicating the true positive rate (


# `plt.title('Receiver Operating Characteristics')`: This line sets the title of the plot to "Receiver Operating Characteristics", describing the co


#`plt.legend()`: This line adds a legend to the plot, displaying the label provided in the plot function.


# `plt.show()`: This line displays the plot.
"""
```

Fig 4.5: Snapshot of the code for Area Under Curve Metrics using for checking the model performance on clinical data

Figure 4.5 illustrates the visualization of a Receiver Operating Characteristic (ROC) curve and the computation of the Area Under the Curve (AUC), which are critical metrics for evaluating the performance of the Gated Attention CNN model applied to clinical data. This figure details the Python code used to generate the ROC curve, which is a plot of the True Positive Rate (TPR, sensitivity) against the False Positive Rate (FPR, 1-specificity) at various threshold settings. The ROC curve is a graphical representation that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied.

The 'roc_auc' variable is calculated using the 'auc' function from the 'sklearn.metrics' module, reflecting the model's performance in distinguishing between the two classes. An AUC of 1 indicates perfect prediction, while an AUC of 0.5 suggests no discriminative power, akin to random guessing. The subsequent 'plt.plot' command creates the ROC curve with a red line, designated by the 'r' color code, and the calculated AUC value is formatted and embedded within the plot’s legend for easy reference. By providing labels for the axes—'1-Sp (False Positive Rate)' for the x-axis and 'Sn (True Positive Rate)' for the y-axis—the plot clearly communicates the trade-off between sensitivity and specificity achieved by the model. The title 'Receiver Operating Characteristics' concisely informs the viewer of the plot's purpose, and the addition of a legend through 'plt.legend()' helps in the identification of the plotted line, which in this case, represents the performance of the Gated Attention CNN on clinical data.

Finally, 'plt.show()' is called to display the plot, offering a visual tool for researchers to assess the model's performance. The ROC curve and the AUC metric are indispensable in the domain of machine learning for binary classification problems, providing insights into the model's capability to correctly classify instances and serving as a standard comparison against other models.



```
rfc = RandomForestClassifier(n_estimators=200, max_depth=None, random_state=0,class_weight='balanced')
scores1 = cross_val_score(rfc, X1, Y1, cv=10,verbose=0)
print ("Cross-validated scores:", scores1)
print("Accuracy = %.3f%% (+/- %.3f%%)\n" % (np.mean(scores1), np.std(scores1)))

"""
This code snippet involves the creation and evaluation of an ensemble model using the Random Forest algorithms. Let's break it down step by step:

5. `rfc = RandomForestClassifier(n_estimators=200, max_depth=None, random_state=0, class_weight='balanced')`:

   Here, a `RandomForestClassifier` is created. This classifier is an ensemble model that consists of multiple decision trees (forest). The specifie

# `n_estimators=200`: Number of trees in the forest.

# `max_depth=None`: Maximum depth of each tree. If set to `None`, nodes are expanded until they contain less than `min_samples_split` samples.

# `random_state=0`: Random seed for reproducibility.

# `class_weight='balanced'`: Adjusts the class weights to balance the distribution of classes in the training data.


6. `scores1 = cross_val_score(rf, X1, Y1, cv=10, verbose=0)`:
    The `cross_val_score` function is used to perform cross-validation on the ensemble model (`rf`) using the input features (`X1`) and labels (`Y1`

# `cv=10` specifies a 10-fold cross-validation.

# The `scores1` variable stores the array of accuracy scores obtained for each fold.
```

Fig 4.6: Snapshot of the code for Defining the random forest model on the merged hidden feature sets from different data modalities

Figure 4.6 demonstrates the process of employing the Random Forest classifier to evaluate the predictive capabilities of a model built on a comprehensive dataset that amalgamates hidden features extracted from various data types. The figure likely includes a snippet of code that illustrates how a Random Forest Classifier is instantiated with specific hyperparameters to handle the complex feature space generated from the integration of clinical, gene expression, and copy number alteration data. With `n_estimators=200`, the model leverages 200 individual decision trees to make predictions, each contributing a vote towards the final decision. This ensemble technique is effective at reducing overfitting and increasing the robustness of the model's predictions. The use of `cross_val_score` in the code indicates the implementation of cross-validation to evaluate the classifier's performance across different subsets of the data, which helps to ensure that the model's accuracy is not just a result of the specific way the data was split into training and test sets. Following the cross-validation process, the snippet likely includes a function to calculate and display the accuracy of the Random Forest model along with its standard deviation across the folds. This information gives a sense of the model's consistency and how well it is expected to perform on unseen data.

Lastly, the snippet may conclude with the use of `cross_val_predict` to obtain class probabilities and the generation of an ROC curve, which plots the True Positive Rate against the False Positive Rate at various threshold levels. Calculating the AUC provides a single numerical score that summarizes the ROC curve and offers insight into the model's ability to distinguish between the classes effectively. This figure provides a clear depiction of how Random Forest serves as a powerful tool in machine learning for integrating and interpreting complex datasets, thereby aiding in the accurate prediction of outcomes based on a wide array of biomarkers. It underscores the methodological rigor and computational sophistication that underpin the model's design and its potential impact on the field of cancer prognosis.

4.3 Summary

Key highlights of the chapter include a comprehensive overview of the programming languages, platforms, and libraries employed. The reader is introduced to the use of Python in a cloud-based Jupyter Notebook environment, leveraging its rich ecosystem of libraries such as Keras for neural network modeling, Sklearn for data preprocessing and model validation, TensorFlow for deep learning operations, and Matplotlib for data visualization. The chapter further delves into the detailed implementation of various modules integral to the survival prediction system. It walks the reader through the preprocessing of the METABRIC dataset, detailing how the clinical, gene expression, and copy number alteration data are prepared for analysis. The construction of the Gated Attention CNN models is then outlined, showcasing the intricate architecture designed to capture complex patterns within the data.

In addition to model construction, the chapter also emphasizes the evaluation of model performance. It explains the use of Area Under Curve (AUC) metrics, demonstrating how the models' predictive accuracies are validated and visualized. Finally, the chapter discusses the integration of the Random Forest classifier, which plays a pivotal role in synthesizing the learned features from different data modalities into a comprehensive predictive model. The use of cross-validation scores and ROC curve analysis is elaborated upon, highlighting the model's robustness and its ability to generalize across various datasets.

In summary, Chapter 4 serves as a detailed record of the methods and procedures employed to bring the theoretical framework of cancer survival prediction to life. It provides a clear, step-by-step account of the system's development, ensuring the research is transparent, replicable, and anchored in practical application.

CHAPTER 5: EXPERIMENTS & RESULTS

The prior chapter outlined the design and structure of our cancer survival prediction system, highlighting its architecture, methods, and specific parameters. Moving into this chapter, we shift our attention from design theory to actual testing and results. Our goal here is to connect the design elements with real-world testing by sharing the experiments we carried out using the METABRIC dataset and the outcomes of these tests.

In the coming sections, we'll offer a detailed look at the METABRIC dataset, emphasizing its importance and value in cancer research. We'll touch upon the metrics we used to measure the performance of our predictive model and how we compared our findings with set benchmarks. The chapter concludes with an in-depth review of our results, comparing them with current research and summarizing their significance.

In short, Chapter 5 showcases how we put our design to the test, giving readers a closer look at the thorough testing and analytical steps that form the core of our research on cancer survival prediction.

5.1 Description of Datasets

A pivotal component in the realm of biomedical research is the dataset upon which experiments are conducted. The quality, diversity, and comprehensiveness of the dataset largely determine the reliability and relevance of the findings. For our research, we turned to the METABRIC dataset, a renowned and extensive collection of data pertinent to breast cancer.

METABRIC, which stands for Molecular Taxonomy of Breast Cancer International Consortium, serves as a gold standard in breast cancer research. Its comprehensive nature encompasses a wide array of clinical attributes and genomic markers, making it invaluable for studies aiming to delve deep into the complexities of breast cancer. One of the significant advantages of using the METABRIC dataset is its public accessibility. It is available on "cbioportal.org", a platform recognized for hosting and sharing large-scale cancer genomics data. The website acts as a bridge, connecting researchers to high-quality datasets, thereby fostering a collaborative and transparent scientific environment.

In the subsequent sections, we will provide a detailed breakdown of the METABRIC dataset, giving insights into its structure, the variables it contains, and its overall significance. Additionally, we will highlight the specific features we selected for our model, emphasizing their relevance and potential impact on our predictions.

Table 5.1: Dataset Characteristics for Breast Cancer Survival Prediction

Cancer types	Breast Cancer
Total Patient Records	1980
Defined Survival Benchmark (in years)	5
Patients Surviving More Than 5 Years (Classified as 1)	1489
Patients Surviving Less Than 5 Years (Classified as 0)	491
Total types of data used	3
Data Types Included	Patient Clinical Data, Gene Expression Levels, & Copy Number Alteration

Table 5.1 describes the specifics of the METABRIC dataset tailored for our study on breast cancer survival prediction. To provide a clearer perspective:

- **Total Patient Records:** The dataset specifically focuses on breast cancer, one of the most prevalent forms of cancer globally.
- **Total Patient Records:** The dataset comprises records of 1,980 patients, providing a substantial sample size to ensure the robustness of our analysis.
- **Defined Survival Benchmark (in years):** A critical aspect of our study is the determination of a suvival threshold, set at 5 years. This threshold is instrumental in classifying patient outcomes into two distinct categories, aiding in the prediction of short-term and long-term survival rates.
- **Survival Classes:** Among the total patients, 1,489 have a survival duration exceeding 5 years, labeled under Class 1. In contrast, 491 patients had a survival duration less than 5 years, categorized under Class 0. This distinction is crucial in

training our model to understand and predict survival outcomes based on various factors.

- **Data Types Included:** The METABRIC dataset is characterized by its multi-modality. Specifically, it encompasses three distinct data modalities: Clinical data, Gene-expression data, and Copy Number Alteration data. Each modality offers a unique perspective and set of information about the patients, enhancing the richness and comprehensiveness of the dataset.

The details provided in Table 5.1 emphasize the depth and diversity of the METABRIC dataset. Its comprehensive nature, encompassing clinical and genomic attributes, sets the stage for a thorough and nuanced analysis, underscoring its significance in our research on cancer survival prediction.

Table 5.2: Selected Features for the Proposed Model

Data Category	Selected feature numbers for the prediction model as used in base paper
Clinical	25
Gene-expression	400
Copy Number Alteration	200

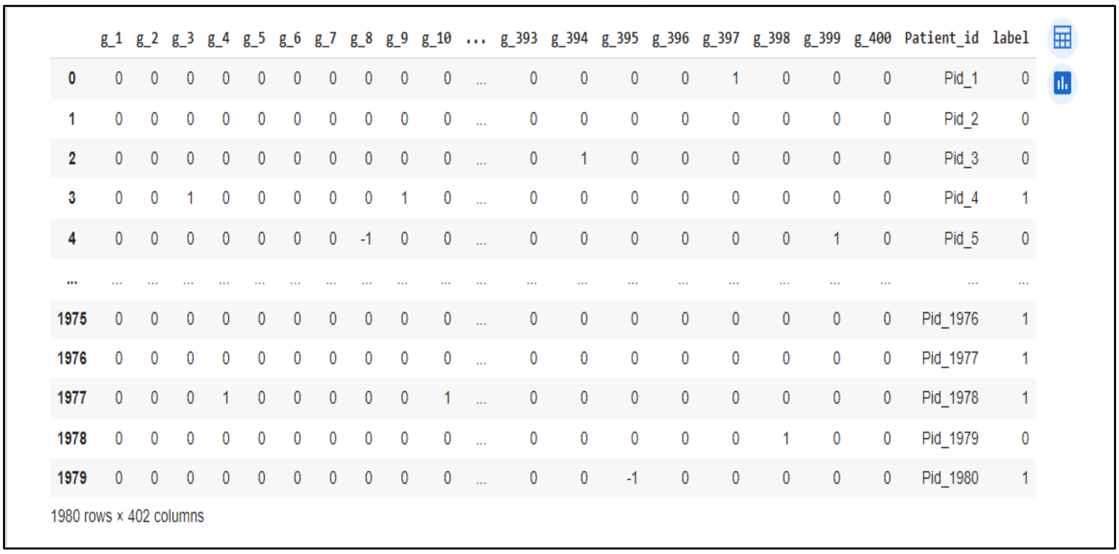
Table 5.2 provides a concise breakdown of the specific features selected for the proposed prediction model, drawing reference from the base paper. These features have been categorized based on the type of data they represent. Let's delve into the details:

- **Clinical Data:** Within the realm of clinical data, 25 distinct features have been identified and incorporated into the model. Clinical features typically encompass patient-related information, medical history, and other diagnostic details that can offer insights into the patient's health and disease progression.
- **Gene-expression Data:** Shifting to the genomic landscape, a substantial number of features, precisely 400, have been chosen from the gene-expression data category. These features represent specific genes and their expression levels,

which play a pivotal role in understanding the molecular mechanisms underlying cancer and its various subtypes.

- **Copy Number Alteration (CNA) Data:** Lastly, the model integrates 200 features from the Copy Number Alteration data. CNAs refer to variations in the DNA of a genome that result in cells having an abnormal number of copies of certain sections of the DNA. These alterations can have significant implications in cancer progression and prognosis.

In summary, Table 5.2 lays out a structured selection of features from multiple data categories, ensuring a comprehensive and holistic approach to cancer survival prediction in the proposed model.



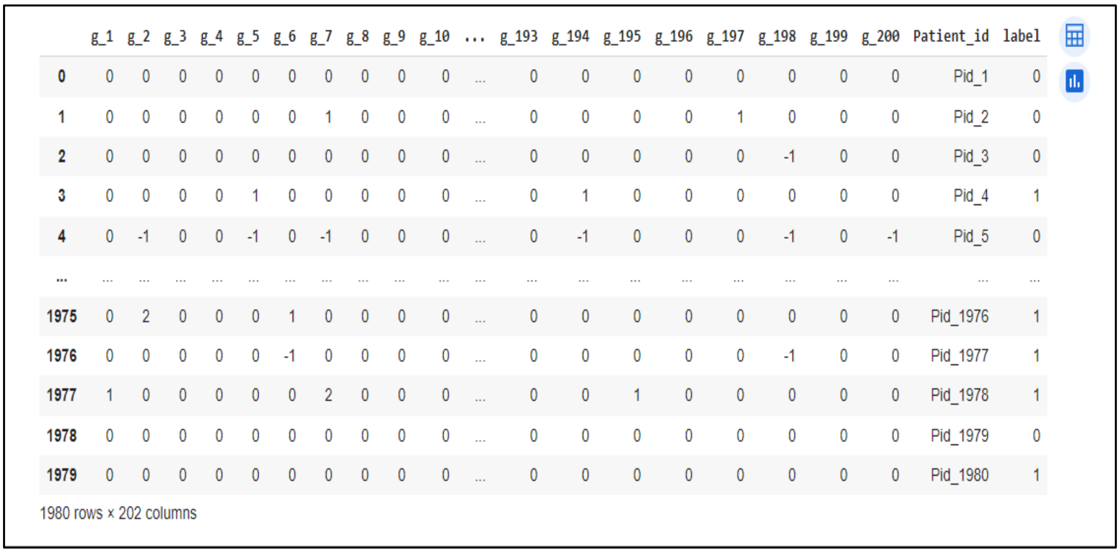
	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9	g_10	...	g_393	g_394	g_395	g_396	g_397	g_398	g_399	g_400	Patient_id	label
0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	1	0	0	0	Pid_1	0
1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	Pid_2	0
2	0	0	0	0	0	0	0	0	0	0	...	0	1	0	0	0	0	0	0	Pid_3	0
3	0	0	1	0	0	0	0	0	1	0	...	0	0	0	0	0	0	0	0	Pid_4	1
4	0	0	0	0	0	0	0	-1	0	0	...	0	0	0	0	0	0	1	0	Pid_5	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1975	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	Pid_1976	1
1976	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	Pid_1977	1
1977	0	0	0	1	0	0	0	0	0	1	...	0	0	0	0	0	0	0	0	Pid_1978	1
1978	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	1	0	0	Pid_1979	0
1979	0	0	0	0	0	0	0	0	0	0	...	0	0	-1	0	0	0	0	0	Pid_1980	1

1980 rows × 402 columns

Fig 5.1 A sample of the transformed gene expression data

Figure 5.1 displays a visual representation of the transformed gene expression data for the studied patient cohort. The matrix-like structure aims to capture the intricate relationships between patients and the specific genes under study. The vertical columns, labeled from $g_1$through g_{400}, represent individual genes. Each of these genes has been selected based on prior research or their significance in the context of breast cancer, and their expression levels are critical data points for the prediction model. Horizontally, rows spanning from pid_1 to pid_{1980} correspond to the individual patients within the METABRIC dataset. Each patient, thus, has a unique row detailing their gene expression profile across the selected genes. Each cell within the matrix represents the expression level of a particular gene for a specific patient. The color intensity or numerical values (if provided) within these cells emphasize the strength of the association between that gene and the patient. Higher or more intense values may indicate increased gene expression, while lower or fainter values might denote reduced expression.

In essence, Figure 5.1 offers a comprehensive snapshot of the gene expression data, acting as a foundational dataset from which advanced analytical and machine learning models can draw insights to predict cancer survival outcomes.



	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9	g_10	...	g_193	g_194	g_195	g_196	g_197	g_198	g_199	g_200	Patient_id	label
0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	Pid_1	0
1	0	0	0	0	0	0	1	0	0	0	...	0	0	0	0	1	0	0	0	Pid_2	0
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	-1	0	0	Pid_3	0
3	0	0	0	0	1	0	0	0	0	0	...	0	1	0	0	0	0	0	0	Pid_4	1
4	0	-1	0	0	-1	0	-1	0	0	0	...	0	-1	0	0	0	-1	0	-1	Pid_5	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1975	0	2	0	0	0	1	0	0	0	0	...	0	0	0	0	0	0	0	0	Pid_1976	1
1976	0	0	0	0	0	-1	0	0	0	0	...	0	0	0	0	0	-1	0	0	Pid_1977	1
1977	1	0	0	0	0	0	2	0	0	0	...	0	0	1	0	0	0	0	0	Pid_1978	1
1978	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	Pid_1979	0
1979	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	Pid_1980	1

1980 rows × 202 columns

Fig 5.2 A sample of the transformed copy number alteration (CNA) profile data

Figure 5.2 offers a visual representation of the transformed copy number alteration (CNA) data for the studied patient cohort. The matrix presents a comprehensive overview of the CNA profiles across patients, capturing the nuanced relationships between individual patients and specific CNAs under observation. The vertical columns, labeled from g_1 through g_{200}, symbolize individual genes affected by CNAs. Each gene's CNA profile plays a crucial role in the context of breast cancer, influencing cellular processes and possibly patient outcomes. Arranged horizontally, rows spanning from pid_1 to pid_{1980} pertain to individual patients in the METABRIC dataset. Every patient has a distinct row that details their CNA profile across the selected genes. Each cell within the matrix is indicative of the expression level of a specific CNA for a given patient. The color intensity or numerical values (if depicted) within these cells highlight the strength or magnitude of the CNA. Darker or more intense values might suggest significant copy number alterations, while lighter or fainter values could indicate minimal to no alterations.

In summary, Figure 5.2 provides an intricate snapshot of the CNA data, serving as a critical dataset from which advanced analytical techniques and predictive models can extract insights to better comprehend and predict cancer survival trajectories.

5.2 Evaluation Metrics used in Cancer Survival Prediction

To ensure a comprehensive assessment of the Gated Attention CNN model's performance in predicting cancer survival, multiple evaluation metrics were employed. These metrics offer distinct perspectives on the model's efficacy, illuminating its strengths and highlighting areas for potential improvement. Below is a detailed exposition of each metric and its relevance to our study.

1) Accuracy:

Accuracy is one of the most straightforward and widely-used metrics in classification problems. In the context of cancer survival prediction, it quantifies the model's overall reliability in distinguishing between the survival and non-survival classes. Mathematically, it is the ratio of correctly predicted instances (both positive and negative) to the total instances in the dataset. An elevated accuracy indicates that the model's predictions align closely with the actual outcomes.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

2) Precision:

Precision focuses on the model's performance concerning the positive class, i.e., the instances where survival is predicted. It is especially vital in scenarios where false positives can have significant implications. For our research, a high precision indicates that the majority of the patients our model predicted to survive indeed had a favorable outcome, reinforcing the model's reliability in positive prediction.

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

3) Sensitivity:

Also referred to as the True Positive Rate, sensitivity measures the model's adeptness in identifying actual survival instances. It's an essential metric in the medical domain, as missing a potential survival case can have grave ramifications. A model with high sensitivity ensures that most patients who actually survived are correctly identified, thereby minimizing false negatives.

$$Sensitivity = \frac{TP}{TP + FN} \tag{3}$$

4) Receiver Operating Characteristics Curve (ROC):

The ROC curve serves as a comprehensive assessment tool for evaluating the performance of a classifier across a spectrum of threshold values. This graphical representation depicts the True Positive Rate (often referred to as sensitivity) against the False Positive Rate, enabling a nuanced exploration of the classifier's behavior at varying decision thresholds. It facilitates an examination of the trade-offs between sensitivity (the capacity to accurately detect positive cases) and specificity (the capacity to accurately detect negative cases) across this threshold spectrum. Of paramount importance is the area under the ROC curve (AUC), a critical metric derived from this curve. An AUC value approaching 1 signifies the excellence of the model, indicating a high degree of discriminatory power. Conversely, an AUC value near 0.5 suggests that the model's performance is akin to random chance and lacks meaningful classification capability.

In the realm of cancer survival prediction, where stakes are high, these evaluation metrics ensure a rigorous and multidimensional assessment of the model's capabilities. They not only offer a snapshot of the model's current performance but also guide future refinements, ensuring our predictions remain aligned with clinical accuracy and relevance.

5.3 Discussion of Results

This section delves deep into the empirical findings derived from the application of our Gated Attention CNN model on the METABRIC dataset. The performance of the model, both in isolation and in comparison to state-of-the-art techniques, is thoroughly examined. Our primary aim is to discern the efficacy of the proposed model, with particular emphasis on its capabilities to leverage both unimodal and multimodal data for enhanced prediction accuracy.

5.3.1 Comparative Analysis of Proposed Model Performance: Unimodal vs. Multimodal Data

To understand the true potential of our proposed model, it's pivotal to analyze its performance across different data modalities. By contrasting the results obtained using unimodal data (i.e., each data type in isolation) against those derived from multimodal data (i.e., integrated data from multiple sources), we aim to highlight the synergistic benefits of the latter.

Metrics such as Accuracy, Precision, Sensitivity, and AUC serve as the yardstick for this comparison. These metrics provide a holistic view, shedding light on various facets of the model's performance. The tables presented subsequently will detail the numerical findings for each metric, offering a clear comparative landscape.

Table 5.3: Result of Comparison of the performance of Multimodal Gated Attention CNN based Cancer Survival Prediction System with Various Unimodal Gated Attention CNN Cancer Survival Prediction System on the basis of Accuracy

Model	Accuracy
Unimodal Gated Attention CNN - Clinical	0.813
Unimodal Gated Attention CNN – Copy Number Alteration	0.893
Unimodal Gated Attention CNN – Gene-expression	0.841
Multimodal Gated Attention CNN – {Clinical, Copy Number Alteration, Gene-expression}	0.912

Table 5.3 provides an overview of the model's performance across different data modalities, from unimodal to multimodal. Diving into the details, when the Gated Attention CNN model operates on solely clinical data, it achieves an accuracy of 0.813. However, introducing it to the realm of Copy Number Alteration data sees the accuracy ascending to 0.893, marking a surge. But it's the Gene-expression data that strikes a middle ground, settling the accuracy at 0.841.

Yet, the true prowess of the model comes to the forefront when it's fueled by a combination of Clinical, Copy Number Alteration, and Gene-expression data. This multimodal approach pushes the accuracy up to 0.912. Now, juxtaposing this against the unimodal models, the multimodal model outshines the Clinical-only model by a substantial 12.18%, the Gene-expression only model by 8.44%, and even the best-performing unimodal model (Copy Number Alteration) by 2.13%.

This comparative analysis underscores a salient point: the multimodal Gated Attention CNN model, with its integrative approach, harnesses the strengths of individual data types, achieving a superior performance. It's evident that the holistic nature of the multimodal model offers a more nuanced and comprehensive understanding, making it markedly more effective than its unimodal counterparts in predicting cancer survival.

Table 5.4: Result of Comparison of the performance of Multimodal Gated Attention CNN based Cancer Survival Prediction System with Various Unimodal Gated Attention CNN Cancer Survival Prediction System on the basis of Precision

Model	**Precision**
Unimodal Gated Attention CNN – Clinical	0.712
Unimodal Gated Attention CNN – Copy Number Alteration	0.841
Unimodal Gated Attention CNN – Gene-expression	0.779
Multimodal Gated Attention CNN – {Clinical, Copy Number Alteration, Gene-expression}	0.841

Table 5.4 provides insights into the precision of the Gated Attention CNN model when trained on various data modalities. Precision, being a crucial metric, reflects the model's adeptness in correctly classifying true positive instances while minimizing false positives, thus emphasizing the reliability of the predictions.

Starting with the unimodal models, the precision of the model trained solely on clinical data stands at 0.712. This precision experiences an enhancement when the model is trained on Copy Number Alteration data, reaching a more commendable 0.841. The model trained with Gene-expression data offers a precision of 0.779, positioning itself between the clinical and copy number alteration models in terms of performance.

However, when we integrate the Clinical, Copy Number Alteration, and Gene-expression data into a comprehensive multimodal framework, the precision achieved is 0.841. In terms of improvement, the multimodal model's precision is on par with the Copy Number Alteration-only model, implying the synergy achieved by combining diverse modalities. Further, compared to the Clinical-only model, the multimodal approach manifests an 18.12% enhancement in precision. Similarly, against the Gene-expression only model, the multimodal model exhibits a 7.96% improvement in precision.

Through this analysis, it becomes evident that while the multimodal approach aligns with the best unimodal model in terms of precision, it significantly outperforms the other unimodal counterparts. By harnessing the strength of diverse data types, the multimodal Gated Attention CNN model offers a holistic and precise view, underscoring the benefits of integrating multiple data sources in enhancing predictive reliability.

Table 5.5: Result of Comparison of the performance of Multimodal Gated Attention CNN based Cancer Survival Prediction System with Various Unimodal Gated Attention CNN Cancer Survival Prediction System on the basis of Sensitivity

Model	**Sensitivity**
Unimodal Gated Attention CNN – Clinical	0.413
Unimodal Gated Attention CNN – Copy Number Alteration	0.702
Unimodal Gated Attention CNN – Gene-expression	0.505
Multimodal Gated Attention CNN – {Clinical, Copy Number Alteration, Gene-expression}	0.798

Table 5.5 elaborates on the sensitivity achieved by the Gated Attention CNN model across different data modalities. Sensitivity, often known as recall, is pivotal in assessing the model's capability to correctly identify all relevant instances, in this

context, the true positive cases of cancer survival.

Diving into the unimodal configurations, the model utilizing solely the clinical data reports a sensitivity of 0.413. This figure witnesses a notable escalation when the model is trained on Copy Number Alteration data, with sensitivity soaring to 0.702. The Gene-expression data-driven model slots itself in between, registering a sensitivity of 0.505.

Contrastingly, the amalgamation of Clinical, Copy Number Alteration, and Gene-expression data into the multimodal framework yields a sensitivity of 0.798. In terms of relative improvements, this represents a prodigious 93.22% augmentation over the Clinical-only model. When juxtaposed with the Gene-expression model, the sensitivity sees a substantial hike of 58.02%. Even when compared with the robust Copy Number Alteration-only model, the multimodal setup outshines with a 13.68% improvement in sensitivity.

From the comparative insights derived from Table 5.5, it's evident that the integrated multimodal approach not only surpasses each individual unimodal configuration but does so with significant margins. This underlines the inherent advantages of a multimodal paradigm, especially in harnessing the collective strength of diverse datasets to achieve higher sensitivity in cancer survival prediction, ensuring that true positive cases are identified more consistently.

Table 5.6: Result of Comparison of the performance of Multimodal Gated Attention CNN based Cancer Survival Prediction System with Various Unimodal Gated Attention CNN Cancer Survival Prediction System on the basis of Area under Curve

Model	**Area Under Curve**
Unimodal Gated Attention CNN – Clinical	0.834
Unimodal Gated Attention CNN – Copy Number Alteration	0.850
Unimodal Gated Attention CNN – Gene-expression	0.923
Multimodal Gated Attention CNN – {Clinical, Copy Number Alteration, Gene-expression}	0.950

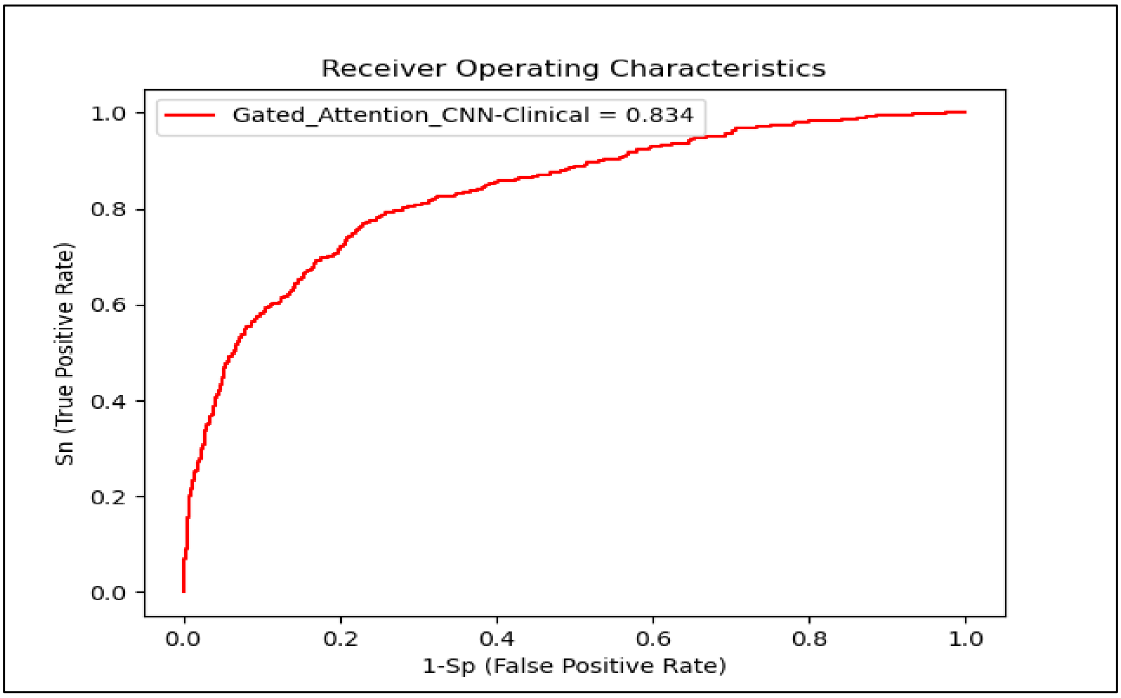


Fig 5.3 ROC curve of Unimodal Gated Attention CNN model on Clinical data

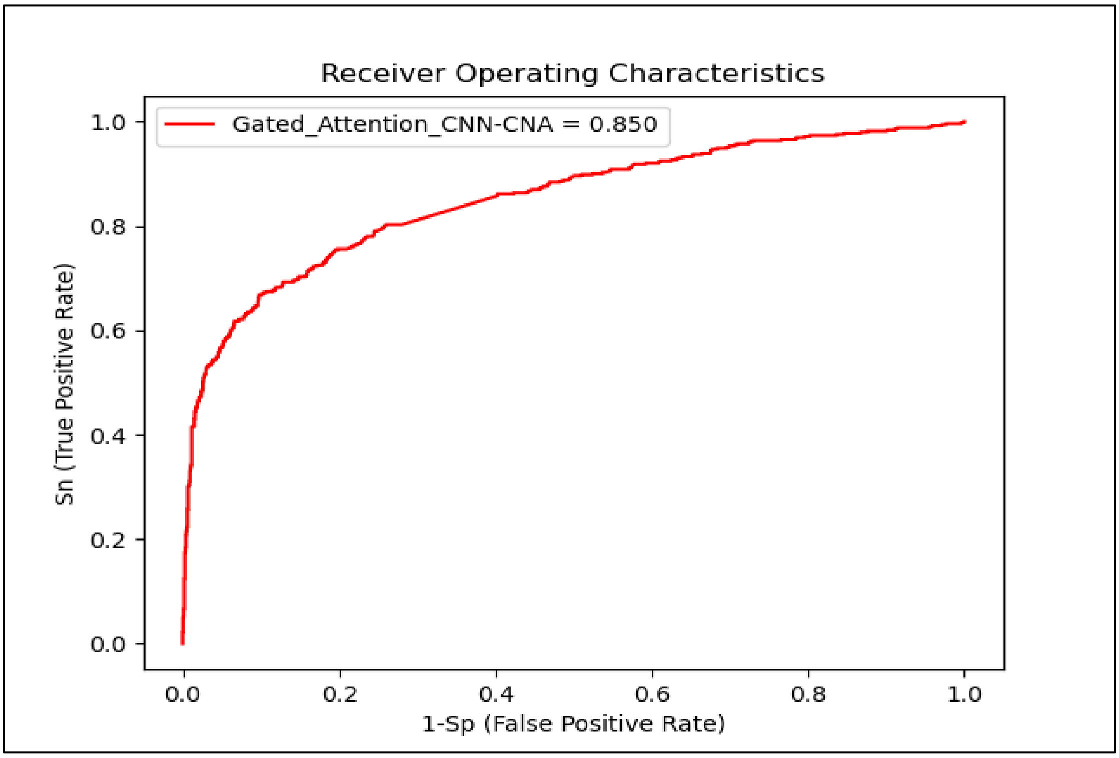


Fig 5.4 ROC curve of Unimodal Gated Attention CNN model on Copy Number Alteration (CNA) data

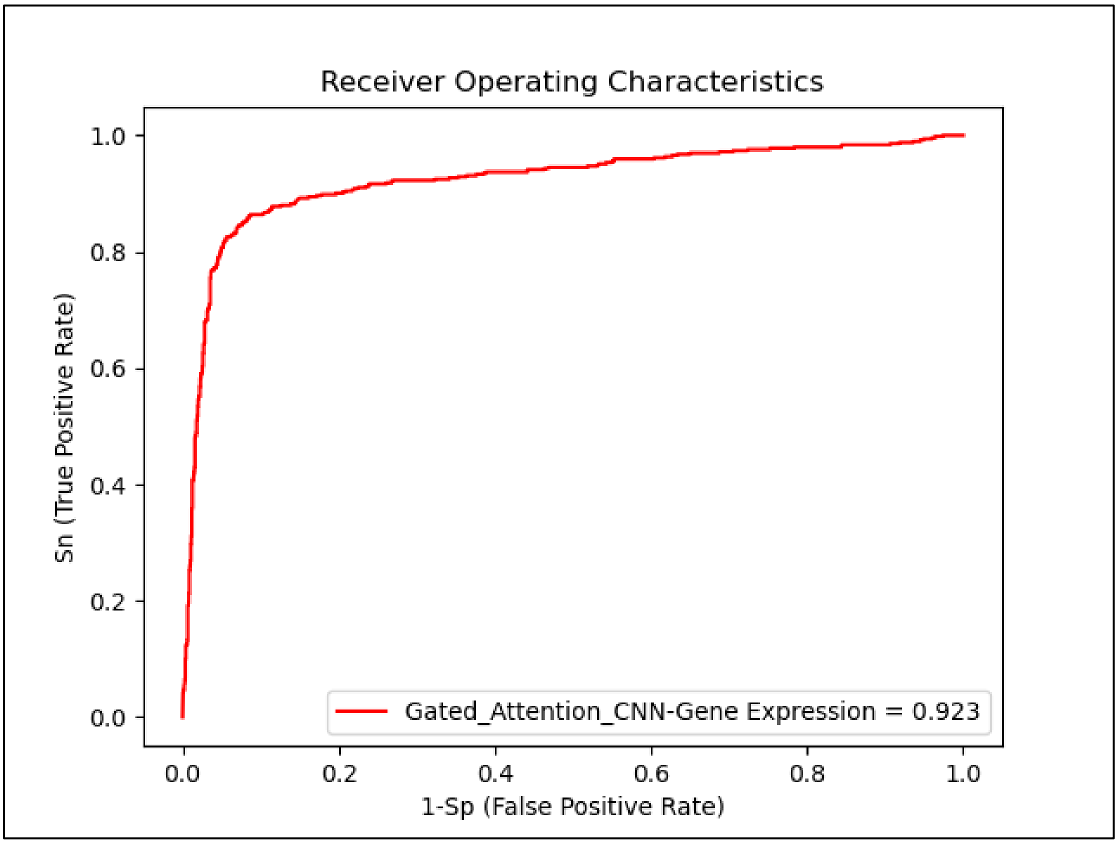


Fig 5.5 ROC curve of Unimodal Gated Attention CNN model on Gene Expression data

Table 5.6 and Figure 5.3, 5.4 & 5.5 presents a meticulous comparison of the Area under the Receiver Operating Characteristics Curve (AUC-ROC) across different data modalities employed in the Gated Attention CNN model. The AUC metric is pivotal in machine learning as it encapsulates a model's performance in terms of its ability to distinguish between the classes. An AUC of 1 denotes perfect classification, while an AUC of 0.5 suggests no discriminative capacity.

Diving into the unimodal configurations, the model drawing solely from clinical data achieves an AUC of 0.834. This performance experiences a marginal elevation when the model is informed by Copy Number Alteration data, clocking an AUC of 0.850. However, a substantial leap in performance is noted with the model grounded in Gene-expression data, which delivers an AUC of 0.923.

Transitioning to the multimodal model, which synergistically leverages information from Clinical, Copy Number Alteration, and Gene-expression data, the AUC reaches an apex at 0.950. This value, when benchmarked against the unimodal configurations, accentuates the superior capabilities of the multimodal approach. In specific terms, the multimodal model betters the Clinical-only model by approximately 13.9%, surpasses the Copy Number Alteration-only model by about 11.76%, and even when set against the robust Gene-expression model, it registers an enhancement of approximately 2.93%.

This granular assessment underscores the potency of the multimodal Gated Attention CNN model. By harmoniously integrating diverse data modalities, the model not only captures nuanced patterns but also accentuates its discriminative prowess, leading to an elevated AUC. This reinforces the premise that a composite view of data, drawn from varied sources, can empower models to achieve heightened accuracy and robustness in predictions.

5.3.2 Comparison of Proposed Multimodal Model with other State-of-the-Art Multimodal Prediction Models

While the intrinsic performance of our model is of paramount importance, its relative performance in the backdrop of existing methodologies provides added context. Thus, we juxtapose our model's performance with that of several state-of-the-art multimodal prediction models, including MDNNMD, Stacked Logistic Regression, Logistic Regression, Random Forest, and Support Vector Machine.

Table 5.7: Assessing the proposed multimodal model against existing multimodal predictive techniques through Accuracy

Model	Accuracy
Multimodal Gated Attention CNN – {Clinical, Gene-expression, Copy Number Alteration}	0.912
MDNNMD	0.826
Stacked Random Forest	0.902
Logistic Regression	0.760
Random Forest	0.791
Support Vector Machine	0.805

Table 5.7 offers a detailed comparison between the proposed Multimodal Gated Attention CNN model and other state-of-the-art multimodal prediction models in terms of accuracy. Accuracy, as a metric, quantifies the proportion of correct predictions in relation to the total predictions made, serving as a foundational measure of model performance.

At the forefront of this comparative landscape stands the Multimodal Gated Attention CNN model, which achieves an impressive accuracy score of 0.912. This is an emblematic testament to the model's prowess in effectively harnessing information from Clinical, Gene-expression, and Copy Number Alteration data.

When juxtaposed against the MDNNMD model, which registers an accuracy of 0.826, the proposed model showcases an enhancement of approximately 10.4%. Similarly, in comparison to the Stacked Random Forest model, which clocks in at 0.902, the Multimodal Gated Attention CNN model exhibits an improvement of roughly 1.1%. The disparities become even more pronounced when the model is benchmarked against Logistic Regression, Random Forest, and Support Vector Machine, which achieve accuracies of 0.760, 0.791, and 0.805, respectively. In these comparisons, the proposed model outstrips Logistic Regression by a staggering 20%, Random Forest by around 15.3% and the Support Vector Machine by approximately 13.3%.

In essence, this comparative analysis accentuates the formidable capabilities of the proposed Multimodal Gated Attention CNN model. By seamlessly integrating diverse data modalities and leveraging state-of-the-art convolutional mechanisms, it not only surpasses individual state-of-the-art multimodal models but sets a new benchmark in terms of accuracy for cancer survival prediction.

Table 5.8: Assessing the proposed multimodal model against existing multimodal predictive techniques through Precision

Model	Precision
Multimodal Gated Attention CNN – {Clinical, Gene-expression, Copy Number Alteration}	0.841
MDNNMD	0.749
Stacked Random Forest	0.841
Logistic Regression	0.549
Random Forest	0.766
Support Vector Machine	0.708

Table 5.8 positions the Multimodal Gated Attention CNN model in relation to other renowned multimodal prediction models based on precision. The proposed model, by amalgamating Clinical, Gene-expression, and Copy Number Alteration data, achieves a precision of 0.841.

Against the backdrop of MDNNMD, which has a precision of 0.749, the proposed model showcases a superior performance, marking an improvement of approximately 12.3%. When contrasted with the Stacked Random Forest, both models exhibit an identical precision of 0.841, suggesting that the proposed model can match the precision of other top-tier models while potentially offering advantages in other metrics.

However, the real distinction of the Multimodal Gated Attention CNN model becomes evident when compared against Logistic Regression, Random Forest, and Support Vector Machine, which register precisions of 0.549, 0.766, and 0.708, respectively. The proposed model surpasses Logistic Regression by a significant 53.2%, outperforms Random Forest by around 9.8%, and edges ahead of the Support Vector Machine by approximately 18.8%.

This comparative exploration underscores the precision-centric strengths of the proposed Multimodal Gated Attention CNN model. Its ability to make accurate positive predictions, especially in the nuanced domain of cancer survival prediction, emphasizes its potential as a reliable tool for researchers and medical professionals alike.

Table 5.9: Assessing the proposed multimodal model against existing multimodal predictive techniques through Sensitivity

Model	**Sensitivity**
Multimodal Gated Attention CNN – {Clinical, Gene-expression, Copy Number Alteration}	0.798
MDNNMD	0.450
Stacked Random Forest	0.747
Logistic Regression	0.183
Random Forest	0.226
Support Vector Machine	0.365

Table 5.9 offers a comparison of the sensitivity achieved by the Multimodal Gated Attention CNN model against other established multimodal prediction models. The proposed model, integrating Clinical, Gene-expression, and Copy Number Alteration data, attains a commendable sensitivity score of 0.798.

When compare with MDNNMD, which records a sensitivity of 0.450, the proposed model stands out with a staggering 77.3% improvement in performance. Against the Stacked Random Forest, the proposed model also exhibits superior capabilities, surpassing it by approximately 6.8%.

The differentiation becomes even more pronounced when comparing against Logistic Regression, Random Forest, and Support Vector Machine. These models register sensitivity scores of 0.183, 0.226, and 0.365 respectively. The Multimodal Gated Attention CNN model outstrips Logistic Regression by an impressive 335.5%, overshadows Random Forest by about 253.1%, and excels over the Support Vector Machine by approximately 118.6%.

This evaluation has the ability of the proposed Multimodal Gated Attention CNN model to accurately identify true positive cases. Given the importance of sensitivity in cancer survival prediction, the model's high sensitivity score solidifies its potential as an indispensable tool in the domain of medical research and patient care.

Table 5.10: Assessing the proposed multimodal model against existing multimodal predictive techniques through Area under Curve

Model	**Area under Curve**
Multimodal Gated Attention CNN – {Clinical, Gene-expression, Copy Number Alteration}	0.946
MDNNMD	0.845
Stacked Random Forest	0.930
Logistic Regression	0.663
Random Forest	0.801
Support Vector Machine	0.810

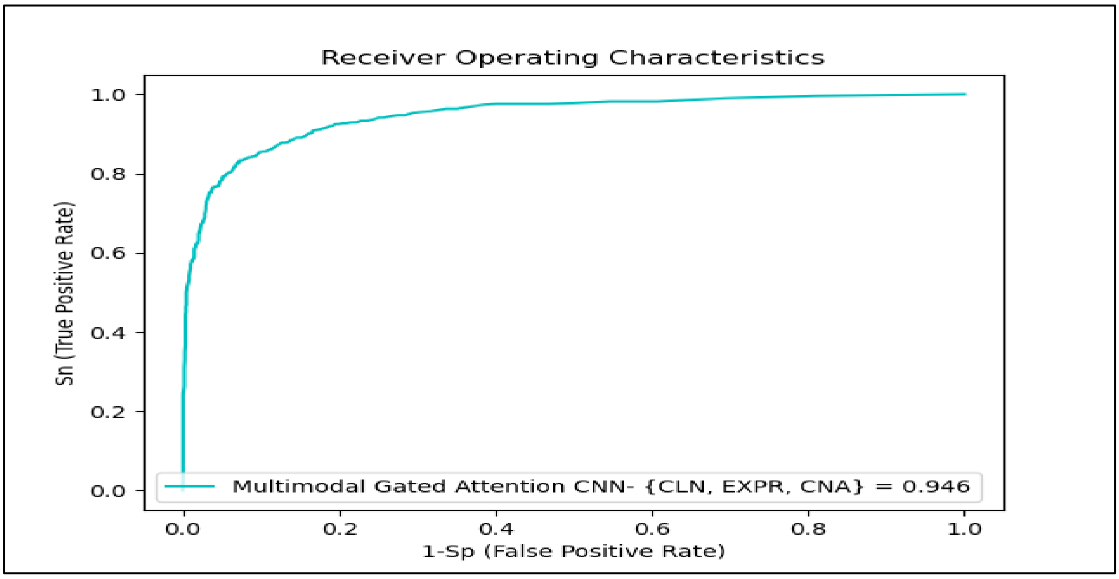


Fig 5.6 ROC curve of Multimodal Gated Attention CNN model on combination of all the three modality

Table 5.10 and Figure 26 describe the AUC scores of the proposed Multimodal Gated Attention CNN model vis-à-vis other leading multimodal prediction models. The amalgamation of Clinical, Gene-expression, and Copy Number Alteration data in our model culminates in an AUC score of 0.950, nearing the pinnacle of classification excellence.

When this score is contrasted against the MDNNMD's AUC of 0.845, the proposed model manifests a noteworthy enhancement of approximately 12.4%. While the Stacked Random Forest achieves a commendable AUC of 0.930, our model still surpasses it, albeit by a narrower margin of about 2.2%.

The superiority of the Multimodal Gated Attention CNN model becomes even more evident when compared with Logistic Regression, Random Forest, and Support Vector Machine, which register AUCs of 0.663, 0.801, and 0.810 respectively. Our model trumps Logistic Regression by a striking 43.3%, outperforms Random Forest by 18.6%, and edges out the Support Vector Machine by approximately 17.3%.

In summation, the AUC scores emphatically underscore the proposed model's superior classification ability. With its near-perfect AUC, the Multimodal Gated Attention CNN model demonstrates its robustness and reliability in the realm of cancer survival prediction, underscoring its potential as a benchmark in the field.

5.4 Summary

This chapter embarked on a journey through the experimental landscape of our cancer survival prediction research. It commenced with a thorough overview of the METABRIC dataset, elucidating its relevance and depth. The technological underpinnings, encompassing the tools, techniques, and software libraries instrumental to our analytical process, were meticulously detailed, providing a comprehensive backdrop to our experimentation phase.

A pivotal segment of this chapter was dedicated to evaluation metrics, where we illuminated the benchmarks against which our model's performance was assessed. This set the stage for a rigorous comparative analysis, where the mettle of the proposed Multimodal Gated Attention CNN model was tested against both its unimodal counterparts and other state-of-the-art multimodal prediction models. Through a series of tables and in-depth discussions, the model showcased its superior performance, particularly in its ability to harness the synergistic potential of multimodal data.

The results presented in this chapter are a testament to the robustness of the proposed model. Its consistent outperformance across various metrics underscores its potential as a reliable tool in the realm of cancer survival prediction. As we move forward, these findings provide both validation for our research endeavors and motivation for future explorations, with the overarching aim of further refining and enhancing the predictive capabilities of cancer survival models.

CHAPTER 6: CONCLUSION & FUTURE DIRECTIONS

This chapter concludes this dissertation by summarizing our key discoveries in applying deep learning to predict cancer patient survival and proposes areas for future exploration. It highlights our notable contributions to the field and suggests pathways for continued research aimed at improving both the precision of predictions and patient outcomes in cancer treatment.

6.1 Conclusion

In this dissertation, we commenced a transformative journey in the domain of cancer survival prediction, utilizing the potential of advanced data analytics and deep learning techniques. Our comprehensive goal was to contribute to the expanding body of knowledge that seeks to enhance our understanding of cancer prognosis and ultimately advance patient outcomes.

At the beginning of our research, we established clear objectives to guide our efforts. We aimed to bridge the gap between classical machine learning methods and modern deep learning approaches in cancer survival prediction. This aspiration emerged from the recognition of the limitations inherent in conventional methods, which often face challenges with high-dimensional or nonlinear data. However, the limitations of these traditional methods became increasingly evident, especially in the era of big data and personalized medicine. High-dimensional datasets, comprising diverse modalities such as clinical records, genomic data, and medical images, presented formidable challenges. These datasets were not only vast but also complex, with intricate interactions between variables that challenge linear assumptions.

Against this backdrop, we turned to deep learning, a subset of artificial intelligence, to revolutionize cancer survival prediction. Deep learning models, inspired by the neural structures of the human brain, held the promise of automatically recognizing complex patterns and structures within extensive and intricate datasets. The attraction of deep learning lies in its ability to process substantial quantities of data, recognize non-linear relationships, and achieve this without the requirement for manual feature engineering. Central to our research was the development of an innovative multimodal prediction model: the Multimodal Gated Attention Convolutional Neural Network. This model represented a paradigm shift in cancer survival prediction, as it not only leveraged the capabilities of deep learning but also addressed the challenge of integrating multimodal data.

The proposed model was meticulously designed to navigate the complexities of multimodal datasets, which are increasingly prevalent in contemporary medical research. Instead of treating all data modalities as uniform, our model adopted a segmented approach, creating separate branches for clinical data, gene expression data, and copy number alteration data. This approach recognized the unique characteristics and challenges posed by each data type, allowing for more focused and tailored analyses. A notable approach within the proposed model was the inclusion of a "gated attention" mechanism. Inspired by human cognitive processes, this mechanism enabled the model to dynamically focus on different parts of the input data, thereby enhancing its ability to detect intricate patterns and relationships. This architectural choice played a crucial role in deciphering the nuances within multimodal data, where different modalities often contained complementary information. Our research emphasized the importance of integrating multimodal data in cancer survival prediction. Through the training and validation of the proposed model, we demonstrated its exceptional performance in harnessing the collaborative potential of diverse data sources. The model consistently outperformed both unimodal counterparts and other state-of-the-art prediction models, highlighting its adaptability and precision in handling the complexities of multimodal data.

A crucial aspect of our analysis was the exploratory data analysis (EDA) phase, where we delved deep into the intricacies of clinical and genomic datasets. This phase revealed patterns, anomalies, and potential correlations that were pivotal in understanding cancer prognosis. The insights acquired from EDA informed the design of the Gated Attention CNN model, ensuring that it was not merely a computational tool but a scientific instrument for unraveling the molecular intricacies of cancer survival. The utilization of the proposed model yielded results that signify a significant advancement in the field of cancer survival prediction. Our model's performance, evaluated through various evaluation metrics, consistently exceeded that of traditional machine learning methods and other state of the art deep learning models.

One notable achievement was the model's ability to seamlessly integrate clinical data, gene expression data, and copy number alteration data. This multimodal integration allowed for a holistic analysis that considered the diverse facets of cancer biology.

Notably, the model excelled in capturing intricate relationships between clinical variables, molecular profiles, and survival outcomes. Furthermore, the gated attention mechanism played a pivotal role in identifying nuanced patterns within the data. The dynamic focus on relevant features enabled the model to uncover previously undetected prognostic factors and molecular markers. These findings have the potential to revolutionize the identification of biomarkers for cancer prognosis and inform targeted therapies.

The consequences of our research extend well beyond the domain of computational science. Improved prediction accuracy has the potential to transform cancer care in several ways. Firstly, it can optimize the allocation of healthcare resources, ensuring that interventions are directed toward patients who need them the most. Secondly, accurate survival predictions empower clinicians to make informed decisions, facilitating personalized treatment plans and early identification of high-risk patients. These capabilities have the potential to significantly improve patient outcomes and the overall efficiency of healthcare systems.

In conclusion, this dissertation signifies the dawn of a new era in cancer survival prediction. We have traversed the historical landscape of survival analysis, embraced the transformative potential of deep learning, and uncovered the capability of multimodal data integration. The proposed model represents a paradigm shift in our approach to understanding and predicting cancer prognosis. Our research endeavors have not only expanded the boundaries of computational science but have also paved the way for more informed, personalized, and effective cancer care. As we progress, the ongoing journey beckons, with the assurance of further enhancements, collaborations, and discoveries that will continue to shape the future of cancer survival prediction and, ultimately, the lives of cancer patients worldwide.

6.2 Future Work

In our ongoing pursuit of advancing cancer survival prediction, several avenues for future research emerge. Firstly, we plan to explore alternative deep learning architectures beyond our successful Multimodal Gated Attention Convolutional Neural Network model, with a focus on recurrent neural networks (RNNs), transformers, and hybrid models to potentially enhance prediction accuracy. Additionally, the integration of transfer learning techniques holds promise in expediting model training and knowledge transfer from pre-trained models, further improving our predictive capabilities. As we continue to integrate diverse data modalities, addressing overfitting challenges becomes paramount, and we will investigate advanced regularization methods and dropout strategies. Collaboration with patient advocacy groups is essential to ensure our model aligns with patient needs, refining prediction criteria to be both technically robust and patient-centric.

Diversifying our dataset remains a priority, and we aim to establish collaborations with additional healthcare institutions to gather a more comprehensive and diverse patient data repository. This expansion encompasses not only demographic diversity but also the inclusion of various medical data types, such as MRI and PET scans, to enrich our model's input features and provide a more holistic assessment of patient health. Moreover, extending the model's applicability to different forms of cancer is a promising avenue for broadening its relevance in the field of oncology.

To enhance model interpretability and explainability, we will develop visualization techniques to elucidate the decision-making processes of the model, aiding clinicians in understanding predictions. Feature importance analysis will be implemented to identify critical predictive factors, empowering clinicians with actionable insights. Addressing biases in predictions is a priority, and methods to detect and mitigate biases will be explored to ensure equitable treatment recommendations across diverse patient demographics.

In conclusion, the future of cancer survival prediction holds immense promise. Collaboration, continuous research, and interdisciplinary partnerships will be instrumental in realizing the full potential of our model. Our commitment to pushing the boundaries of this vital field underscores our dedication to improving the lives of cancer patients worldwide.

REFERENCES

1. Clark, T. G., Bradburn, M. J., Love, S. B., & Altman, D. G. (2003). Survival analysis part I: basic concepts and first analyses. British journal of cancer, 89(2), 232–238. https://doi.org/10.1038/sj.bjc.6601118
2. Alan J Gross, Parametric methods in the analysis of survival data, Microelectronics Reliability, Volume 20, Issue 4, 1980, Pages 477-481, ISSN 0026-2714, https://doi.org/10.1016/0026-2714(80)90592-2.
3. Chakraborty, A. and Tsokos, C. (2021) Parametric and Non-Parametric Survival Analysis of Patients with Acute Myeloid Leukemia (AML). Open Journal of Applied Sciences, 11, 126-148. doi: 10.4236/ojapps.2021.111009
4. Andrew Wey, John Connett, Kyle Rudser, (2015), Combining parametric, semi-parametric, and non-parametric survival models with stacked survival models, Biostatistics, Volume 16, Issue 3, Pages 537–549, https://doi.org/10.1093/biostatistics/kxv001
5. Zhu W, Xie L, Han J, Guo X. (2020). The Application of Deep Learning in Cancer Prognosis Prediction. Cancers 12, no. 3: 603. https://doi.org/10.3390/cancers12030603
6. Huang SC, Pareek A, Seyyedi S, Banerjee I, Lungren MP. (2020). Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. NPJ Digit Med. 3:136. doi:10.1038/s41746-020-00341-z
7. Si Y, Du J, Li Z, Jiang X, Miller T, Wang F, Jim Zheng W, Roberts K. (2021). Deep representation learning of patient data from Electronic Health Records (EHR): A systematic review. J Biomed Inform. 115:103671. doi: 10.1016/j.jbi.2020.103671
8. Khoa A. Tran, Olga Kondrashova, Andrew Bradley, Elizabeth D. Williams, John V. Pearson & Nicola Waddell. (2021). Deep learning in cancer diagnosis, prognosis and treatment selection. Genome Med 13, 152. https://doi.org/10.1186/s13073-021-00968-x
9. Wu, X., Shi, Y., Wang, M., & Li, A. (2023). CAMR: cross-aligned multimodal representation learning for cancer survival prediction. Bioinformatics (Oxford, England), 39(1), btad025. https://doi.org/10.1093/bioinformatics/btad025

10. Vale-Silva, L.A., Rohr, K. (2021) Long-term cancer survival prediction using multimodal deep learning. Sci Rep 11, 13505. https://doi.org/10.1038/s41598-021-92799-4

11. Fatimah Abdulazim Altuhaifa, Khin Than Win, Guoxin Su, Predicting lung cancer survival based on clinical data using machine learning: A review, Computers in Biology and Medicine, Volume 165, 2023, 107338, ISSN 0010-4825, https://doi.org/10.1016/j.compbiomed.2023.107338.

12. Bashiri, A., Ghazisaeedi, M., Safdari, R., Shahmoradi, L., & Ehtesham, H. (2017). Improving the Prediction of Survival in Cancer Patients by Using Machine Learning Techniques: Experience of Gene Expression Data: A Narrative Review. Iranian journal of public health, 46(2), 165–172.

13. Graf RP, Eskander R, Brueggeman L, Stupack DG. Association of Copy Number Variation Signature and Survival in Patients With Serous Ovarian Cancer. JAMA Netw Open. 2021;4(6):e2114162. doi:10.1001/jamanetworkopen.2021.14162

14. Lee, S., & Lim, H. (2019). Review of statistical methods for survival analysis using genomic data. Genomics & informatics, 17(4), e41. https://doi.org/10.5808/GI.2019.17.4.e41

15. Berg J, Robbins G. The failure of a model to predict cancer survival. J Chron Dis. 1967;20:809–814. doi: 10.1016/0021-9681(67)90093-8.

16. Kim YJ, Yoon SJ, Suh S-Y, Hiratsuka Y, Kang B, Lee SW, et al. (2022) Performance of clinician prediction of survival in oncology outpatients with advanced cancer. PLoS ONE 17(4): e0267467. https://doi.org/10.1371/journal.pone.0267467

17. Hosni, M., Abnane, I., Idri, A., Carrillo de Gea, J. M., & Fernández Alemán, J. L. (2019). Reviewing ensemble classification methods in breast cancer. Computer methods and programs in biomedicine, 177, 89–112. https://doi.org/10.1016/j.cmpb.2019.05.019

18. Peng, Joanne & Lee, Kuk & Ingersoll, Gary. (2002). An Introduction to Logistic Regression Analysis and Reporting. The Journal of Educational Research, volume 96, pages 3-14. https://doi.org/10.1080/00220670209598786

19. Shichao Zhang, Debo Cheng, Zhenyun Deng, Ming Zong, Xuelian Deng. (2018). A novel kNN algorithm with data-driven k parameter computation, Pattern

Recognition Letters, Volume 109, Pages 44-54, ISSN 0167-8655, https://doi.org/10.1016/j.patrec.2017.09.036

20. Lynch, C. M., Abdollahi, B., Fuqua, J. D., de Carlo, A. R., Bartholomai, J. A., Balgemann, R. N., van Berkel, V. H., & Frieboes, H. B. (2017). Prediction of lung cancer patient survival via supervised machine learning classification techniques. International journal of medical informatics, 108, 1–8. https://doi.org/10.1016/j.ijmedinf.2017.09.013
21. Cruz JA, Wishart DS. (2007). Applications of machine learning in cancer prediction and prognosis. Cancer Inform. 2, 59-77. PMID: 19458758; PMCID: PMC2675494
22. Lynch, C. M., Abdollahi, B., Fuqua, J. D., de Carlo, A. R., Bartholomai, J. A., Balgemann, R. N., van Berkel, V. H., & Frieboes, H. B. (2017). Prediction of lung cancer patient survival via supervised machine learning classification techniques. International journal of medical informatics, 108, 1–8. https://doi.org/10.1016/j.ijmedinf.2017.09.013
23. Hosni, M., Abnane, I., Idri, A., Carrillo de Gea, J. M., & Fernández Alemán, J. L. (2019). Reviewing ensemble classification methods in breast cancer. Computer methods and programs in biomedicine, 177, 89–112. https://doi.org/10.1016/j.cmpb.2019.05.019
24. Breiman, L. (1996). Bagging predictors. Machine Learning 24, 123–140. https://doi.org/10.1007/BF00058655
25. N. Fatima, L. Liu, S. Hong and H. Ahmed. (2020). Prediction of Breast Cancer, Comparative Review of Machine Learning Techniques, and Their Analysis, in IEEE Access, vol. 8, pp. 150360-150376, doi: 10.1109/ACCESS.2020.3016715
26. Hosni, M., Abnane, I., Idri, A., Carrillo de Gea, J. M., & Fernández Alemán, J. L. (2019). Reviewing ensemble classification methods in breast cancer. Computer methods and programs in biomedicine, 177, 89–112. https://doi.org/10.1016/j.cmpb.2019.05.019
27. Lynch, C. M., Abdollahi, B., Fuqua, J. D., de Carlo, A. R., Bartholomai, J. A., Balgemann, R. N., van Berkel, V. H., & Frieboes, H. B. (2017). Prediction of lung cancer patient survival via supervised machine learning classification techniques. International journal of medical informatics, 108, 1–8.

https://doi.org/10.1016/j.ijmedinf.2017.09.013

28. Bozkurt, Caner & Aşuroğlu, Tunç. (2023). Mortality Prediction of Various Cancer Patients via Relevant Feature Analysis and Machine Learning. SN Computer Science. 4. 10.1007/s42979-023-01720-5

29. Arya, Nikhilanand & Saha, Sriparna & Mathur, Archana & Saha, Snehanshu. (2023). Improving the robustness and stability of a machine learning model for breast cancer prognosis through the use of multi-modal classifiers. Scientific Reports. 13. 10.1038/s41598-023-30143-8

30. Zolfaghari, Behrouz & Mirsadeghi, Leila & Bibak, Khodakhast & Kavousi, Kaveh. (2023). Cancer Prognosis and Diagnosis Methods Based on Ensemble Learning. ACM Computing Surveys. 55. 10.1145/3580218

31. Sorayaie Azar, Amir & Babaei Rikan, Samin, (2022). Application of machine learning techniques for predicting survival in ovarian cancer. BMC Medical Informatics and Decision Making. 22. 10.1186/s12911-022-02087-y

32. Yan, F., Feng, Y. (2022). A two-stage stacked-based heterogeneous ensemble learning for cancer survival prediction. *Complex Intell. Syst.* **8**, 4619–4639. https://doi.org/10.1007/s40747-022-00791-w

33. S P, S., I, S., A H, K., R, H., S, K., H, G., & A, S. Z. (2020). Predicting Lung Cancer Patients' Survival Time via Logistic Regression-based Models in a Quantitative Radiomic Framework. *Journal of biomedical physics & engineering*, *10*(4), 479–492. https://doi.org/10.31661/JBPE.V0I0.1027

34. Bartholomai, J. A., & Frieboes, H. B. (2018). Lung Cancer Survival Prediction via Machine Learning Regression, Classification, and Statistical Techniques. IEEE International Symposium on Signal Processing and Information Technology, 2018, 632–637. https://doi.org/10.1109/ISSPIT.2018.8642753

35. Mostavi, M., Chiu, YC., Huang, Y. et al. (2020). Convolutional neural network models for cancer type prediction based on gene expression. BMC Med Genomics 13, article 44. https://doi.org/10.1186/s12920-020-0677-2

36. Kumar, Y., Gupta, S., Singla, R. et al. (2022). A Systematic Review of Artificial Intelligence Techniques in Cancer Prediction and Diagnosis. Arch Computat Methods Eng 29, 2043–2070. https://doi.org/10.1007/s11831-021-09648-w

37. Lee M. (2023). Deep Learning Techniques with Genomic Data in Cancer

Prognosis: A Comprehensive Review of the 2021–2023 Literature. Biology; 12(7):893. https://doi.org/10.3390/biology12070893

38. Shen Junjie, Li Huijun, Yu Xinghao, Bai Lu, Dong Yongfei, Cao Jianping, Lu Ke, Tang Zaixiang. (2023). Efficient feature extraction from highly sparse binary genotype data for cancer prognosis prediction using an auto-encoder. Frontiers in Oncology. Volume = 12. DOI=10.3389/fonc.2022.1091767

39. Wu, X., & Fang, Q. (2022). Stacked Autoencoder Based Multi-Omics Data Integration for Cancer Survival Prediction. ArXiv, abs/2207.04878.

40. Xu, J., Xiang, L., Liu, Q., Gilmore, H., Wu, J., Tang, J., & Madabhushi, A. (2016). Stacked Sparse Autoencoder (SSAE) for Nuclei Detection on Breast Cancer Histopathology Images. IEEE transactions on medical imaging, 35(1), 119–130. https://doi.org/10.1109/TMI.2015.2458702

41. Li, M. M., Huang, K., & Zitnik, M. (2022). Graph representation learning in biomedicine and healthcare. Nature biomedical engineering, 6(12), 1353–1369. https://doi.org/10.1038/s41551-022-00942-x

42. Ahmedt-Aristizabal, D., Armin, M.A., Denman, S., Fookes, C., & Petersson, L. (2021). A Survey on Graph-Based Deep Learning for Computational Histopathology. Computerized medical imaging and graphics: the official journal of the Computerized Medical Imaging Society, 95, 102027

43. Ahmedt-Aristizabal, David & Armin, Ali & Denman, Simon & Fookes, Clinton & Petersson, Lars. (2021). A Survey on Graph-Based Deep Learning for Computational Histopathology. Computerized Medical Imaging and Graphics. 95. 102027. 10.1016/j.compmedimag.2021.102027

44. Wu, X., Shi, Y., Wang, M., & Li, A. (2023). CAMR: cross-aligned multimodal representation learning for cancer survival prediction. Bioinformatics (Oxford, England), 39(1), btad025. https://doi.org/10.1093/bioinformatics/btad025

45. J. Gao, et al., (2022) Predicting the Survival of Cancer Patients With Multimodal Graph Neural Network in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 19, no.02, pp.699-709. doi: 10.1109/TCBB.2021.3083566

46. Su-Fang Zhang, Jun-Hai Zhai, Bo-Jun Xie, Yan Zhan and Xin Wang. (2019). Multimodal Representation Learning: Advances, Trends and Challenges. 2019

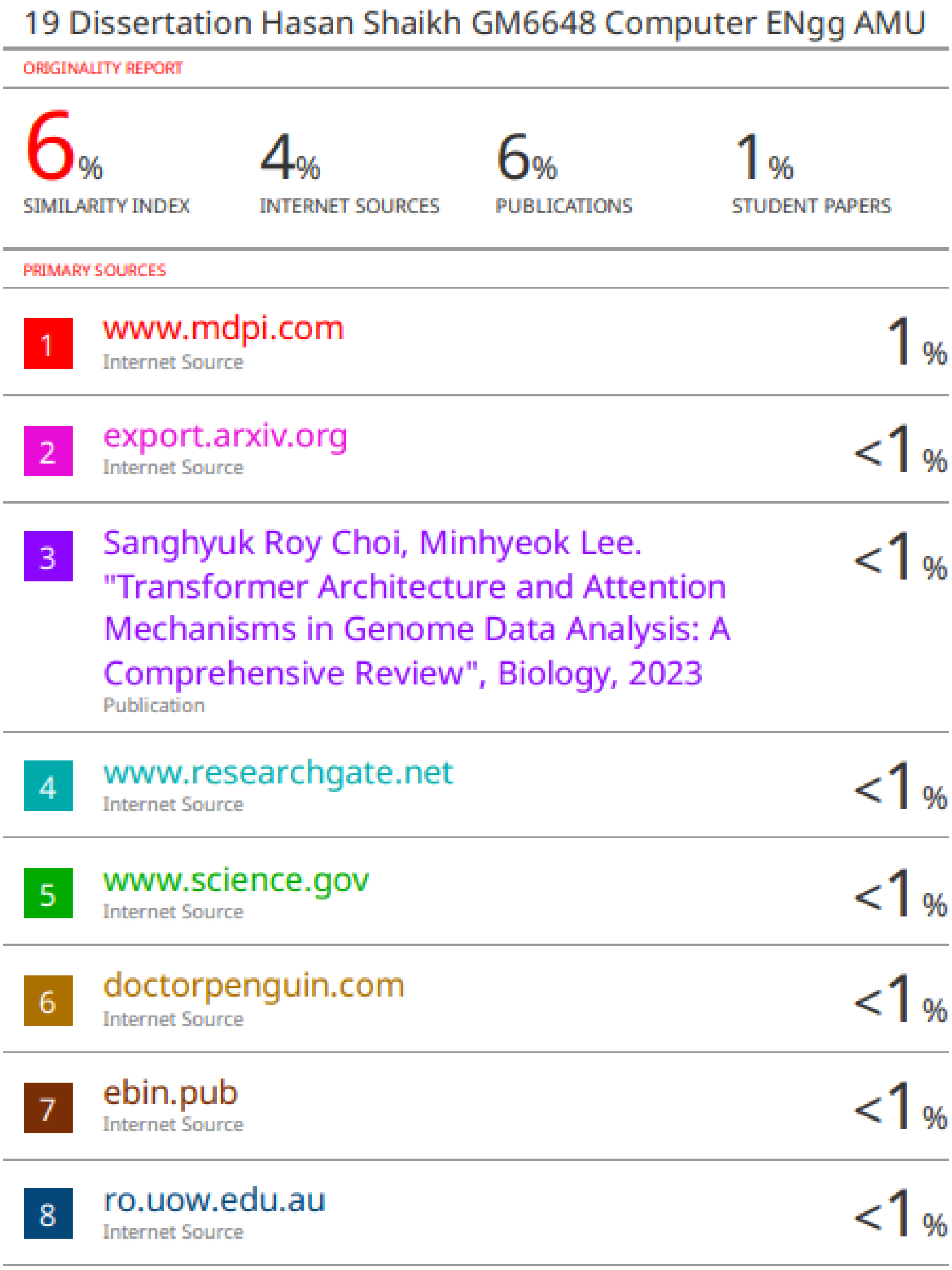
International Conference on Machine Learning and Cybernetics (ICMLC), Kobe, Japan, pp. 1-6, doi: 10.1109/ICMLC48188.2019.8949228

47. Chen, H., Gao, M., Zhang, Y., Liang, W., & Zou, X. (2019). Attention-Based Multi-NMF Deep Neural Network with Multimodality Data for Breast Cancer Prognosis Model. BioMed research international. https://doi.org/10.1155/2019/9523719
48. Wu, X., Shi, Y., Wang, M., & Li, A. (2023). CAMR: cross-aligned multimodal representation learning for cancer survival prediction. *Bioinformatics (Oxford, England)*, *39*(1), btad025. https://doi.org/10.1093/bioinformatics/btad025
49. Fan, Z., Jiang, Z., Liang, H., & Han, C. (2023). Pancancer survival prediction using a deep learning architecture with multimodal representation and integration. *Bioinformatics advances*, *3*(1), vbad006. https://doi.org/10.1093/bioadv/vbad006
50. Arya, N., & Saha, S. (2022). Multi-Modal Classification for Human Breast Cancer Prognosis Prediction: Proposal of Deep-Learning Based Stacked Ensemble Model. *IEEE/ACM transactions on computational biology and bioinformatics*, *19*(2), 1032–1041. https://doi.org/10.1109/TCBB.2020.3018467
51. Summrina Kanwal, Faiza Khan, Sultan Alamri. (2022). A multimodal deep learning infused with artificial algae algorithm – An architecture of advanced E-health system for cancer prognosis prediction, Journal of King Saud University - Computer and Information Sciences, Volume 34, Issue 6, Part A, Pages 2707-2719, ISSN 1319-1578, https://doi.org/10.1016/j.jksuci.2022.03.011.
52. Li, X., Jonnagaddala, J., Cen, M., Zhang, H., & Xu, S. (2022). Colorectal Cancer Survival Prediction Using Deep Distribution Based Multiple-Instance Learning. *Entropy (Basel, Switzerland)*, *24*(11), 1669. https://doi.org/10.3390/e24111669
53. Wu, Xing & Fang, Qiulian. (2022). Stacked Autoencoder Based Multi-Omics Data Integration for Cancer Survival Prediction. 10.48550/arXiv.2207.04878
54. Cheerla, A., & Gevaert, O. (2019). Deep learning with multimodal representation for pancancer prognosis prediction. *Bioinformatics (Oxford, England)*, *35*(14), i446–i454. https://doi.org/10.1093/bioinformatics/btz342
55. Huang, Zhi & Zhan, Xiaohui & Xiang & Huang, Kun. (2019). SALMON:

Survival Analysis Learning With Multi-Omics Neural Networks on Breast Cancer. Frontiers in Genetics. 10. 10.3389/fgene.2019.00166

56. Sun, D., Wang, M., & Li, A. (2018). A multimodal deep neural network for human breast cancer prognosis prediction by integrating multi-dimensional data. IEEE/ACM transactions on computational biology and bioinformatics, https://doi.org/10.1109/TCBB.2018.2806438

APPENDIX



19 Dissertation Hasan Shaikh GM6648 Computer ENgg AMU

ORIGINALITY REPORT

6%	4%	6%	1%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.mdpi.com Internet Source	1%
2	export.arxiv.org Internet Source	<1%
3	Sanghyuk Roy Choi, Minhyeok Lee. "Transformer Architecture and Attention Mechanisms in Genome Data Analysis: A Comprehensive Review", Biology, 2023 Publication	<1%
4	www.researchgate.net Internet Source	<1%
5	www.science.gov Internet Source	<1%
6	doctorpenguin.com Internet Source	<1%
7	ebin.pub Internet Source	<1%
8	ro.uow.edu.au Internet Source	<1%