GraphSAGE-based Approach for Age-specific Multi-omics Biomarker Identification in Bladder Cancer

by

Name: Usman Fakhar (Student ID: 1221662)

A THESIS
SUBMITTED TO THE DEPARTMENT OF COMPUTER SCIENCE
AND THE FACULTY OF GRADUATE STUDIES
OF LAKEHEAD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
Master OF Computer Science

Lakehead University
Thunder Bay, ON, Canada

Supervisory Committee

Presented by

Usman Fakhar

(*Student, Lakehead University, Thunder Bay, Ontario, Canada.*
ID: 1221662)

Internal Reader

M. Mazhar Rathore

(*Assistant Professor, Department of Computer Science, Lakehead University, Thunder Bay, Ontario, Canada.*)

External Reader

Dr. Yong Deng

(*Assistant Professor, Software Engineering, Lakehead University, Thunder Bay, Ontario, Canada.*
Email: yong.deng@lakeheadu.ca)

Supervisor

Dr. Abedalrhman Alkhateeb

(*Assistant Professor, Department of Computer Science, Lakehead University, Thunder Bay, Ontario, Canada.*
Email: aalkhate@lakeheadu.ca)

ABSTRACT

Bladder cancer is a highly prevalent malignancy with substantial morbidity and mortality, emphasizing the urgent need for early detection and personalized treatment strategies. Although recent advances in cancer genomics have enhanced our understanding of tumor biology, the role of age-related genomic variations in bladder cancer progression remains largely unexplored. In this study, we present a novel framework that combines multi-omics data integration with Graph Neural Networks (GNNs) to identify age-specific biomarkers associated with bladder cancer prognosis. We integrate copy number alterations (CNA), DNA methylation, and mRNA expression profiles into graph-based representations, where nodes denote genomic features and edges encode molecular interactions. Unlike conventional statistical or machine learning approaches, our method incorporates age both as a stratification factor and as a graph-level feature, enabling the model to learn distinct molecular signatures across different patient age groups. Using survival outcomes, we determined 64 years as the optimal threshold for age stratification, revealing significant differences in mortality between patients aged $\leq$64 years (30.46%) and those $>$ 64 years (51.74%), thereby highlighting the prognostic value of age in bladder cancer. To enhance model interpretability and performance, we implemented a robust feature selection pipeline involving variance thresholding, ANOVA F-scores, L1 regularization, and Recursive Feature Elimination with Cross-Validation (RFECV). Among several models tested, GraphSAGE consistently achieved the highest accuracy, F1-score, and AUC, demonstrating the effectiveness of graph-based learning in capturing complex biological relationships. Furthermore, SHAP (SHapley Additive exPlanations) analysis revealed key age-associated biomarkers such as *SNRPN*, *LINC01091*, and *DHX36*, which are strongly implicated in patient survival and may inform future therapeutic targeting. This study introduces a comprehensive, age-aware graph learning framework for biomarker discovery in bladder cancer, offering a powerful tool for advancing personalized diagnosis, prognosis, and treatment planning. Beyond bladder cancer, this methodology has the potential to be generalized to other cancer types where age significantly influences disease trajectory, thereby contributing to the broader field of precision oncology. By bridging age-specific genomic variation with multi-modal data and explainable machine learning, our approach opens new avenues for developing clinically actionable insights and enhancing patient-specific management strategies in oncology.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest gratitude to my father and mother, whose unwavering support, sacrifices, and prayers have been the foundation of my academic journey. Their love, encouragement, and belief in my abilities have been a constant source of motivation throughout my life.

I would also like to extend my sincere appreciation to my supervisor, Dr. Abedalrhman Alkhateeb, for his invaluable guidance, continuous support, and insightful feedback during the course of my research. His mentorship, patience, and encouragement have been instrumental in shaping this thesis and expanding my knowledge in the field. I am truly grateful for the opportunity to learn from his expertise and experience.

This work would not have been possible without the inspiration, support, and mentorship I have received from each of them. I am forever indebted to their kindness and belief in me.

List of Acronyms

GNN	Graph Neural Network
GCN	Graph Convolutional Network
GAT	Graph Attention Network
SHAP	SHapley Additive exPlanations
CNA	Copy Number Alteration
ML	Machine Learning
PCA	Principal Component Analysis
RFECV	Recursive Feature Elimination with Cross-Validation
SVM	Support Vector Machine
ANN	Artificial Neural Network
EMR	Electronic Medical Record
AUC	Area Under the Curve
ROC	Receiver Operating Characteristic
ctDNA	Circulating Tumor DNA
aCGH	Array Comparative Genomic Hybridization
NHIB	Network Modeling Analysis in Health Informatics and Bioinformatics
BIBM	IEEE International Conference on Bioinformatics and Biomedicine

PUBLICATIONS

Parts of this thesis have been submitted for peer-review, published, or accepted for publication:

Published

- **Machine Learning Model to Predict Autism Spectrum Disorder Using Eye Gaze Tracking** is published in the 2023 IEEE International Conference on Bioinformatics and Biomedicine [1].
- **Machine Learning Model for Anxiety Disorder Diagnosis Based on Sensory Time-Series Data** is published in Bioinformatics and Biomedical Engineering [2].

Submitted for Publication

- **GraphSAGE-based Approach for Age-specific Multi-omics Biomarker Identification in Bladder Cancer** has been submitted for publication to Network Modeling Analysis in Health Informatics and Bioinformatics (NHIB) on February 26, 2025.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Cancer continues to be one of the leading causes of death globally, characterized by its heterogeneity and complex changes that enable it to develop and progress. Bladder cancer is also a complex cancer that is equally characterized by its recurrence and variable prognosis, highlighting the need for enhanced diagnostic and prognostication methods. In this chapter, the underlying rationale for the study—utilizing multi-omics data alongside machine learning methods to support the early detection and stratification of bladder cancer cases—is articulated. It starts by outlining the clinician's view of cancer and the biological processes that occur in tumor progression. It then focuses on how the incorporation of multi-omics, including genomic, transcriptomic, proteomic, and epigenomic features, transforms cancer biology by providing a complete overview of the topic. Accordingly, special emphasis is placed on the ability of machine learning in identifying meaningful patterns in this high-dimensional data, thus enabling new biomarker discovery and design of individualized modalities of treatment. Finally, the chapter concludes by summarizing the overall problem statements, goals, and contributions of the study, which involve the evaluation of feature selection methods, graph-based modeling, and interpretable machine learning, all in the pursuit of advancing computational oncology.

1.1 Cancer: An Overview

Cancer is as a heterogeneous group of diseases with uncontrolled cell mutation growth and the ability to invade or metastasize to other parts of the body [3]. This disease is further characterized by the very unsettling ability of such cancer cells to not only invade nearby tissues but to metastasize and migrate to other areas in other parts of the body. If we look into the molecular complexities underlying cancer progression, we find that this disease is a result of a vast number of mutations in genes with a very crucial role in regulating normal cell processes [4]. Such fundamental processes include not only regulation of cell cycles but include processes responsible for cell death, scientifically referred to as apoptosis, and processes responsible for repairing damaged DNA [5]. This intricate interplay between genetic mutations and epigenetic alterations leads to a state of uncontrolled and deregulated cell growth, eventually culminating in malignant tumor formation. Approximately 10 million individuals die annually from cancer, making cancer one of the most leading causes of mortality in the world with its effects being felt in a variety of communities and populations [6].

A fundamental aspect that distinctly defines the complicated nature of cancer is its inherent heterogeneity. This essentially refers to the fact that in the extensive and multifarious world of cancer, there are many tumors that although all belong to a specific type of cancer, show significant variations in their molecular changes and their respective genomic variations. This inbuilt diversity within the tumors itself then translates into a vast range of clinical presentations, and hence, significant variations in how different patients react to different treatment modalities. Such variations then results in the complexity in controlling cancer effectively [7]. With our progressive knowledge and ability to define cancers according to their distinct molecular characteristics, as opposed to conventional histopathological features, has brought in revolutionary advancements in the discipline of precision oncology. With this paradigm-shifting approach, treatments are carefully tailored to match the distinct genetic profile of each individual patient, thus greatly improving the overall efficacy of the therapeutic interventions that are given to them [8].

Recent advancements, along with major breakthroughs achieved in high-throughput technologies, such as next-generation sequencing (NGS), along with microarray-based methods and mass spectrometry, have highly empowered researchers and scientists. These cutting-edge technologies have helped them gain a multi-faceted and integrative understanding of intricate features that exist in the fields of genetics, transcriptomics, proteomics, and metabolomics that occur in different types of tumors [9]. This integrative multi-omics approach has highly increased our overall knowledge regarding cancer biology in totality, and has also introduced new research avenues. This research is centered on identifying new biomarkers that can be used for early detection, effective prognostication, and better therapeutic strategies to control cancer more effectively [10].

A significant and serious challenge that working researchers in the field of oncology face is the daunting and overwhelming challenge to interpret the evolution of tumors, and to comprehend the intricate interactions that tumors have with the surrounding microenvironment in which tumors develop and establish themselves. Tumors do not exist in a static state and are merely not changing; instead, tumors are dynamic and undergo a myriad of evolutionary pathways throughout time, continuously picking up new mutations while adapting and remodeling themselves in response to selective pressures, such as those from chemotherapy treatments and immune system attacks. This extremely dynamic nature of cancer, and the heterogeneity within tumors themselves, makes it extremely difficult to predict how patients will react to treatments and to develop effective therapeutic strategies that can effectively battle this multi-faceted disease [11]. With such overwhelming and challenging tasks in front of them, researchers increasingly find themselves turning to multi-omics strategies, which are cutting-edge and innovative methodologies that integrate and synthesize information obtained from multiple molecular levels, all in an attempt to better comprehend and more fully interpret the intricate and complicated processes that eventually lead to cancer progression [12].

1.2 Bladder Cancer

Bladder cancer, is a serious and serious form of cancer that arises from urothelium. Urothelium is a specialized lining responsible for playing a crucial role in protecting and lining the inner bladder surfaces. It is noteworthy to point out that bladder cancer has a distinct feature in that it is the 10th most commonly diagnosed cancer in the world. This fact is rather alarming, considering that it is accompanied by an estimated incidence rate of about 573,000 new cases diagnosed annually. Besides this alarming number, one should point out that approximately 212,000 deaths occur each year as a direct consequence of this particular illness [6]. Bladder cancer is generally divided into two important groups depending on the degree of invasion that individuals who have developed this kind of cancer have to undergo. These two groups are known as non-muscle invasive bladder cancer (NMIBC) and muscle-invasive bladder cancer (MIBC). Out of the two important groups, one should point out that NMIBC is responsible for almost 70% of all reported bladder cancer cases, while MIBC is a more serious form of this illness. Unfortunately, this more invasive form is characterized by a much worse prognosis for individuals who are diagnosed with this specific form of cancer [13].

Bladder cancer staging is necessary to determine the extent of the disease and to decide on management. The TNM system is frequently used, in which T describes the invasion of the tumor into the bladder wall, N informs us about lymph node involvement, and M indicates metastasis to other parts of the body. The different stages are shwon in Table 1.1.

Table 1.1: Bladder Cancer Staging

Stage	**Location**	**Severity**	**Description**
Ta	Inner bladder lining (urothelium)	Non-invasive	Cancer is confined to the bladder lining and has not invaded deeper layers. Usually presents as papillary growths with a lower risk of progression.
T1	Connective tissue beneath the bladder lining	Non-muscle-invasive	Cancer has grown into the connective tissue layer beneath the lining of the bladder but has not invaded the muscle layer. This stage carries a higher risk of progression than earlier stages.

Continued on next page

Table 1.1 – continued from previous page

Stage	Location	Severity	Description
T2	Muscle layer of the bladder wall	Muscle-invasive	Cancer has spread into the muscle layer of the bladder wall. It is subdivided into: T2a, where cancer has invaded the superficial muscle, and T2b, where it has spread to the deeper muscle. T2 bladder cancers are muscle-invasive and typically require more intensive treatment.
T3	Fatty tissue surrounding the bladder	Advanced local disease	Cancer has penetrated through the muscle layer into the adjacent fat. It is subdivided into: T3a, where the invasion into the fat layer is microscopic, and T3b, where the invasion is macroscopic and evident on imaging or during surgery. This stage suggests more advanced disease with a higher likelihood of regional spread.
T4	Nearby organs (prostate, uterus, vagina, pelvic/abdominal wall)	Most advanced local disease	The tumor has spread beyond the bladder into adjacent organs or structures. T4a indicates invasion of contiguous organs like the prostate, uterus, or vagina, while T4b indicates invasion into the pelvic or abdominal wall. This stage typically requires systemic treatments and, depending on the extent of metastatic disease, may involve palliative care.

Understanding these stages is important for clinicians to devise appropriate treatment plans and provide patients with accurate prognostic information. As bladder cancer progresses through these stages, the treatment options become increasingly aggressive, and the overall prognosis generally worsens [14]. Figure 1.1 shows the progression of bladder cancer through various stages:

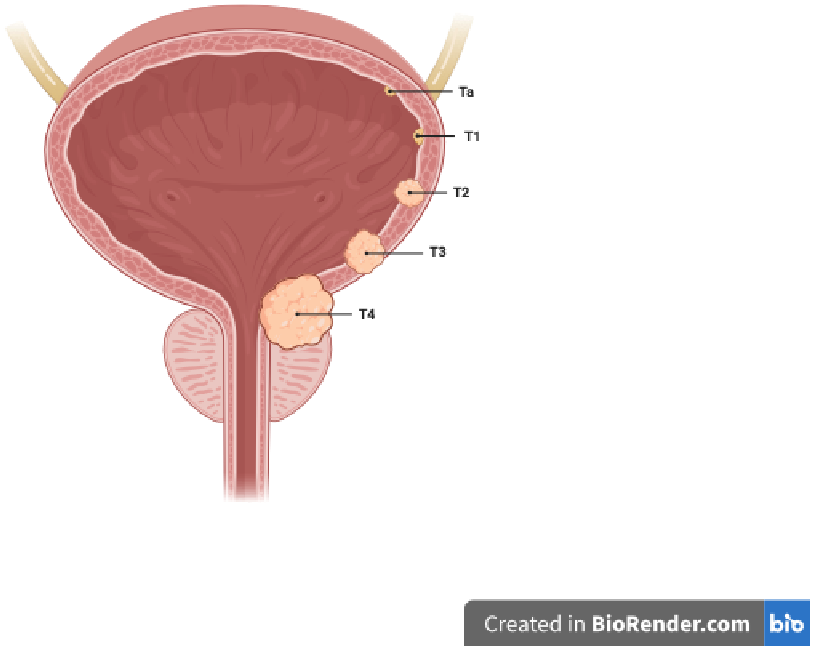

Figure 1.1: Bladder Cancer Stages.

The intricacies in dealing with and managing bladder cancer are further complicated with alarmingly high rates of recurrences in individuals who have been diagnosed with NMIBC. Additionally, there is a basic issue in properly assessing whether certain cases of NMIBC will develop into the more serious variant known as MIBC or whether or not they will be localized and not develop further [15].

The various factors that contribute to bladder cancer include a vast range of factors, with one being tobacco smoking, which is a known and potent cause for this disease. Besides tobacco use, exposure to certain industrial dyes, working in certain occupations within the rubber and textile industries, and exposure to certain noxious industrial chemicals pose a significant risk for those who are in danger. Additionally, individuals who experience prolonged infections within the urinary tract system, or who have a history of bladder stones or chronic inflammation, have a greater chance of developing bladder cancer as well [16]. Among all identified factors for bladder cancer, tobacco use is the most salient modifiable one, accounting for about half of all reported cases for bladder cancer and thus making it a public health concern that warrants notice and action [17]. Another important factor to take into account is age, as studies have established that the probability for bladder cancer increases greatly for individuals who are older than 60 years old [18]. It has been noted, further, that bladder cancer is more apt to occur in men than in women, with a remarkable ratio of 2:1 to 4:1 for different counties for different ethnicity with different diets and lifestyles [19].

The genetic basis for the onset and progression of bladder cancer is not only multifactorial in nature but is highly intricate, with a vast range of genetic alterations being responsible for causing the disease. Among them are serious mutations in important genes, namely TP53, RB1, and FGFR3, along with a range of structural changes within chromosomes and the presence of copy number variations [20]. Nonetheless, apart from the above-stated mutations, extensive research has established that numerous other genes, which are responsible for important processes in life such as regulating the cell cycle, processes responsible for repairing DNA, and apoptosis, have been found to harbor mutations. Such genetic changes are crucial, as they are responsible for initiating the disease and promoting progression in bladder cancer [21]. However, despite mounting knowledge about the complicated molecular processes responsible for bladder cancer formation, several important challenges persist and are still being met. These include challenges with respect to detecting biomarkers consistently that could be crucial for enabling early disease detection, accurately being able to predict responses to different treatments, and ultimately improving outcomes in those who have been diagnosed with this specific type of cancer [22].

Multi-omics approaches, which include a variety of different types of biological data, ranging from genomic, transcriptomic, and proteomic information, have the incredible ability to provide information regarding intricate molecular pathways that have a crucial role to play in bladder cancer pathology. Most of the research studies carried out over the last few decades have repeatedly proved that such integrated strategies have a ability to discover important genetic drivers as well as changes in important signaling pathways that are instrumental in understanding this disease. Such breakthrough findings have potential therapeutic targets that could be utilized for the creation of personalized treatment regimens specifically designed to address the specific demands of individual patients [23].

Recent research integrating genomic and transcriptomic information has revealed new mutations as well as extensive immune microenvironment changes linked with bladder cancer. This find is a thrilling opportunity for the seamless integration of immune checkpoint therapy in overall treatment regimens and thus improving their overall efficacy and effectiveness [24]. However, despite all these advancements in bladder cancer research, one big challenge still remains: efficiently integrating currently available omics data sets. This challenge is extremely crucial, since successful integration is a prerequisite for developing clinically useful biomarkers and treatment strategies that can be effectively translated to useful applications in the management and care of individuals affected with this disease.

1.3 Multi-Omics Data in Cancer

Multi-omics is a developing discipline in the biological sciences that combines findings from a variety of 'omics' such as genomics, transcriptomics, proteomics, metabolomics, and epigenomics. The integrated approach aims to provide a complete outlook on biological systems and their complex interactions [25]. The emergence of high-throughput technologies has revolutionized our capacity to generate large-scale datasets in multiple omics domains, thus facilitating explanations of biological events with a previously unparalled level of complexity and profundity [26].

The utilization of several omics methodologies has several advantages over single-omics-based approaches. Overall, it increases understanding of cell mechanisms by bringing together information at all biological layers ranging from DNA to RNA to proteins and eventually metabolites [27]. Such a comprehensive understanding is particularly important in complex disorders like cancer in which changes at multiple biological layers combine to affect the phenotypic characteristics of the disease. In addition to this, integration of multi-omics information can reveal emergent properties that would be masked if single layers of omics were to be examined separately [28].

One of the key challenges in the field of multi-omics studies is the integration and analysis of different types of data. Data from each omics layer differ in characteristics, scales, and noise profiles. To surmount this challenge, researchers have developed a variety of computational methods, such as network-based approaches, machine learning methods, and statistical models for data integration [29]. The central aim of these methods is the identification of important patterns and relationships between different omics layers, thus providing insights into the underlying biological mechanisms.

The scope of multi-omics is advancing rapidly with the continuous advent of new technologies and analytical platforms. Multi-omics techniques are providing extremely detailed views of cell diversity and intricate biological functions at the single-cell resolution. In addition, integration of multi-omics with clinic information enables to build more customized approaches to cancer diagnosis, prognostication, and therapeutic intervention [30].

The integration of multi-omics information in cancer studies has been extremely beneficial as it provides a more in-depth and integrated understanding of the complex molecular mechanisms involved in tumorigenesis (the development of normal cells into malignant cells) and metastasis (the spread of cancer cells from their point of origin to other parts of the body). All omics aspects provide different views of the molecular characteristics of cancer cells and hence augment the integrated understanding of mechanisms involved in cancer initiation and progression.

The discipline of genomics, particularly the study of copy number variations (CNAs), has been vital to identify and describe cancer. Importantly, CNAs with considerable gains or losses in genomic parts represent distinguishing markers of cancer genomes and can have a tremendous impact on gene and cell function [31]. The detection approaches with high resolution to identify CNAs have included array comparative genomic hybridization (aCGH) and next-generation sequencing [32].

The integration of datasets related to copy number alterations (CNA), mRNA expression, and DNA methylation is a useful multi-omics tool in the field of cancer studies as follows:

- Kim and Lee [33] analyzed mRNA expression profiles in breast cancer, identifying 20 differentially expressed mRNAs with excellent diagnostic performance.
  - 14 downregulated and 6 upregulated mRNAs were identified in breast cancer tissues compared to non-cancerous tissues.

  - This study demonstrated the potential of mRNA expression analysis in cancer detection.

- Bhattacharya et al. [34] developed the TACNA (Transcriptional Adaptation to CNA) profiling method to analyze the transcriptional effects of CNAs across human cancers.
  - Their study revealed that CNAs can promote tumor progression by altering gene expression levels.
  - This highlights the importance of CNAs in cancer development and detection.

- Holm et al. [35] examined DNA methylation patterns in breast cancer subtypes, identifying distinct methylation profiles for basal-like, luminal A, and luminal B tumors.
  - Their research suggested that a large fraction of genes with subtype-specific expression patterns may be regulated through methylation.
  - This emphasizes the role of DNA methylation in cancer subtype classification and detection.

These studies collectively demonstrate how CNAs, mRNA expression, and DNA methylation contribute to cancer biology and serve as potential biomarkers for improved cancer detection and classification. By integrating these three omics layers, researchers can explore the complex interplay between genomic, transcriptomic, and epigenomic alterations in cancer. This comprehensive approach reveals how CNAs influence gene expression patterns, highlights the role of DNA methylation changes in transcriptional dysregulation, and provides insights into how these molecular alterations collectively drive cancer phenotypes.

The technologies and procedures being developed in this day and age is for non-invasive cancer detection. New advances in liquid biopsy technologies have enabled new possibilities for non-invasive cancer detection through application of omics technologies. An example is the detection of circulating tumor DNA (ctDNA), enabling detection of cancer-related genetic and epigenetic alterations in blood samples and thus potentially increasing possibilities of early malignancy detection and therapeutic efficacy evaluation [36].

The application of machine learning and artificial intelligence approaches to interpret omics data has drastically improved our ability to diagnose and type cancer. The machine learning algorithms have the capability to detect complex patterns and associations in large omics datasets and hence lead to more accurate and efficient cancer diagnostic instruments [37]. The development of omics technologies is poised to augment their application in cancer detection.

1.4 Machine Learning in Cancer

Machine learning (ML), is greatly revolutionizing how we approach complex and multi-faceted issues and high dimensional data problems. It is a branch in the wider field of artificial intelligence, that allows computers to learn from large volumes of data and then predict or make decisions based on information without having to be programmed for every eventuality. Unlike conventional software systems that use hard and pre-programmed rules to function, machine learning has the amazing ability to learn and improve with increased experience over time. This feature renders Machine Learning extremely useful for tasks that include pattern classification, informed decision-making, and process automation [38].

Machine learning can be broadly categorized into three main types: supervised learning, unsupervised learning, and reinforcement learning. Figure 1.2 shows the different types of machine learning.

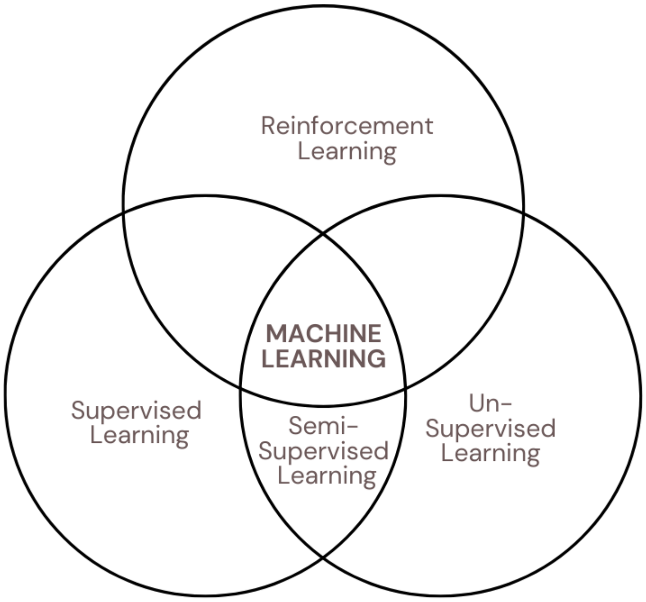

Figure 1.2: Types of Machine Learning Algorithms.

Supervised learning involves training a model on labeled data, where the algorithm learns from input-output pairs to make accurate predictions on new data. Common supervised learning tasks include classification, such as cancer detection, and regression, like predicting patient survival rates. Unsupervised learning, on the other hand, deals with unlabeled data, where the model identifies hidden patterns and structures without explicit guidance. Clustering techniques, such as hierarchical clustering for gene expression analysis, and dimensionality reduction methods, like principal component analysis (PCA), are widely used in biomedical research. Reinforcement learning takes a different approach, where an agent learns optimal actions by interacting with an environment and receiving rewards or penalties based on its decisions. This type of learning is frequently applied in robotics, autonomous systems, and treatment optimization in healthcare. Additionally, semi-supervised learning, which combines both labeled and unlabeled data, and self-supervised learning, where models generate their own labels, have gained traction in recent years, particularly in domains requiring large-scale data analysis, such as genomics and medical imaging. These different types of machine learning provide diverse and powerful tools for tackling complex problems across various fields, including cancer research and personalized medicine.

Over the last decade, machine learning, has revolutionized a vast number of industries, ranging from finance to retail, healthcare to autonomous systems. Within healthcare, however, the influence of ML has been particularly significant and noteworthy, as ML helps doctors to diagnose diseases with unprecedented accuracy, create treatments that are specifically designed to meet individual patient needs, and even forecast the course of different diseases with unprecedented accuracy. Of all of these significant uses, the field of cancer detection is one of the most significant in which machine learning can actually save lives and have a tangible impact on outcomes for sufferers and allows us to mitigate the bias when dealing with patient solutions.

Traditionally, the diagnosis of cancer has relied on a blend of several significant methods, which include a variety of medical imaging tests, extensive pathology tests, and a range of tests in laboratories that are aimed to scrutinize samples. Although all of these methods have been effective in most instances, they do have certain limitations inherent to them in terms of reliability—mistakes can and do occur, results from these tests can often be a while in being produced, and early cancers usually pose significant difficulties for being detected due to their elusive nature. Conversely, machine learning is a cutting-edge solution that uses the amazing ability to scrutinize vast amounts of medical data, to spot patterns that cannot be detected with the naked eye, and to provide diagnoses that are not only quicker but more accurate [39].

At its most basic level, machine learning, works by taking algorithms and training them extensively on past data so that they can learn to recognize certain and significant features. For example, an ML model can be trained to correctly distinguish between benign tumors, which are benign, and malignant tumors, which are cancerous, by extensively studying thousands of past cases. The more data that the model is given to process and analyze, the more proficient it will be in making accurate predictions for future cases. This amazing ability to learn from data and improve over time places ML in a revolutionary position to revolutionize cancer detection since it greatly enhances accuracy in diagnoses while, at the same time, lowering the overall workload placed on medical professionals who have to interpret such results.

Developing a machine learning model for cancer detection requires a thorough and systematic approach to ensure accuracy and reliability. First and foremost, a vast amount of medical data is to be compiled, not just images but genetic data and detailed patient records as well, all of which have to undergo a rigorous process of cleaning to eliminate any inconsistencies or mistakes that would taint the validity of the data. After this collection process, important features like tumor size or shape are systematically extracted to enable the model to focus on the most relevant information to its mission. Once the model has been trained using this cleaned dataset, it is then thoroughly tested using new and unseen data to assess how well it can perform in detecting cancer. If the model proves to meet accuracy measures in these tests, then it can then be deployed with confidence in actual clinical environments, where it is an important aid in helping doctors to arrive at quicker and more reliable diagnoses and hence improve overall efficiency in detecting cancer in medical practice [40].

That being stated, however, one should not deny that machine learning (ML) is not without faults and limitations. One main issue that can occur is that models can be biased, particularly when trained using data sets that have a narrow scope or do not reflect the diversity of a real-world population. It should be noted that deep learning models are essentially black boxes and thus explaining their decision-making to healthcare professionals like doctors is extremely challenging. All that being said, notwithstanding all these different challenges and barriers, it is reassuring to see researchers working hard to improve Machine learning models interpretability. They are working hard to see that the models generalize well to a very wide variety of different patient populations [41].

The application of machine learning (ML) in the niche field of cancer detection is a breathtakingly revolutionary development in oncology. This is owing to machine learning having a distinctive capability to meaningfully leverage enormous processing power that is readily available in modern technological resources. Through this capability, machine learning facilitates a detailed examination of exceedingly complex datasets to significantly improve diagnostic processes entailed in cancer detection. Machine learning has been associated with significant successes through advanced algorithms that are skilled to make crucial differentiation between cancerous patterns of medical images and different medical images with a rich diversity of patterns and patterns found in complex gene expression and complicated patient histories. Indeed, it has been proven that advanced machine learning algorithms are more effective than traditional diagnostic methods that have been used for scores of decades. By adapting cutting-edge methods like convolutional neural networks (CNNs) and support vector machines (SVMs), scientists and researchers have made tremendous progress in precisely diagnosing different forms of cancers like breast cancer, lung cancer, and prostate cancer. They are attaining diagnostic accuracies that were considered impossible. One good example of this development is a recent AI-based system that was said to register a stunning accuracy rate of 97% in lung cancer diagnosis by extensively examining tissue specimens. This quite vividly demonstrates enormous potential that machine learning technology has to revolutionize and fundamentally restructure the practice of cancer diagnostics [42] [43].

One of the most significant and impactful contributions that has been developed from the domain of machine learning, which is also popularly abbreviated as ML, is its phenomenal and pioneering usage in the crucial field of medical imaging. To be more precise, deep learning models, with a specific focus on convolutional neural networks, which are also known as CNNs, have been widely and successfully applied to interpret a varied assortment of radiological images. This encompasses, but is not confined to, mammograms, CT scans, and MRIs. These extremely advanced models possess the astounding capability of identifying faint abnormalities and irregularities that can readily escape the sharp eyes of even the most experienced and competent human practitioners operating in the domain of medicine. For example, CNN-based methodologies have outstandingly recorded sensitivity and specificity rates of more than 96% in identifying early-stage breast cancer by scrupulously analyzing histopathological images, as has been reported in previous studies [44]. Likewise, newer methodologies in the domain of machine learning, including radiomics, are gaining considerable traction in their endeavors to mine high-dimensional features from various medical images, thus revealing disease-related patterns that are not visible to the naked eye of human observers [45].

Outside of imaging modalities, machine learning (ML) has become a revolutionary in the genomic analysis field, especially for cancer detection. With advanced algorithms having the powerful ability to thoroughly scan large gene expression data to categorize tumors into different and particular molecular signatures. Among different methodologies used in this context, artificial neural network (ANN) methods have been used with great success for different cancer prediction with a high accuracy rate of 96% for different cancers including mesothelioma [42] [44]. With this great accuracy, it is attained by identifying key genetic markers that are informative and relevant. Apart from this, ML models have also been found to hold great potential to stratify patients based on different levels of cancer risks to individual patients. This is done through synthesising and deeply analysing trends obtained through electronic medical records (EMRs) and other relevant clinical sources [43]. All of these developments are combined in a way to play a crucial role in enabling cancer detection in advance and also in formulating individualized screening methods based on individual patients' requirements.

The exceptional scalability and impressive flexibility that are typical of machine learning (ML), make it a very powerful tool in the quest to address the complex challenges involved in cancer detection. Unlike traditional methods that are prone to rely heavily on human experts and professionals and are hence susceptible to variability and inconsistency in their results, ML algorithms are blessed with a special advantage. They can learn and refine themselves with each and every additional and incoming input of information. Consequently, these algorithms become dramatically more effective with time as they get to see and deal with more information. This special ability has been found to be highly useful in the very important mission of minimizing false positives and false negatives in a variety of cancer screening programs. Artificially powered systems, for instance, have been able to reduce the rate of false positives in reading mammograms by a significant 6% while still maintaining a very high diagnostic accuracy as supported by different studies [45].

Why Machine Learning Stands Out in Cancer Detection

- **Enhanced Accuracy:** Machine learning algorithms achieve high diagnostic accuracy, often surpassing human specialists. For example, AI systems have demonstrated up to 98% sensitivity and specificity in detecting prostate cancer [42].
  - ML models can outperform radiologists in detecting specific cancer types with higher accuracy.
  - This highlights the potential of AI in improving diagnostic performance.

- **Early Detection:** ML models can identify subtle patterns or biomarkers that indicate early-stage cancers, significantly improving patient survival rates for aggressive cancers like pancreatic or lung cancer [43] [45].
  - Early-stage cancers, such as pancreatic, can often be undetected by traditional methods, but ML algorithms can detect early biomarkers.
  - This can lead to earlier interventions and significantly better outcomes.
- **Personalized Medicine:** By analyzing genomic and clinical data, ML enables tailored treatment plans that optimize therapeutic outcomes while minimizing side effects [44].
  - Machine learning can predict how a patient will respond to a treatment, leading to personalized therapies.
  - This minimizes adverse effects and maximizes therapeutic efficacy.
- **Efficiency:** ML-powered systems process large datasets rapidly, reducing diagnostic time from days to minutes and expediting treatment decisions [42] [43].
  - AI systems help in processing and analyzing data much faster than traditional methods, reducing turnaround times.
  - This results in quicker treatment decisions, which is critical in time-sensitive cancer cases.
- **Cost-Effectiveness:** By focusing resources on high-risk populations through stratified screening programs, ML reduces unnecessary tests and improves healthcare efficiency [45].
  - ML helps identify high-risk individuals early, reducing the need for costly, unnecessary diagnostic tests.
  - This leads to a more efficient use of resources and reduces overall healthcare costs.
- **Error Reduction:** AI significantly lowers false positives/negatives in diagnostics. For example, deep learning models have improved the accuracy of breast cancer detection while reducing diagnostic errors [44] [45].
  - Deep learning models reduce the rate of false positives in breast cancer screening, thus avoiding unnecessary biopsies.

  - This improves both the efficiency and accuracy of cancer detection.

Machine learning's integration with multimodal data is transforming oncology by enhancing diagnostic precision, enabling early detection, and personalizing treatment strategies. ML models surpass traditional methods in accuracy, often achieving diagnostic sensitivity and specificity levels exceeding those of human specialists. Their ability to identify subtle biomarkers in genomic, imaging, and clinical data allows for the early detection of aggressive cancers, significantly improving survival rates. By analyzing multi-omics data, ML facilitates personalized medicine, optimizing treatment efficacy while minimizing adverse effects. Additionally, AI-driven systems streamline large-scale data analysis, reducing diagnostic turnaround times and expediting treatment decisions. This not only enhances efficiency but also makes cancer care more cost-effective by prioritizing high-risk patients and reducing unnecessary testing. Furthermore, deep learning models reduce diagnostic errors, minimizing false positives and negatives in cancer screening. Despite existing challenges like data heterogeneity and interpretability, advancements in self-supervised learning, federated learning, and explainable AI will continue refining ML's role in multimodal oncology research, ultimately leading to improved patient outcomes and more effective cancer therapies.

1.5 Problem Statement

Bladder cancer is still a major global health problem owing to its heterogeneous clinical presentations and heterogeneity in survival rates among patients. Despite advancements in diagnostic methods and treatment modalities, precision in survival prediction for patients with bladder cancer is still a major focus for future studies. Tumor stage, molecular signatures, and demographic characteristics like age are among various factors identified as survival determinants. However, age is generally regarded as a generic risk factor in existing literature and has not been explored for what it can do as a unique predictive factor on examination with large multi-omics data.

The integration of multi-omics data sets that consist of mRNA, CNA and DNA methylation data has the capability to enhance our understanding of complex biological processes connected with bladder cancer. Nevertheless, the high dimensionality and internal variability of these data sets and widespread missingness of points are major challenges to refining feature choice and creation of prediction models. Consequently, this has limited clinical utility of prediction models and impeded detection of clear-cut biomarkers that would increase accuracy of prediction.

In addition, while traditional machine learning methods have been used in cancer prediction, they are often unable to properly capture complex interplay between clinical features and omics data. Advanced machine learning methods like Graph Neural Networks (GNNs) offer a great chance to address these limitations by having an ability to clarify complex interrelationships and derive valuable insights from highly connected data [46]. However, there has not been a deep exploration of using GNNs for survival prediction in bladder cancer with special focus on age as a key prognostic factor. In summary, the key problems that this study seeks to address are:

- The under-exploration of age as a specific prognostic indicator in bladder cancer when integrated with multi-omics data.
- The challenges in effectively selecting and reducing features from complex, high-dimensional multi-omics datasets to build reliable predictive models.
- The need for advanced machine learning frameworks, such as GNNs, to capture the non-linear and complex interactions among clinical variables and omics data.
- The requirement for interpretable models that provide actionable insights into the biological mechanisms of bladder cancer, thereby facilitating personalized therapeutic approaches.

By addressing these challenges, this research aims to establish an optimal age cut-off for predicting bladder cancer outcomes, develop a robust feature selection framework for multi-omics data, and leverage advanced GNN architectures - enhanced by SHAP-based interpretability - to improve survival prediction and inform personalized treatment strategies.

1.6 Research Objective

1.6.1 Age as a Prognostic Indicator: An Investigation

The main objective of this current study is to determine the importance of age as a key predictive variable in predicting survival in bladder cancer patients. While most previous studies treated age as a general risk factor, this study aims to determine the most critical cut-off point for age when utilized as a prediction variable in combination with multi-omics information.

Specifically, the study formalizes the following hypotheses:

- The integration of multi-omics information reveals that increasing age is a strong survival predictor.
- There is an optimal cut point that is capable of classifying people according to their differential survival.
- The establishment of this cut-point should improve both clinical understanding and biological knowledge about the role played by aging in bladder cancer prognosis.

To achieve this aim, the study utilized the *log-rank test* and *Kaplan-Meier survival analysis* in order to compare various cut-off values by age. Using this iterative process, it was ascertained that the most favorable cut-off point was **64**, which demonstrated a statistically significant difference in the survival cohorts.

1.6.2 Formulating a Solid Foundation in Feature Selection

The complexity in multi-omics data, with its high dimension, inherent biological variability, and substantial numbers of missing values, calls for a careful approach in feature selection. The main aim in this part of the study is developing a strong feature selection framework that improves the dataset by ensuring that the predictive model incorporates only the most salient and informative features.

Firstly, we perform data cleaning to improve the quality and consistency of the data. Then we need to apply feature selection techniques to be able to distinguish between the relevant features which will have an impact on the final predictor variable and the redundant variable which are adding noise to the dataset. Using the optimum features the dimension of the dataset should be reduced by choosing the most relevant features, thus highlighting features that play a critical role in the prediction process, and eliminating unnecessary and redundant information.

By this way of systematic improvement in the dataset, this framework aims at increasing the efficacy and predictability of the model, thus making it more robust and relevant for clinical usage in bladder cancer prediction.

1.6.3 Utilization of Graph Neural Networks

The current study makes use of advanced Graph Neural Networks (GNNs) in order to demonstrate the intricate relations existing in multi-omics information. GNNs offer a strong framework in defining the biological layers and their inter-dependencies [46].

In our study we have made use of:

- Graph Convolutional Networks (GCNs) were used because of their ability to capture neighborhood patterns through fixed weight aggregation [47].
- Graph Attention Networks (GATs) incorporate the use of the attention mechanism which allows them to prioritize influential neighbors, improving node classification and link prediction accuracy [48].
- GraphSAGE was used because it employs localized neighborhood sampling and feature aggregation to generalize to unseen nodes or entirely new graphs. By sampling a fixed number of neighbors per node and applying pre-trained aggregator functions (e.g., mean, LSTM, or pooling), it generates embeddings for new nodes based solely on their immediate connections. This approach drastically reduces computational overhead, enabling scalability to massive or dynamic graphs [49].

In our study, we identified that GraphSAGE is the best-performing model, which showed improved capacity in predicting patient survival outcomes from multi-omics data and age-related features.

1.6.4 Ensuring Model Interpretability

To enhance the clinical relevance of the predictions produced by the model, this study combined a comprehensive SHAP (SHapley Additive Explanations) explanation with the best-performing GraphSAGE model. The purpose of this interpretative explanation was to:

- Identify critical markers that significantly influence the predictions made by the model.
- Several genes that were particularly important in determining survival were identified.

The results of this interpretive study emphasized the relevance of the chosen biomarkers, thus providing practical implications for future studies that seek to uncover the biological determinants that predict bladder cancer prognosis.

1.6.5 List of Contributions

The main goal of this study is to bridge the gap between informatics and medical practice by improving clinical decision-making through the use of advanced data analysis techniques. The key contributions of this research are as follows:

- **Informatics Contribution:** Development and application of machine learning techniques, such as graph neural networks (GNNs), feature selection, and survival analysis, to optimize patient stratification in bladder cancer based on molecular and clinical data.

- **Medical Contribution:** Identification of critical biomarkers (*SNRPN*, *LINC01091*, *DHX36*, etc.) and determination of optimal prognostic age cut-offs, which offer insights into bladder cancer progression and patient survival.
- **Framework for Personalized Medicine:** Creation of a novel methodology that merges multi-omics analysis with survival data to improve clinical decision-making, enabling personalized treatment approaches based on age.
- **Facilitating Precision Oncology:** Use of statistical modeling and machine learning methods to uncover key biological markers, which can guide targeted therapies, including immunotherapy and chemotherapy, for bladder cancer patients.
- **Clinical Decision Support:** Provision of a data-driven framework to support oncologists in categorizing patients based on risk profiles and designing individualized treatment plans, advancing the adoption of precision medicine.

1.7 Novelty of the Research

This research introduces novel methodologies and perspectives that advance the field of bladder cancer prognosis, particularly through the integration of multi-omics data and the application of cutting-edge machine learning techniques. The main novel aspects of the study are as follows:

1. **Empirical Age Cutoff Determination:** The current research brings with it a new method of empirically determining an age cutoff of great prognosis relevance for bladder cancer. Through the combination of Kaplan-Meier survival analysis with iterative log-rank testing, 64 was determined to be the most distinguishing threshold for survival outcomes that has not been reported in existing research studies.
2. **Comprehensive Multi-Omics Feature Selection:** A feature selection technique has been proposed to tackle the intricacies related to multi-omics data. This method promises the identification of only the most related features for

prognosis using combined state-of-the-art approaches and thus improves model performance without losing interpretability from the biological perspective.

3. **Graph Neural Network (GNN) Advancements:** This research expands the application of GNNs in the field of medical prognosis with the improvement of existing architecture as well as introducing their remarkable performance on bladder cancer data. In this context, GraphSAGE is particularly innovative with better generalization of the model as well as better predictive accuracy compared to traditional approaches.

4. **Interpretability through SHAP Analysis:** A improvement of the current study is the use of SHAP analysis to improve the interpretability of predictions from complex machine learning models. This allows clinicians to understand the underlying factors that influence survival predictions, thus improving transparency and clinical relevance.

The novelty of the current study is that it is the first to combine advanced statistical analyses, multi-omics data processing methods, and novel machine learning approaches with the age varaible in bladder cancer, thus constituting a significant leap in the field of bladder cancer prognosis.

Chapter 2

Background

This chapter lays the crucial theoretical and conceptual groundwork for the study by investigating the literature in the fields of cancer informatics, multi-omics analysis, and machine learning. It begins by investigating cancer biology, highlighting the value of high-throughput omics technologies in unravelling the complicated nature of tumors. It then performs a critical appraisal of previous studies that applied omics data for cancer prognosis and classification, qualitatively identifying both strengths and weaknesses. Next, it describes in detail feature selection methods, explaining how their application can reduce the curse of dimensionality that often plagues omics data. It then introduces graph neural networks (GNNs) as an efficient tool in modelling interacting biological entities, especially for gene interactions and pathway-level analysis. Finally, in response to the growing desire for interpretable results in the clinic, it concludes by introducing the new field of explainable AI (XAI). Through the synthesis of previous literature and identification of gaps in what is known, this chapter lays the groundwork for later methodological developments and laboratory procedures that are further described in later chapters.

2.1 Literature Review and Relared Works

Bladder cancer research has seen significant advancements, driven by the integration of multi-omics data, radiomic analysis, and machine learning techniques. These innovations have led to better predictive modeling and personalized treatment approaches. This section explores the most relevant studies and methodologies in bladder cancer research, focusing on tumor mutation burden (TMB), the tumor microenvironment (TME), and radiomics, as well as the integration of these elements into predictive models and treatment strategies.

Radiomic Feature Extraction for Tumor Mutation Burden Prediction
Tang et al. [50] proposed the use of radiomic features extracted from pelvic contrast-enhanced computed tomography (CECT) images to predict tumor mutation burden (TMB) in bladder cancer. In their pilot study with 75 patients, six radiomic features were selected through logistic regression with backward elimination and LASSO regression. Their findings demonstrated the potential of radiomics as a non-invasive alternative to traditional biopsy-based methods for predicting genetic features such as TMB.

Tumor Microenvironment and TMB Integration
Cao et al. [51] explored the role of the tumor microenvironment (TME) in muscle-invasive bladder cancer (MIBC), emphasizing its significance in the prognosis and treatment of the disease. Their study combined TME-related signatures (TMERS) with TMB to enhance prognostic models, revealing how TME can affect the efficacy of immune checkpoint inhibitors (ICIs). This integrated approach suggests that the TME is a critical factor in predicting responses to immunotherapy, and its integration with TMB could personalize treatment strategies.

Blood-Based Tumor Mutation Burden as a Predictor
Nan et al. [52] conducted a systematic review and meta-analysis involving 1,525 patients across 11 studies, focusing on the use of blood-based tumor mutation burden (bTMB) as a biomarker for immunotherapy response. Their analysis found that bTMB was a stronger predictor than tissue TMB, offering a non-invasive alternative for patient stratification. This development is particularly valuable for reducing the need for invasive tissue biopsies, making it easier to predict a patient's response to

immune checkpoint inhibitors.

Multi-Omics Integration and Feature Selection Frameworks

Al-Ghafer et al. [53] introduced a multi-omics integration framework using non-negative matrix factorization (NMF) in conjunction with genetic algorithms. This framework facilitated dimensionality reduction and highlighted key biomarkers necessary for distinguishing TMB in bladder cancer patients. Their work underscored the potential of integrating genomic, transcriptomic, and radiomic data to enhance predictive models, improving the accuracy of cancer prognosis.

Personalized Treatment and Multi-Omics Integration

Chen et al. [54] reviewed various strategies for integrating multi-omics data in bladder cancer to optimize personalized treatment. They demonstrated that combining genomic, proteomic, and metabolomic data could significantly enhance therapeutic outcomes by identifying specific biomarkers that guide treatment decisions. This approach is pivotal in precision oncology, where personalized therapies are tailored to the individual tumor profile.

Graph Neural Networks for Modeling Biological Interactions

Wang et al. [55] applied graph neural networks (GNNs) to model complex biological interactions, such as genetic mutations and protein-protein interactions. Their study demonstrated that GNNs can effectively analyze high-dimensional data, providing insights into cancer biology that are difficult to capture through traditional analysis methods. This technique offers a promising approach for understanding bladder cancer and enhancing the predictive power of cancer models.

Survival Analysis in Bladder Cancer Prognosis

Lee et al. [47] employed survival analysis to identify key prognostic factors for bladder cancer. Their work was instrumental in improving personalized treatment by identifying biomarkers predictive of patient outcomes. Survival analysis methods are essential for risk stratification, enabling the identification of patients who may benefit from more aggressive treatments.

Feature Selection Techniques in Cancer Research

Kim et al. [56] and Smith et al. [49] explored various feature selection techniques in cancer research, demonstrating their effectiveness in removing irrelevant or redundant features from predictive models. This process improves model accuracy by retaining only the most relevant biomarkers, enhancing the efficiency and performance of cancer prognosis models.

Advancements in Deep Learning and Graph-Based Approaches
Garcia et al. [57] utilized deep learning methods to identify biomarkers in oncology, enabling the extraction of complex patterns from large datasets. This approach has significantly enhanced our ability to analyze and interpret cancer data. Chen et al. [58] further contributed by applying graph-based methods to improve model explainability, offering new insights into how cancer models function and making them more interpretable.

Immunotherapy and Predictive Modeling in Cancer Treatment
Davis et al. [59] highlighted the importance of predictive modeling in determining the effectiveness of immunotherapy. By integrating predictive models with immunotherapy strategies, these studies have led to improved patient stratification, allowing for better-targeted treatments in bladder cancer.

Related Works
A number of seminal works have significantly contributed to the evolution of multi-omics integration and prognostic modeling in cancer research. Tang et al. [50] and Cao et al. [51] laid the groundwork for integrating radiomic and TME data with TMB predictions in bladder cancer, providing crucial insights into how these variables influence patient outcomes. Nan et al. [52] expanded this perspective by systematically reviewing the predictive efficacy of TMB in NSCLC, highlighting potential translational applications across cancer types.

In addition, Al-Ghafer et al. [53] illustrated that the integration of non-negative matrix factorization (NMF) and genetic algorithms is not just a feasible method but, more importantly, a very pertinent approach towards feature enhancement, especially in the context of multi-omics studies. This contribution has significantly enriched the scholarly discourse in this area and provided a vital reference point for the design and application of novel methodologies from the integrative point of view. It has set a strong platform for future studies in the area of multi-omics integration.

Follow-up studies by Chen et al. [54] and Wang et al. [55] have investigated novel methods for using multi-omics information in developing personalized treatment approaches that address the unique needs of cancer patients. In addition, studies by

Lee et al. [47] and Kim et al. [56] have provided important insights into the mechanisms that allow for the selection and identification of prognostic markers. These mechanisms include sophisticated statistical methods that optimize the validity and reliability of the results.

There has been a recent trend in scholarly works in observing increased interest in deep learning approaches, with special focus on graph-oriented approaches. Garcia et al. [57] and Chen et al. [58] demonstrated the utility of graph neural networks (GNNs) in capturing the complexity of biological interactions and in biomarker identification. In addition, extensive reviews by Smith et al. [49] summarized the various approaches utilized in multi-omics integration, supporting the progression of more complex and explainable models.

Davis et al. [59] underscored the value of predictive modeling in the scope of immunotherapy responses, a consideration that relates intimately with our focus on bladder cancer prognosis. Collectively, these studies depict the foremost strides in cancer studies and call for further innovation in integrating multiple data sources in order to achieve superior patient results.

Table (2.1) below summarizes the papers along with their methodologies, techniques, and key findings:

Table 2.1: Summary of Literature Review

Author(s)	**Study Focus**	**Methodology/ Techniques**	**Key Findings**
Tang et al. [50]	Radiomic feature extraction from pelvic CECT images in bladder cancer	Logistic Regression (with backward elimination) and LASSO regression on six selected features	Developed a predictive model for tumor mutation burden (TMB) in a feasibility study of 75 patients

Continued on next page

Table 2.1 – continued from previous page

Author(s)	**Study Focus**	**Methodology/ Techniques**	**Key Findings**
Cao et al. [51]	Exploring the tumor microenvironment (TME) in muscle-invasive bladder cancer (MIBC)	Combined TME-related signature (TMERS) with TMB	Provided prognostic information and predicted response to immune checkpoint inhibitors (ICI)
Nan et al. [52]	Systematic review and meta-analysis on TMB stratification in NSCLC	Meta-analysis of 11 studies involving 1,525 patients	Demonstrated that blood-based TMB (bTMB) is a superior stratifier compared to tissue TMB for immunotherapy
Al-Ghafer et al. [53]	Multi-omics integration for distinguishing TMB in bladder cancer	Non-negative matrix factorization (NMF) combined with a genetic algorithm	Successfully reduced data dimensionality while preserving critical biomarkers
Chen et al. [54]	Overview of multi-omics integration for personalized therapy in bladder cancer	Comprehensive review of integration techniques	Provided strategies to optimize personalized therapy approaches
Wang et al. [55]	Application of graph neural networks (GNNs) in oncology	Employed GNNs to process high-dimensional datasets capturing biological interactions	Highlighted the potential of GNNs in analyzing complex cancer datasets

Continued on next page

Table 2.1 – continued from previous page

Author(s)	**Study Focus**	**Methodology/ Techniques**	**Key Findings**
Lee et al. [47]	Identification of prognostic biomarkers for bladder cancer	Survival analysis techniques	Discovered key biomarkers for effective patient stratification based on treatment needs
Kim et al. [56]	Feature selection strategies in oncology research	Review of feature selection methods	Emphasized reducing data complexity while maintaining predictive accuracy
Smith et al. [49]	Feature selection approaches in data-intensive cancer research	Review of feature selection techniques	Discussed methods to balance data reduction with effective prediction
Garcia et al. [57]	Deep learning approaches for biomarker discovery in oncology	Application of deep learning methodologies	Demonstrated improved biomarker discovery and enhanced result interpretability
Chen et al. [58]	Graph-based methodologies in biomarker discovery	Utilized graph-based approaches	Highlighted the potential of these methods to enhance biomarker discovery in cancer research

2.2 Survival Analysis

Traditionally, survival analysis has played a central and critical role in the quest for precise survival probability estimations. Moreover, these methods have greatly expanded our understanding of patient outcomes across a variety of medical situations and settings. Classic methods, such as the *Kaplan-Meier estimator* and the *log-rank* *test*, remain invaluable tools for producing accurate estimates of survival distributions and enabling credible comparisons across patient populations.

In addition to these traditional methods, numerous recent developments in survival analysis have emerged. Notably, the introduction of *multi-state models* and *multi-state frailty models* has greatly expanded the range of analytical tools available to researchers. These innovative approaches are particularly effective in detecting and examining dynamic changes in patient status. As a result, they contribute to the creation of more advanced risk assessment protocols and predictive tools that are essential for clinical environments.

It is important to note that survival analysis methods also provide valuable tools for determining relevant cut-off values, such as the cut-off age used in our study. These cut-off values play a profound role in assessing clinical outcomes in various medical contexts. This utility is particularly striking in bladder cancer, where survival outcomes are influenced by the complex interaction of clinical and molecular parameters. By leveraging advanced analytical tools with substantial expertise, our study is well-positioned to accurately classify patients into distinct risk groups, enabling more precise individualized assessments. This, in turn, allows for the optimization of patient management and tailoring of treatment approaches to maximize outcomes.

2.3 Multi-Omics Data Integration

In parallel with advancements in survival analysis, the field of biological studies has also been revolutionized by unprecedented progress in high-throughput sequencing technologies. These advanced technologies provide researchers with the tools to generate vast and complex multi-omics datasets that are rich in biological information. These datasets encompass a wide range of data types, including genomic sequences that reveal the complexities of gene architecture, transcriptomic profiles that describe dynamic gene expression, and methylomic data that reveal key patterns in DNA methylation—an important regulatory function in genes.

This vast array of biological information offers valuable insights into the complex mechanisms of pathogenesis and aids in identifying potential targets for therapeutic intervention. A paradigmatic example of successful integration across multiple omics fields is the rapidly developing and innovative field of **radiogenomics**. Liu et al. [60] have made significant contributions to this discipline, which involves the systematic integration of imaging and genomic data. This integrated approach enables the construction of unified cancer models that capture a broad range of disease-related factors, greatly improving our ability to accurately characterize tumors. Consequently, it allows for more efficient and personalized therapeutic approaches for cancer patients.

In addition, Miller et al. [61] emphasized the critical role of stringent feature selection methods in overcoming the challenges posed by multi-dimensional datasets. These methods are essential for filtering out background noise, which often contaminates such datasets, and for pinpointing relevant information. The use of advanced computational techniques is instrumental in extracting important biomarkers from complex multi-omics data, ensuring that vital information is not overlooked during the analytical process.

2.4 Integrative Approach

By innovatively combining survival analysis methods with a multi-omics paradigm, our study creates a comprehensive framework that effectively identifies and distinguishes prognostic biomarkers. This approach maximizes the efficacy of patient stratification procedures, enabling more informed and targeted treatment strategies. Moreover, this multidisciplinary and integrative approach not only enhances predictive modeling but also expands our understanding of the complex molecular mechanisms driving bladder cancer.

The developments presented in this study ultimately lead to more clinically informed decision-making, resulting in better patient outcomes for individuals facing this specific health challenge. Through this pioneering integration of survival analysis and multi-omics, we aim to contribute significantly to advancing personalized medicine in bladder cancer.

2.5 Research Gap

In spite of the noteworthy and meaningful advances made in this specific area of research, it is apparent that several significant gaps remain in the literature. One of the most conspicuous gaps is the lack of adequate focus on age as a crucial prognostic variable in bladder cancer. Although age is consistently noted and reported in clinical observations and assessments, a well-defined and systematic cutoff for stratification purposes—particularly in relation to age—has yet to be comprehensively developed using advanced computational methods. Such methods could provide more sophisticated and fine-grained information.

Johnson et al. [59] succinctly emphasize that molecular alterations resulting from aging are generally under-investigated and not explored in sufficient detail. This lack of exploration creates a considerable gap in our overall understanding of how age-related molecular changes contribute to survival outcomes in bladder cancer patients.

Another significant research gap that has been recognized within the field is the challenge researchers face when integrating heterogeneous multi-omics data into a coherent and practical predictive model. Much of the past research has focused on single-modal data, which restricts the applicability of the findings and significantly limits their scope. Additionally, many studies have used simple feature selection methods that fail to adequately capture the complex interactions between different omics layers.

Moreover, while graph neural networks (GNNs) have shown vast potential in a variety of biomedical applications [55, 57], their application in bladder cancer prognostication—particularly when combined with survival analysis methods—remains largely unexplored and untapped.

Addressing these significant gaps is imperative in the quest to develop predictive models that are not only more precise but also more interpretable. Such models will enable accurate stratification of individuals, thus informing personalized treatment strategies tailored to each patient's specific needs. Table 2.2 summarizes all the approaches and findings that are present in the previous studies conducted.

Table 2.2: Summary of Key Research Gaps

Study	**Age as Prognostic Variable**	**Multi-Omics Integration**	**GNN Application**	**Survival Analysis**
Research Gap Requirement	Advanced computational determination of age cutoff and exploration of age-related molecular changes.	Sophisticated integration of heterogeneous omics data.	Incorporation of GNNs.	Seamless integration with survival analysis methods.
Tang et al. [50]	✗	✓	✗	✓
Cao et al. [51]	✗	✓	✗	✓
Nan et al. [52]	✗	✓	✗	✓
Al-Ghafer et al. [53]	✗	✓	✗	✓
Chen et al. [54]	✗	✓	✗	✓
Wang et al. [55]	✗	✓	**P**	✗
Lee et al. [47]	✗	✓	✗	✓
Kim et al. [56]	✗	✓	✗	✓
Smith et al. [49]	✗	✓	✗	✓
Garcia et al. [57]	✗	✓	**P**	✗

Continued on next page

Table 2.2 – continued from previous page

Study	Age as Prognostic Variable	Multi-Omics Integration	GNN Application	Survival Analysis
Chen et al. [58]	✗	✓	**P**	✗
Davis et al. [59]	✗	✓	✗	✓

Legend:

- ✓ (Full): The study fully addresses the criterion.
- **P** (Partial): The study partially addresses the criterion.
- ✗ (None): The study does not address the criterion.

2.6 Contributions

The current study represents a set of novel and revolutionary developments with important implications for bladder cancer prognosis, which is a vital component in the improvement in patient outcomes. The main contributions in this study are outlined below:

- **Age Cutoff based on Survival Analysis:** The current study demonstrates a systematic, empirical method for the determination of the single most important cutoff point that distinguishes individuals according to their survival status. Using a combination of Kaplan-Meier survival assessment with iterative log-rank testing methods, our results establish that the most critical cutoff point is at the age of 64. This is supported by strong statistical methods and has strong clinical implications, marking a valuable advancement in bladder cancer prognostication.

- **Feature Selection Pipeline:** The current study proposes a highly rigorous and robust framework for feature selection in the context of multi-omics datasets, involving a number of critical stages:

    - **Variance Thresholding:** A critical feature selection process component.
    - **Univariate Feature Selection (ANOVA F-scores):** Picking features that play crucial roles.
    - **Recursive Feature Elimination with Cross-Validation (RFECV):** Enables efficient and accurate feature selection.

  The suggested method successfully reduces the increased dimensionality found in datasets without compromising biologically meaningful attributes. This improvement in accuracy, as well as in explainability, results in providing insightful information that is highly valuable for medical practitioners and researchers alike. Previous studies by Kim et al. [56] and Smith et al. [49] highlighted the importance of careful feature selection in multi-omics studies.

- **Utilization of Graph Neural Network (GNN) Architectures:** In this study, we employ established GNN architectures to model the complex correlations inherent in multi-omics datasets for bladder cancer prognosis. Our approach utilizes:

    - **Graph Convolutional Networks (GCNs):** Our GCN implementation uses a linear encoder with batch normalization and ReLU, followed by two graph convolution layers and regularization (dropout and early stopping) to aggregate neighboring information. The final log-softmax layer outputs class probabilities.
    - **Graph Attention Networks (GATs):** Building on a similar encoding framework, the GAT model incorporates multi-head attention in two graph attention layers to selectively focus on key neighbors. Regularization is applied similarly, with a log-softmax activation for classification.
    - **GraphSAGE:** Our GraphSAGE approach also employs the same encoder structure and two graph convolution layers. Its inductive learning capability allows for generating embeddings for unseen nodes, resulting in superior performance in accuracy, F1-score, and AUC.

Based on extensive experimentation, we conclude that the GraphSAGE model, with its in-built inductive learning capability, outperforms all the rest in all major metrics such as accuracy, F1-score, and AUC.

Chapter 3

Preliminary Concepts and Research Context

This chapter provides a detailed description of the experimental pipeline employed in this study, covering all stages from data acquisition and preprocessing to model development and evaluation. It begins by introducing the publicly available multi-omics data obtained from The Cancer Genome Atlas (TCGA), including mRNA expression, miRNA expression, DNA methylation, and copy number alterations (CNA). The preprocessing steps—such as impurity removal, normalization, and imputation of missing values—are then described to ensure the consistency and quality of the dataset. The chapter further outlines various dimensionality reduction and feature extraction techniques, encompassing both statistical and graph-based methods, for identifying relevant features. Subsequently, machine learning and deep learning models are introduced, with a focus on graph neural networks (GNNs) used to leverage gene interaction structures. Finally, the chapter details model training procedures, cross-validation strategies, performance evaluation metrics, and interpretability approaches (e.g., SHAP values and visualization tools) to promote transparency and reproducibility of results.

3.1 Overview

In this study, we take a systematic, data-driven approach to explore the TCGA Bladder Urothelial Carcinoma (BLCA) dataset, leveraging multi-omics data to gain insights into bladder cancer. The key steps in our methodology include:

- **Dataset:** The data, obtained from the CBIO Portal, includes clinical data, CNA, mRNA, and DNA methylation profiles. This diverse set of omic data provides a comprehensive view of the molecular landscape of bladder cancer.
- **Data Preprocessing:** We begin by cleaning and normalizing the data to ensure consistency across samples and omic types. This step is crucial for making the data suitable for further analysis.
- **Survival Analysis:** To assess clinical outcomes, the Kaplan-Meier estimator is used to estimate survival functions, helping to interpret the impact of selected features on patient survival.
- **Feature Selection:** To reduce dimensionality and enhance model performance, we apply several feature selection techniques:
  - *Variance Thresholding:* Removes features with little variation, which are unlikely to contribute to model prediction.
  - *Univariate Selection:* Uses statistical tests (e.g., ANOVA F-Score) to select features most related to the target variable.
  - *Lasso Regularization (L1-Based):* A regularization technique that both selects important features and prevents overfitting by shrinking less important feature coefficients to zero.
  - *Recursive Feature Elimination with Cross-Validation (RFECV):* An iterative method that removes the least important features based on model performance, validated using cross-validation to ensure stability.
- **Machine Learning Models:** After selecting the most informative features, machine learning models are built to identify potential biomarkers and predict clinical outcomes. We use a variety of algorithms and validate their performance using standard evaluation metrics.

This combination of methods allows us to effectively handle the complexity of multi-omics data and extract meaningful insights that could enhance our understanding of bladder cancer biology and improve clinical predictions.

3.2 Feature Selection Methods

3.2.1 Variance Threshold

This method of feature selection is employed in high dimensional data where it is believed that features with low-variation are believed not to be very important to the model output. The variance of the features is calculated as follows

$$\text{Var}(x_j) = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2 \tag{3.1}$$

In this equation, $\text{Var}(x_j)$ represents the variance of the j-th feature, n is the number of observations, x_{ij} is the value of the j-th feature for the i-th observation, and $\bar{x}_j$ is the mean of the j-th feature across all observations. Variance thresholding is a technique used to remove features with low variance, which may not contribute significantly to the predictive power of a model.

Features that fall below the variance threshold are then eliminated. This process helps us to reduce the dimensionality of multi-omics data and keep only the features (genes in this case) which contribute to the predictive power of the model so we essentially do this to keep rid of noise in data.

3.2.2 Univariate Feature Selection

Univariate feature selection is a method of Feature selection which evaluates each feature to find its relation with the target variable. The select k percentile of features we have used selects the top 'k' percentile of features based on their univariate test scores. In our case, we utilized the ANOVA F-Score, which measures the degree of linear dependency between each feature and the target variable. Features with higher F-score are considered more relevant than those with a lower F-score.

3.2.3 L1-Based Feature Selection (Lasso Regularization)

L1-Based Feature Selection (also knows as Lasso Regularization) is a very important technique in statistical modeling and machine learning when dealing with high-dimensional and multi-omics data. This technique works by being a method for linear regression which performs both regularization and feature selection simultaneously [62]. The Lasso estimator is defined by the following optimization problem:

$$\hat{\beta} = \arg\min_{\beta} \left\{ \sum_{i=1}^{N} (y_i - \mathbf{x}_i^{\top} \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \tag{3.2}$$

In this equation, $\hat{\beta}$ represents the estimated regression coefficients obtained by minimizing the objective function, which consists of two components: the sum of squared residuals and a regularization term. The term y_i denotes the observed response for the i-th observation, while $\mathbf{x}_i$ is the vector of predictor variables for that observation. The regression coefficients β_j indicate the effect of each predictor on the response variable. The non-negative regularization parameter λ controls the strength of the penalty applied to the regression coefficients. This regularization helps prevent overfitting by encouraging sparsity in the model, allowing only the most significant features to influence the predictions. As a result, the model becomes more interpretable and robust, particularly when dealing with high-dimensional datasets.

The model is also shown to minimize the residual sum of squares subject to the sum of the absolute value of the coefficients being less than a constant. Because of the nature of this constraint, it tends to produce some coefficients that are exactly 0 and this helps us with the varaible selection. This method results in reducing the models over-fitting and multicollinearity between the variables which is essential to stable models and gives good results when dealing with high dimensional data.

3.2.4 Recursive Feature Elimination with Cross-Validation

Recursive Feature Elimination with Cross-Validation (RFECV) is a feature selection framework that makes use of Recursive Feature Elimination as well as Cross-Validation. Recursive Feature Elimination (RFE) was introduced by a researcher as a gene selection method based on iterative feature removal [63].

The method starts of with training SVMs on complete set of features. Subsequently, multiple models are trained and the features that are deemed less important to the target variable are recursively removed. To make sure the features selected are stable and to prevent the data leakage we pair the recursive feature elimination technique with cross-validation as this makes sure the training data is divided into training and validation set.

3.3 Machine Learning Models

3.3.1 Kaplan-Meier Estimator

The Kaplan–Meier estimator is well known and highly regarded as a highly prized non-parametric statistic that has the useful role of estimating a survival function from data regarding lifetimes. This useful measure in statistics, known simply as Kaplan-Meier curve, is most commonly used for graphical representation of event times. Additionally, this estimator can be understood in two different manners: either as a survival rate that represents a proportion of surviving subjects over time or as a total survival function that summarizes total survival experience [64]. Below is mathematically stated formula defining Kaplan-Meier estimator:

$$\hat{S}(t) = \prod_{i:t_i \leq t} \left(1 - \frac{d_i}{n_i}\right) \tag{3.3}$$

where t_i is the observed time points where events occur, d_i is the number of events (e.g., deaths) at t_i, and n_i is the number of individuals at risk just before t_i.

3.3.2 Random Forest

The Random Forrest classifier is an ensemble learning machine learning model which has received a lot of recognition because of its ability to improve the algorithm accuracy by focusing on the concept of Bootstrap Aggregation (Bagging) on the principle that results in the algorithm growing an ensemble of trees and letting them vote for the most popular class, this results in a better output than a single tree. In this way aggregating the predictions of multiple decision trees, this approach provides enhanced accuracy, robustness, and flexibility [65]. The prediction process of a Random Forest classifier can be represented as follows:

$$C = \text{argmax}_i \sum_{t=1}^{T} \delta(c_i, T_t(x)) \tag{3.4}$$

Given an ensemble of T decision trees $T_1, T_2, \ldots, T_T$, each tree predicts the class label c_{T_i} for an instance x. The final predicted class label C for x is determined through majority voting, where T is the total number of decision trees in the ensemble, $T_t(x)$ is the class label predicted by tree T_t for instance x, and $\delta(c_i, T_t(x))$ is the Kronecker delta function that equals 1 if $c_i = T_t(x)$ and 0 otherwise.

3.3.3 Graph Neural Networks (GNN)

3.3.3.1 Graph Convolution Networks (GCN)

Graph Convolution Neural Networks, which are typically referred to as GCNs, are one of the most important and foundational variations of graph neural networks that have been carefully designed to efficiently process and analyze graph-like structured data. Such sophisticated GCNs have a wonderful ability to learn and extract multiple features through a very detailed examination and observation of neighboring nodes surrounding a given node in question. Basically, such innovative networks can be thought of as generalized forms of conventional Convolutional Neural Networks, typically known as CNNs, which were originally developed mainly to process grid-like structures; however, unlike CNNs, GCNs process with a different topology in terms of node connections and have a capacity to process unordered nodes in a graph structure. By extending the convolution operation to be used for graphs, using GCNs enables a process to aggregate information from a variety of different nodes and then distribute that obtained information to all nodes available in a given graph being processed. This extensive and sophisticated process eventually enables the networks to build meaningful representation learning for every individual node placed in the graph framework.

During this sophisticated operation, there are two main inputs—the normalized graph adjacency matrix, known as A', and the node feature matrix, known as F—which are basic building components introduced into the layer. Additionally, a bias vector, known as b, and a weight matrix, known as W, are recognized as trainable variables that have a significant role in controlling the behavior and performance of the layer. The graph convolution operation can be represented by the following formula:

$$H^{(l+1)} = \sigma\left(\hat{A}H^{(l)}W^{(l)} + b^{(l)}\right) \tag{3.5}$$

Where $H^{(l+1)}$ is the output feature matrix of layer $l+1$, $H^{(l)}$ is the input feature matrix of layer l, $\hat{A}$ is the normalized adjacency matrix (with self-loops), $W^{(l)}$ is the weight matrix for layer l, $b^{(l)}$ is the bias vector for layer l, and σ is the activation function (e.g., ReLU).

This formula outlines the graph convolution operation where information is propagated from neighboring nodes through the adjacency matrix and then aggregated at each node to compute new node representations. Figure 3.1 shows a visual representation that illustrates the workflow carried out within Graph Convolutional Models, thus providing a better understanding of how such specific types of networks work and function efficiently.

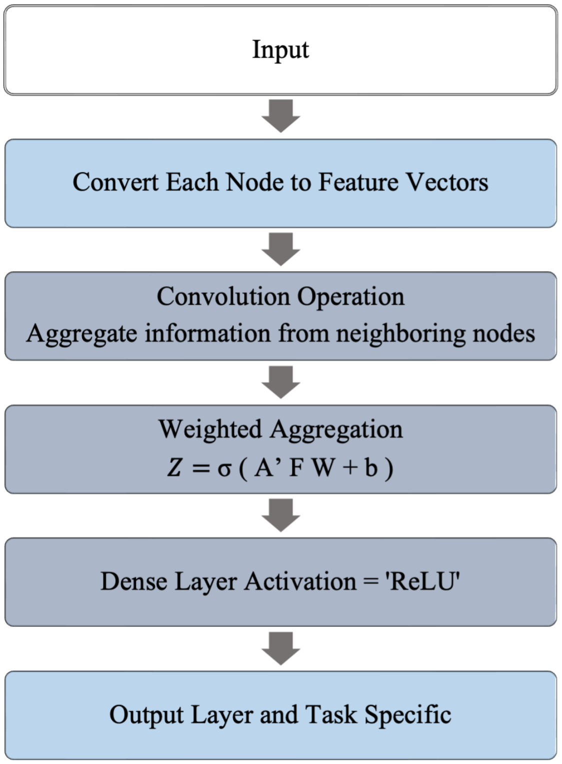
Input
Convert Each Node to Feature Vectors
Convolution Operation
Aggregate information from neighboring nodes
Weighted Aggregation
Z = σ (A' F W + b)
Dense Layer Activation = 'ReLU'
Output Layer and Task Specific

Figure 3.1: The workflow of GCN model.

3.3.3.2 Graph Attention Networks (GAT)

Graph Attention Network (GAT) is a neural network that also works with graph-structured data, designed to extend the mechanism of self-attention to graphs. Unlike Graph Convolution Networks, which rely on convolution layers for aggregating information from neighboring nodes, GATs compute the importance of neighboring nodes dynamically using self-attention mechanisms. This allows GATs to assign different attention weights to the neighbors, making them more flexible and expressive in capturing the relative importance of nodes in a graph.

The self-attention mechanism in GAT assigns attention scores between nodes based on their feature similarities and connectivity. Specifically, the attention score e_{ij} between node i and node j is computed as:

$$e_{ij} = \frac{\exp\left(\text{LeakyReLU}\left(a^T[Wh_i||Wh_j]\right)\right)}{\sum_{k\in\mathcal{N}(i)} \exp\left(\text{LeakyReLU}\left(a^T[Wh_i||Wh_k]\right)\right)} \tag{3.6}$$

Where h_i and h_j are the feature vectors of nodes i and j, respectively, W is the weight matrix, and a is the learnable attention vector. The attention mechanism allows nodes to dynamically adjust their attention weights on neighboring nodes based on their features.

Once the attention scores are computed, the feature vector for node i is updated as follows:

$$h_i' = \sigma\left(\sum_{j\in\mathcal{N}(i)\cup\{i\}} \alpha_{ij} W h_j\right) \tag{3.7}$$

Where α_{ij} is the normalized attention score between nodes i and j, and σ is an activation function (e.g., ReLU). This equation aggregates information from the neighbors of node i, weighted by the attention scores, and updates the feature vector of node i.

GATs enhance this capability by being able to stack a number of such attentional layers to allow nodes to selectively focus on the most informative neighbors with respect to their own feature similarities and connectivity, all this without expensive matrix operations such as inversion, or relying on pre-defined graph structures. The self-attention mechanism, therefore, makes GATs better positioned to work on graphs of various connective patterns and, therefore, more robust in tasks whose node relations are not uniformly distributed.

Another benefit of GATs is their high scalability due to their property of not being dependent on knowing the whole structure of the graph while training. This allows them to perform better on tasks related to noisy, irregular structures [48]. In Figure 3.2, we get a depiction of the workflow of the Graph Attention Models.

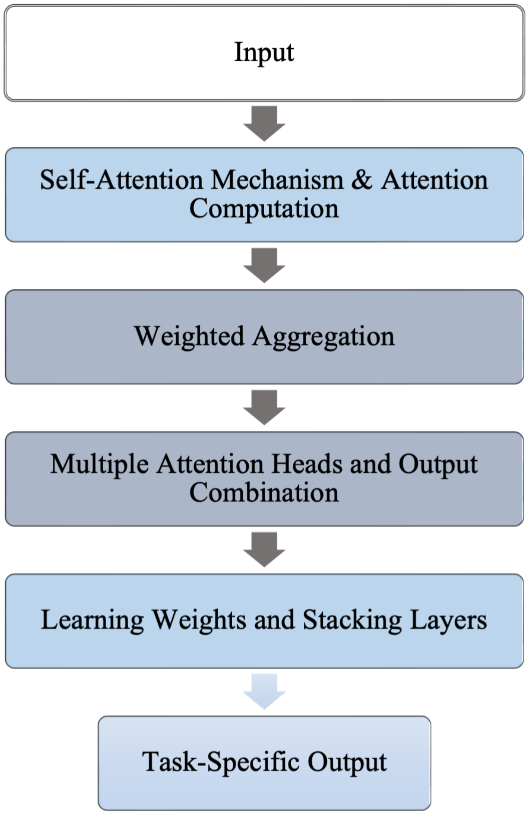
Input
Self-Attention Mechanism & Attention Computation
Weighted Aggregation
Multiple Attention Heads and Output Combination
Learning Weights and Stacking Layers
Task-Specific Output

Figure 3.2: The workflow of GAT model.

3.3.3.3 GraphSAGE Network

GraphSAGE (Graph Sample and Aggregation) is an inductive framework used to learn node embeddings in large graph networks. Unlike the other graph methods, which require knowledge of the entire graph during training, GraphSAGE can generate embeddings for unseen nodes just by using the partial information it gains from node insights. Since it generalizes to unseen parts of the graph, it is especially useful for dynamic graphs when new nodes and edges are added continuously.

The key feature of the GraphSAGE model is to learn a function that aggregates features from a node's local neighborhood. GraphSAGE samples a fixed-size neighborhood for each node and thus avoids processing the entire graph at once, which is efficient for large graphs. This sampling strategy not only reduces computational complexity but also allows the model to scale to graphs with millions of nodes and edges.

GraphSAGE learns the embedding for node v by aggregating information from its neighbors, as follows:

$$h_v^{(k)} = \sigma\left(W^{(k)} \cdot \text{AGGREGATE}\left(\{h_u^{(k-1)} : u \in \mathcal{N}(v) \cup \{v\}\}\right)\right) \tag{3.8}$$

Where $h_v^{(k)}$ is the embedding for node v at the k-th layer, $W^{(k)}$ is the learnable weight matrix at the k-th layer, $\mathcal{N}(v)$ represents the set of neighbors of node v, AGGREGATE is the aggregation function (e.g., mean, LSTM, or pooling), and $h_u^{(k-1)}$ is the embedding of a neighboring node u from the previous layer.

GraphSAGE makes use of aggregation functions such as mean, LSTM, and pooling to integrate the representations of neighbors, which allow the model to learn different aspects of the local graph structure and node features. The aggregation function for mean aggregation is defined as:

$$\text{AGGREGATE}_{\text{mean}}\left(\mathcal{N}(v)\right) = \frac{1}{|\mathcal{N}(v)|} \sum_{u \in \mathcal{N}(v)} h_u^{(k-1)} \tag{3.9}$$

These aggregation functions, together with the sampling strategy, have enabled GraphSAGE to be successfully applied for node classification, link prediction, and graph classification tasks, especially in domains where graphs are large and dynamic [66]. In Figure 3.3, we get a depiction of the workflow of the GraphSAGE Models.

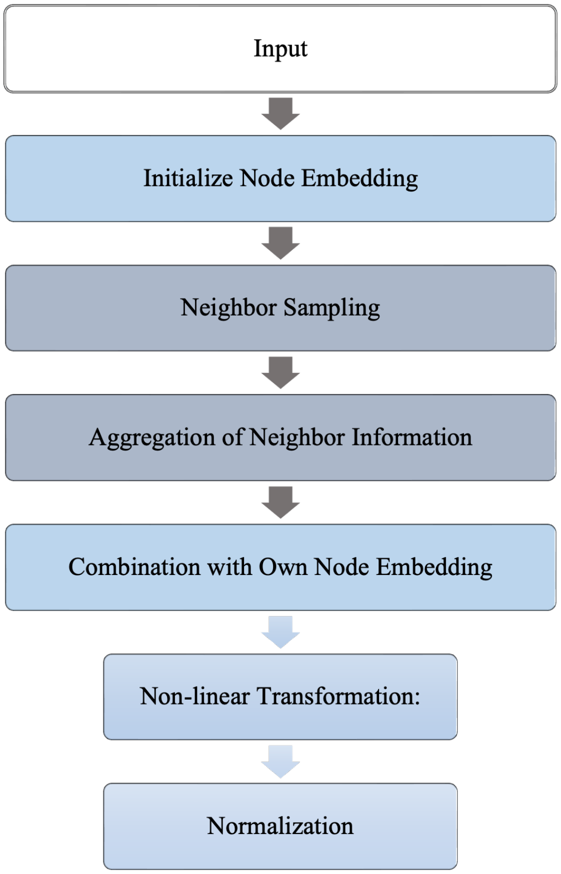
Input
Initialize Node Embedding
Neighbor Sampling
Aggregation of Neighbor Information
Combination with Own Node Embedding
Non-linear Transformation:
Normalization

Figure 3.3: The workflow of GraphSAGE.

3.4 Model Evaluation

A variety of evaluation methods are employed to assess the performance of machine learning models, particularly in the context of classification tasks. These methods provide valuable insights into how well a model generalizes to unseen data, identifying both its strengths and weaknesses. In this section, we discuss the evaluation metrics used to assess the performance of our model, which include accuracy, precision, recall, F1-score, and AUC-ROC. Each of these metrics offers unique insights into different aspects of model performance, which are crucial for determining the model's effectiveness in real-world applications.

3.4.1 Evaluation Metrics

Evaluation metrics are used to give us insights into the performance of machine learning algorithms. The ones we are using are as follows:

- **Accuracy:** Accuracy is one of the most widely used metrics for measuring the performance of a classification model. It is calculated by taking the ratio of correctly classified predictions to the total number of predictions in the dataset. The formula is given by:

  $$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.10}$$

- **Precision:** Precision measures the ratio of correctly predicted positive instances to the total predicted positive instances. It is calculated as:

  $$\text{Precision} = \frac{TP}{TP + FP} \tag{3.11}$$

  Precision is crucial in scenarios where the cost of false positives is high, as it reflects the accuracy of the positive predictions made by the model.

- **Recall:** Recall is the ratio of correctly predicted positive instances to the total actual positive instances in the dataset. The formula is:

  $$\text{Recall} = \frac{TP}{TP + FN} \tag{3.12}$$

- **F1-Score:** The F1-score is the harmonic mean of Precision and Recall. It is calculated as:

  $$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3.13}$$

  The F1-score provides a single metric that balances both precision and recall, making it useful for evaluating models on imbalanced datasets.

- **AUC-ROC:** The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) is a performance measurement for classification problems at various threshold settings. The ROC curve plots the True Positive Rate (Recall) against the False Positive Rate. AUC measures the area under this curve, where a higher AUC indicates better model performance. The formula for AUC is:

$$\mathrm{AUC} = \int_0^1 \mathrm{TPR}(t)\, \mathrm{dFPR}(t) \tag{3.14}$$

In the equations above, TP, TN, FP, and FN stand for True Positive, True Negative, False Positive, and False Negative, respectively. Furthermore, TPR(t) is the True Positive Rate and FPR(t) is the False Positive Rate at the threshold t.

3.5 SHapley Additive Explanations (SHAP)

SHapley Additive Explanations (also known as SHAP) is a tool that is used to help users interpret the predictions of complex machine learning models. SHAP assigns each feature an importance value that quantifies its contribution to a specific prediction, allowing researchers and practitioners to gain insights into complex models, such as deep learning networks and ensemble methods. One of the key advantages of SHAP is its ability to offer both the local (individual prediction) and global (overall model behavior) explanations, making it a widely used technique in high-stakes applications like healthcare [67].

The formula for SHAP values is derived from the concept of Shapley values, which come from cooperative game theory. The Shapley value for a feature f_j in a given instance is calculated as:

$$\phi_j(f) = \sum_{S \subseteq N \setminus \{j\}} \frac{|S|!\,(|N|-|S|-1)!}{|N|!} \left[f(S \cup \{j\}) - f(S) \right] \tag{3.15}$$

Where the SHAP value for feature j, denoted as $\phi_j(f)$, is computed by considering all possible subsets S of features excluding feature j. In this context, N represents the set of all features, and S is any subset of features that does not include feature j. The term $f(S)$ refers to the prediction made by the model using the subset S of features, while $f(S \cup \{j\})$ is the prediction made when feature j is added to the subset S.

The model is also shown to minimize the residual sum of squares subject to the sum of the absolute value of the coefficients being less than a constant. Because of the nature of this constraint, it tends to produce some coefficients that are exactly 0 and this helps us with the variable selection. This method results in reducing the model's overfitting and multi-collinearity between the variables, which is essential for stable models and gives good results when dealing with high-dimensional data.

Chapter 4

Materials and Methods

This chapter details the methodological framework employed in this study, with a focus on the systematic development of predictive models for bladder cancer classification using multi-omics data. It outlines each stage of the process—from data acquisition and preprocessing to feature engineering, model training, and interpretation—providing a comprehensive, reproducible, and scientifically rigorous foundation for the research. By integrating high-throughput data types such as mRNA, miRNA, DNA methylation, and copy number variations, the study seeks to construct machine learning and deep learning models that are not only accurate but also biologically and clinically meaningful. The approach emphasizes the careful curation of input features through advanced selection techniques, followed by the implementation of graph-based neural networks that leverage biological interactions for enhanced prediction. Evaluation strategies, including standard classification metrics and interpretability methods like SHAP, are incorporated to assess model performance and ensure translational relevance. This chapter serves as the technical backbone of the study, bridging computational techniques with biomedical insights to support actionable outcomes in cancer research.

4.1 Dataset

The dataset, comprising 413 samples, was obtained from the CBIO Portal and corresponds to the TCGA Bladder Urothelial Carcinoma (BLCA) cohort. The source data was processed and provided by GDAC Firehose, ensuring high-quality structured genomic information suitable for in-depth analysis [68]. The dataset description is given by Table 4.1 which outlines the structure and key attributes of each data type.

Table 4.1: Dataset description

Omic Name	**Description**
Clinical Data	Clinical data contains 410 samples, they contain the Diagnosis Age, Overall Survival Status and the Overall Survival Months.
CNA	Contains the genes for CNA omics for each Sample ID with 408 samples with 24,776 features.
mRNA	Contains the genes for mRNA omics for each Sample ID with 408 samples with 20,531 features.
DNA Methylation	Contains the genes for DNA methylation omics for each Sample ID with 413 samples with 16,221 features.

4.2 Age Strata Selection

In our study, we began by finding the optimal age strata that we will be using at the target variable moving forward. In our approach, we used a combination of Kaplan-Meier survival analysis and the log-rank test. The dataset included survival time (measured in months), patient age as a continuous variable, and the survival indicator. Iteratively, we evaluated a range of potential age cut-offs at regular intervals (51, 52, 53, etc.) dividing the patients into two groups: patients aged less than or equal to the age threshold and those above the age threshold. Using Kaplan-Meier to calculate the survival curves, and then the log-rank test, allowed us to statistically compare the survival distributions by calculating the p-values for each cut-off.

The optimal age threshold was found by iteratively comparing the different age thresholds, as mentioned before, and choosing the threshold that resulted in the most significant log-rank p-value, indicating the maximum difference in survival probability between the two chosen groups. The results for this approach can be found in Figure 5.1, and it helped us decide the threshold for the age of our patients, leading to the selection of a threshold with statistically greater prediction power in our model.

4.2.1 Kaplan-Meier Survival Analysis

Kaplan-Meier survival analysis was performed to estimate the survival function and visualize the survival curves for different age groups. This non-parametric statistic allowed us to evaluate the probability of survival over time for each group defined by the age threshold. The Kaplan-Meier estimate provides insight into the survival distribution, which is crucial for identifying meaningful differences between the selected groups.

4.2.2 Log-Rank Test

The log-rank test was applied to compare the survival distributions between the two groups created based on the chosen age threshold. The p-value obtained from the log-rank test was used to determine the statistical significance of the difference in survival times between the two groups. This comparison ensured that the selected threshold was the most effective in predicting survival outcomes for our model.

4.3 Feature Selection and Elimination Pipeline

4.3.1 Data Preparation

In our study, as we were dealing with multi-omics data, it was crucial to employ feature selection and elimination techniques to remove noise from the dataset and increase the model's performance, interpretability, and efficiency. The cornerstone of our approach is the feature selection pipeline we used to identify the optimal relevant features for model training. This selection process improved not only the predictive accuracy of our models but also facilitated a deeper understanding of the underlying biological processes by highlighting key features (genes).

4.3.2 Feature Scaling

The importance of feature scaling lies in ensuring that all features contribute equally to the model, preventing dominance by features with larger magnitudes. The following steps were performed for feature scaling:

1. Compute the minimum value:

$$x_{\min} = \min(X), \tag{4.1}$$

2. Apply a shift if the minimum value is non-positive:

$$S = \begin{cases} 0, & \text{if } x_{\min} > 0 \\ |x_{\min}|+1, & \text{if } x_{\min} \leq 0 \end{cases}, \tag{4.2}$$

$$X' = X + S \tag{4.3}$$

3. **Apply the log transformation:**

$$X_{\text{log}} = \log(1 + X'), \tag{4.4}$$

4. **Standardize the transformed data:**

$$Z = \frac{X_{\text{log}} - \mu}{\sigma}, \tag{4.5}$$

where μ and σ are the mean and standard deviation of X_{log}, respectively.

4.3.3 Feature Elimination Methods

To enhance the model's generalization ability and reduce overfitting, we employed various feature elimination techniques. These methods aim to identify and retain the most informative features while eliminating redundant or irrelevant ones. The techniques used are as follows:

- **Variance Threshold**: A threshold of 0.2 was applied to remove features with low variance, as features with minimal variation typically carry less information and may introduce noise into the model.
- **Univariate Feature Selection**: We employed statistical tests such as the ANOVA F-score to select features that exhibit the strongest statistical relationship with the target variable, ensuring that only the most relevant features are retained.
- **L1-based Feature Selection**: By using Lasso regularization (L1 penalty) in logistic regression, we eliminated less important features by driving their coefficients to zero, thereby focusing the model on the most predictive variables.
- **Recursive Feature Elimination (RFE)**: RFE was used in combination with cross-validation (RFECV) to iteratively remove the least important features

based on their impact on model performance. This process ensures that only the most influential features are selected, leading to a more robust and efficient model.

4.4 Model Training Architecture

4.4.1 Overview of Model Architecture

Our model architecture was designed to classify age using multi-omics data. The process starts with loading the pre-processed data after the feature selection pipeline. We then use the K-Nearest Neighbor Algorithm (KNN) to construct graph structures that capture the relationships between data points. Subsequently, three types of graph neural networks (GNNs) — Graph Convolutional Networks (GCN), Graph Attention Networks (GAT), and GraphSAGE — were trained to build comprehensive graph-based models.

4.4.2 Graph Construction with k-Nearest Neighbor (KNN)

The K-Nearest Neighbor (KNN) algorithm was used to create graph structures by connecting each data point with its K nearest neighbors. This construction enables the model to capture the underlying relationships and connectivity between data points, enhancing the ability of the model to learn from the data's inherent structure.

4.4.3 Graph Neural Networks (GNNs) Models

4.4.3.1 Graph Convolutional Network (GCN)

The GCN model begins with an encoder that applies a linear transformation, followed by batch normalization and a ReLU activation function. This allows the model to be less sensitive to the scale of inputs. The model includes two graph convolution layers for aggregating information from neighboring nodes. Regularization techniques such as dropout and early stopping are employed to prevent overfitting. The final output is passed through a log-softmax activation function to compute the class probabilities.

4.4.3.2 Graph Attention Network (GAT)

The GAT model also starts with an encoder, applying a linear transformation, batch normalization, and a ReLU activation function. The model then includes two graph attention layers that use multi-head attention mechanisms to capture the importance of nodes in the graph and focus on the most relevant neighbors. Dropout and early stopping are applied to reduce overfitting. As in the GCN model, the final output is passed through a log-softmax activation function for classification.

4.4.3.3 GraphSAGE Model

The GraphSAGE model follows a similar structure, with an encoder that applies a linear transformation, batch normalization, and a ReLU activation function. The model includes two layers of GraphCONV to learn node embeddings. GraphSAGE is capable of generating embeddings for unseen nodes by using partial information from their neighbors. Dropout and early stopping are again used for regularization, and the output is passed through a log-softmax activation function for classification.

4.4.4 Training Strategy and Evaluation

The training strategy involved using optimization techniques such as Adam, applying the appropriate loss function (cross-entropy loss for classification), and regularization methods like dropout and early stopping to avoid overfitting. We employed Stratified K-Fold Cross-Validation to evaluate the model's performance, ensuring that each fold had a balanced representation of the target classes. Performance metrics such as accuracy, precision, recall, and F1-score were used to evaluate the effectiveness of the model in classifying age groups.

4.4.5 Model Hyperparameter Tuning

To optimize the performance of the models, hyperparameter tuning was performed using Grid Search. We tuned parameters such as learning rate, number of layers, and dropout rate to find the optimal combination that minimized overfitting and maximized model performance. Cross-validation was employed during tuning to ensure the model generalizes well across different subsets of the data. Table 4.2 summarizes the hyperparameters that were tuned using Grid Search.

Table 4.2: Hyperparameter Tuning Results

Model	**Tuned Hyperparameters**	**Best Parameters Found**
GCN	Learning Rate, Hidden Dim, Dropout	LR=0.005, Hidden=32, Dropout=0.2
GAT	Learning Rate, Hidden Dim, Dropout, Attention Heads	LR=0.01, Hidden=16, Dropout=0.4, Heads=4
GraphSAGE	Learning Rate, Hidden Dim, Dropout	LR=0.005, Hidden=32, Dropout=0.2
Random Forest	n_estimators, max_depth, min_samples_split	Estimators=200, Depth=10, Split=4

For model evaluation, we chose the Area Under the Curve (AUC) as the primary metric due to its ability to measure the trade-off between true positive rate (sensitivity) and false positive rate (1-specificity). AUC is particularly valuable in binary classification problems as it provides a comprehensive view of the model's performance across all classification thresholds, making it a reliable metric for models that need to be optimized for both precision and recall. This is especially crucial in imbalanced datasets, where traditional metrics like accuracy may not fully reflect model performance.

4.4.6 Cross-Validation Strategy

We employed Stratified K-Fold Cross-Validation to assess model performance. This technique ensures that each fold contains a proportionate distribution of the target classes, making it particularly useful for imbalanced datasets. Cross-validation helped evaluate the model's generalization ability by training and testing the model on different data subsets, reducing the likelihood of overfitting and providing a more reliable estimate of its real-world performance.

4.4.7 Model Interpretability

To ensure that the results were interpretable and to understand the contribution of each feature, we employed techniques such as SHAP (Shapley Additive Explanations) values and feature importance ranking. For graph-based models, we also explored attention mechanisms within GAT and GraphSAGE to understand how the model focused on different nodes and edges in the graph, providing insights into the most influential factors driving model predictions.

Chapter 5

Experimental Results and Discussion

This chapter presents a detailed account of the experimental outcomes obtained through the computational models introduced earlier. The objective is to not only report the results but also to interpret their significance in the context of biological relevance and clinical applicability to bladder cancer. We begin by analyzing survival trends using statistical tools such as the Kaplan-Meier estimator and the Log-Rank test to determine age thresholds that significantly impact patient outcomes. Through this approach, we identify age 64 as a critical cutoff, which subsequently guides the stratification of patients for downstream analysis. Next, we describe our feature engineering pipeline, which played a pivotal role in managing high-dimensional multi-omics data. By implementing techniques like Variance Thresholding, ANOVA F-score selection, L1-regularization, and Recursive Feature Elimination with Cross-Validation (RFECV), we were able to substantially reduce noise and isolate the most informative features across CNA, mRNA, and DNA methylation datasets. The chapter then transitions into performance evaluation of multiple machine learning models, with a strong emphasis on graph-based approaches including Graph Convolutional Networks (GCN), Graph Attention Networks (GAT), and GraphSAGE. Each model is assessed using robust evaluation metrics such as Accuracy, F1-Score, Precision, Recall, and AUC under a 10-fold cross-validation scheme. Notably, GraphSAGE demonstrated superior performance, attributed to its inductive learning strategy that allows effective generalization to unseen nodes.

5.1 Age-thresholding Experimental analysis

5.1.1 Kaplan-Meier and Log-Rank Test

As shown in Figure 5.1, we have used Kaplan-Meier to calculate the probability of survival over time, stratified by different thresholds of age of the patient. The log-rank test was then used which lets us statistically compare the survival distributions by calculating the p-values for each cut-off, and then finding the optimal log-rank p-value. Using this approach, we identified the optimal threshold that maximally separates the survival curves, which in this case corresponds to patient age 64.

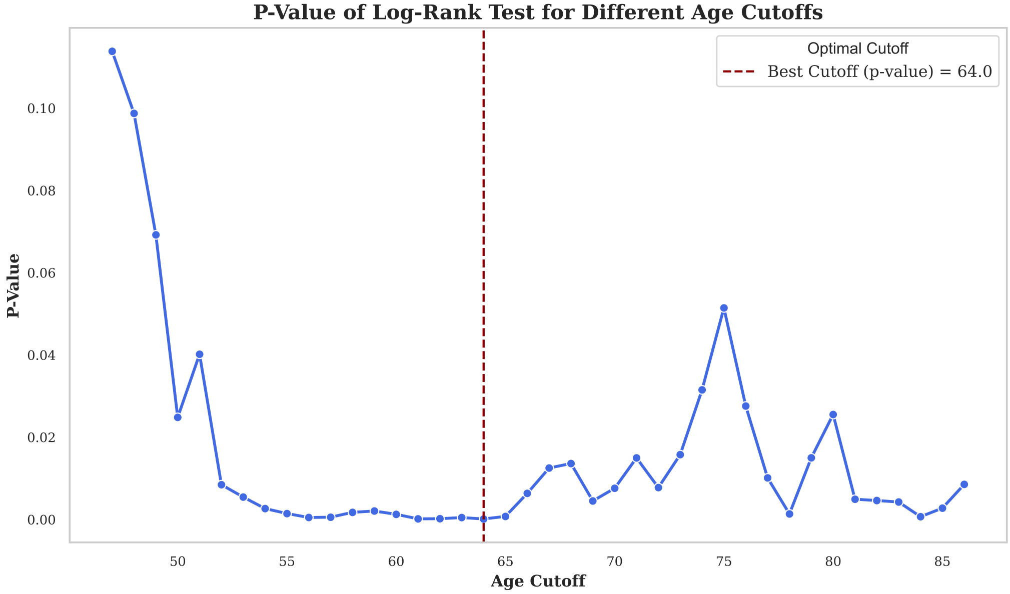

Figure 5.1: Finding Optimum Age Threshold using Kaplan-Meier and Log-Rank Test.

In Figure 5.2, we can see that the age 64 serves as a significant determinant of survival outcomes, with patients above and below this threshold show statistically different survival patterns. This is a good indication of having 64 as as threshold as this gives a good indicator of the predictive power of this threshold in the dataset.

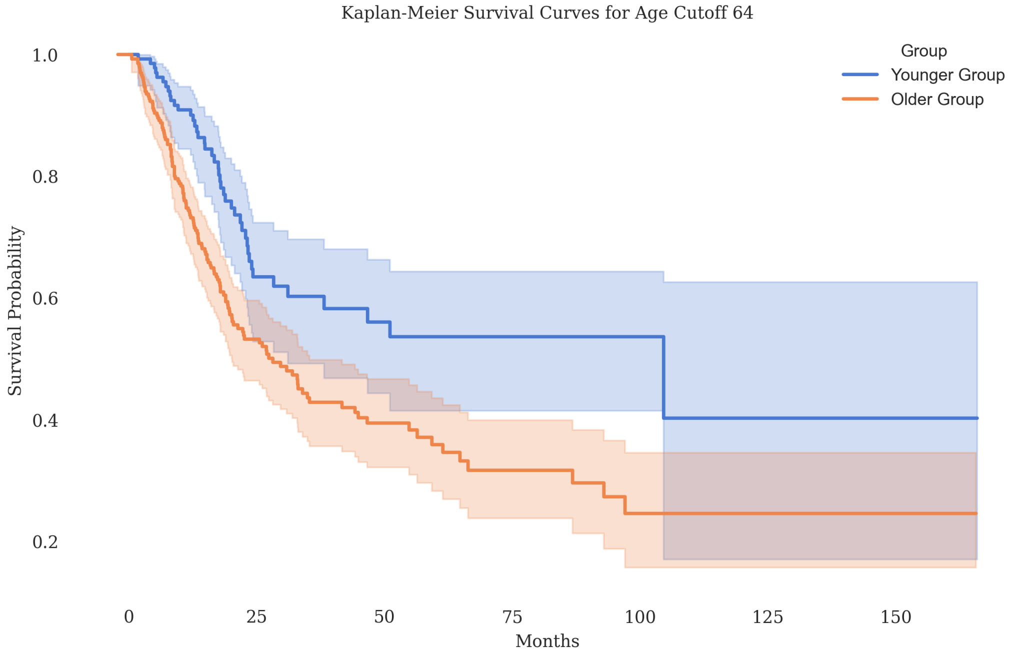

Figure 5.2: Kaplan-Meier Survival Curve based on Age 64 Cut off.

5.1.2 Age Threshold

In Table 5.1, we get insights on the patient mortality based on different age groups thresholds. The Patients have been divided into two groups in this study: Patient aged less than or equal to 64 and patients above the age of 64.

Table 5.1: Mortality statistics based on age groups

Condition	**Patients Deceased**	**Total Patients**	**Percentage Deceased**
Age ≤ 64	46	151	30.46%
Age > 64	134	259	51.74%

Using these statistics, we can confirm that the age variable is a significant factor in determining the survival outcome in patients with bladder cancer as the percentage increases.

5.1.3 Feature Analysis

In our study, as we were dealing with multi-omics data, we had to employ feature selection and elimination techniques to remove noise in the dataset, thereby increasing the model's performance, interpretability, and efficiency. The cornerstone of our study is our feature selection pipeline as demonstrated and referenced in Figure 5.3. in the pipeline we prepare the dataset by catering to missing data. After that we employ our Feature scaling to make sure that the data is normalized thereby preventing the dominance of features with larger magnitudes. After feature scaling we employed various techniques for feature selection such as Variance Thresholding, Univariate Feature Selection (ANOVA F-score method), L1-based feature selection (Lasso Regularization), and Recursive Feature Elimination with Cross-Validation (RFECV). These steps helped us remove features with low variance, eliminate those with weak statistical relationships to the target variable, thereby giving us strongly correlated genes with the target variable.

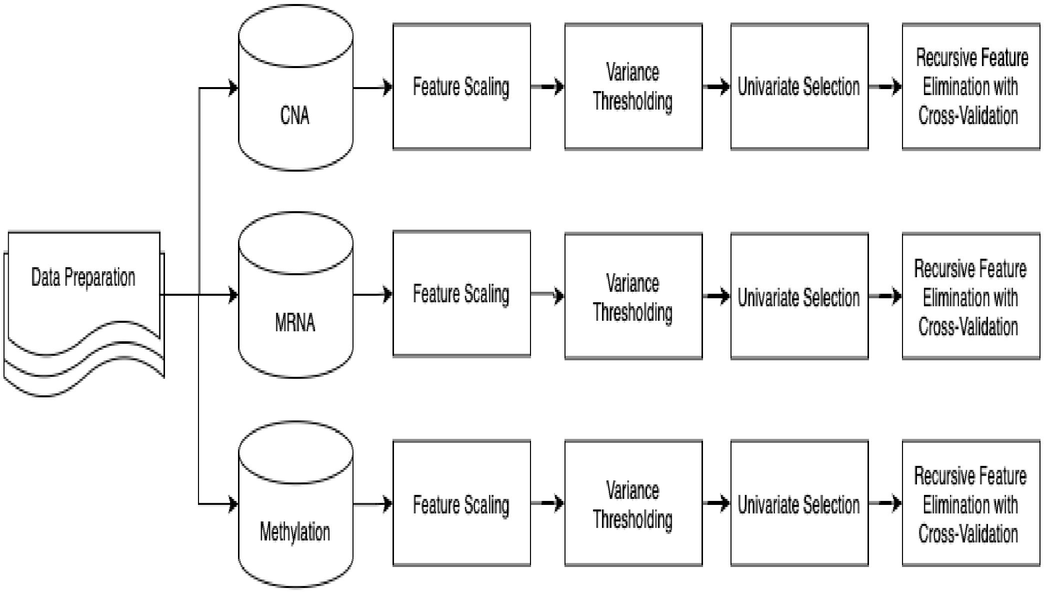

Figure 5.3: Feature Engineering Pipeline Diagram.

Table 5.2: Number of Features Before and After Feature Selection for CNA, mRNA, and Methylation Data.

Data Type	Before Feature Selection	After Feature Selection
CNA	24,776	11
mRNA	20,531	30
Methylation	16,221	76

The effectiveness of our Feature selection pipeline can be observed in Table 5.2 as it turns out not all features are critical in affecting the target variable as significant features are drastically reduced when comparing with the original number of features. Originally, the datasets contained a large number of features 24,776 for CNA, 20,531 for mRNA, and 16,221 for Methylation. However, after applying our feature selection pipeline, the number of retained features drastically reduced to 11 for CNA, 30 for mRNA, and 76 for DNA methylation. This significant reduction highlights the ability of our pipeline to remove redundant or irrelevant features while preserving the most informative ones.

5.2 Model Results and Analysis

5.2.1 Performance Metrics

Table 5.3 presents the results of our models, using performance metrics obtained after 10-fold cross-validation with an 80-20 train-test split. Table 5.3 summarizes Accuracy, F1-Score, Precision, Recall, and Area Under the Curve (AUC) for each model evaluated.

Of all the different models we developed, the GraphSAGE model evidently outshone and was found to be best and most effective among all other alternatives, with a

Table 5.3: 10-Fold Cross-Validation Results for Graph Models.

Model	Accuracy	F1 Score	Precision	Recall	AUC
GCN	0.7264	0.7197	0.7225	0.7264	0.7404
GAT	0.7485	0.7409	0.7558	0.7485	0.7738
GraphSAGE	0.8257	0.8226	0.8289	0.8257	0.8743
Random Forest	0.7340	0.7094	0.7357	0.7340	0.7504

accuracy rate of 82.57%. This performance was further complemented with an F1-score of 82.26% and an area under the curve (AUC) score of 0.8743. Such impressive performance measures evidently reflect GraphSAGE's greater ability to generalize and successfully capture all the intricate patterns within the data, and is most likely due to its cutting-edge inductive learning strategy. This innovative strategy enables it to create meaningful embeddings for unseen nodes within the graph.

5.2.2 Feature Importance

For the best-performing model, GraphSAGE, we conducted a SHAP analysis to identify the top 10 most informative features, as shown in Figure 5.4.

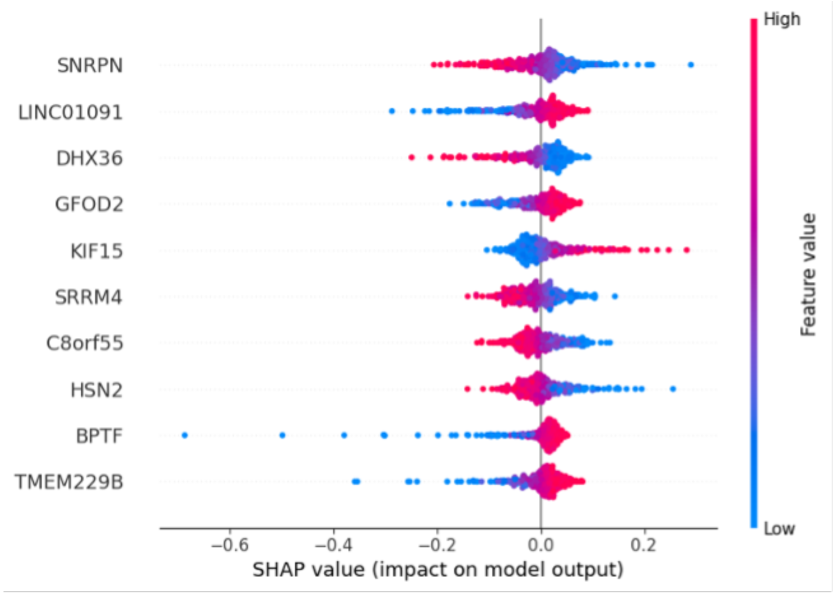

Figure 5.4: SHAP Analysis for top 10 features.

5.3 Discussion and Insights

5.3.1 Age-Threshold Analysis

Our primary research objective was to identify biomarkers linked to age, with a focus on bladder cancer patients. To achieve this, we performed an extensive dataset analysis to determine the optimal age threshold that significantly impacts survival rates.

We utilized the Kaplan-Meier method and the Log-Rank test to identify the optimal threshold that maximally separates survival curves within the sample population. From this analysis, we established that the best threshold is 64 years old.

Individuals aged 64 and below have significantly different survival curves compared to those above 64. The results show a 51.74% mortality rate for individuals above 64 years, compared to 30.46% for those below this threshold. This stark contrast underscores the role of age in bladder cancer prognosis, highlighting how aging-related physiological processes impact treatment outcomes and survival chances.

5.3.2 Model Performance Insights

Among the evaluated models, the GraphSAGE model demonstrated the highest performance across all metrics. With an accuracy of 82.57%, an F1-score of 82.26%, and an AUC of 0.8743 (Table 5.3), GraphSAGE effectively captured complex relationships within the dataset.

GraphSAGE's superior performance can be attributed to its inductive learning approach, which enables better generalization to unseen nodes. Given our dataset's imbalanced nature, relying solely on accuracy is insufficient. Consequently, we assessed model performance using F1-score, Precision, Recall, and AUC to ensure balanced evaluation.

5.3.3 Comparing Graph Models

We observed a clear trend: increased model complexity resulted in improved performance. The GCN model, being simpler in design, recorded the lowest performance, with an F1 score of 71.97% and an AUC of 0.7404. In contrast, the GAT model, which leverages an attention mechanism, achieved improved scores—74.09% (F1) and 0.7738 (AUC)—highlighting the benefits of selective node weighting.

GraphSAGE outperformed all other models due to its inductive learning capabilities, which facilitate better generalization for unseen nodes. Although the Random Forest model achieved an AUC of 0.7504, its lower F1 score of 70.94% reflects struggles with precision and recall, likely due to dataset imbalance.

5.3.4 SHAP Analysis Insights

The SHAP plot (Figure 5.4) illustrates the contributions of individual genes in the GraphSAGE model's predictions. Notably, genes such as SNRPN, LINC01091, and DHX36 emerged as highly significant, strongly influencing the model's decision-making.

Higher values for these features were linked to an increased likelihood of a patient being classified as older than 64. Further investigation into the biological roles of these genes may offer deeper insights into the survival disparities observed between younger and older bladder cancer patients.

Our study highlights the potential of GraphSAGE in effectively identifying key biomarkers linked to age in bladder cancer patients. By integrating survival analysis, feature importance techniques, and advanced graph models, we provided insights that could inform targeted treatment strategies and improve patient outcomes.

5.4 Biological Insights

Our study identified **SNRPN, LINC01091, DHX36, GFOD2, KIF15, SRRM4, C8orf55, HSN2, BPTF, and TMEM229B** as the top genes influencing survival outcomes in bladder cancer, with age-specific implications. These genes are involved in various biological processes such as RNA processing, metabolic regulation, chromatin remodeling, and cell cycle progression [69].

The following genes have previously been linked to cancer and may serve as potential biomarkers or therapeutic targets:

- **SNRPN**: Altered methylation patterns of SNRPN have been associated with the development of germ cell tumors, reflecting abnormal cell differentiation commonly observed in cancer [70] [71].
- **LINC01091**: This gene promotes gastric cancer progression through its involvement in the miR-128-3p/ELF4/CDX2 pathway, suggesting its potential as a therapeutic target [72].
- **DHX36**: As a nucleic acid helicase, DHX36 is linked to cancer progression; in breast cancer, its expression correlates with patient survival, while its knockdown in lung cancer models enhances tumor growth and drug resistance by modulating multiple signaling pathways [73] [74].
- **KIF15**: Overexpression of KIF15 promotes the G1/S phase transition and tumor progression in breast cancer, correlating with larger tumor size, metastasis, and poor prognosis. Its knockdown suppresses proliferation and key oncogenic signaling pathways [75] [76].
- **SRRM4**: Critical for microexon inclusion during RNA splicing, SRRM4 is consistently silenced in tumors, leading to enhanced proliferation. In neuroendocrine prostate cancer, its expression correlates with aggressive disease and poorer survival outcomes [77] [78].
- **C8orf55**: Identified via proteomic analyses, C8orf55 shows higher expression in colon, stomach, and breast cancers, supporting its role as a potential diagnostic and prognostic biomarker [79].

- **BPTF**: Implicated in regulating cell proliferation and survival, BPTF is overexpressed in bladder and lung cancers, where it is associated with poor prognosis and may represent a novel therapeutic target [80] [81] [82].
- **TMEM229B**: Upregulated in bladder cancer, TMEM229B is part of a macrophage M1-related gene signature. Its expression is associated with prognosis and may help guide immunotherapy and chemotherapy decisions [83].

Further investigation into the biological roles of these genes may offer deeper insights into the survival disparities observed between younger and older bladder cancer patients, suggesting potential biomarkers for targeted treatments. These genes could provide deeper insights into survival differences between younger and older bladder cancer patients, suggesting potential biomarkers for targeted treatments.

5.5 Technologies and Tools

The following tools and technologies were employed in this study:

- **Python**: The primary programming language used for data processing, machine learning, and visualization.
- **Pandas**: A powerful library for data manipulation and analysis, particularly for handling tabular data.
- **NumPy**: A library for numerical operations, enabling efficient array and matrix computations.
- **SciPy**: Utilized for statistical tests, including the `logrank_test` for survival analysis.
- **Seaborn**: A statistical data visualization library based on Matplotlib, used for creating informative and attractive plots.
- **Matplotlib**: A widely used library for generating plots and figures in Python.

- **Scikit-learn**: A library for machine learning in Python, providing tools for feature selection, preprocessing, and model training. Specific functions used include:
  - `StandardScaler`: For scaling features to have zero mean and unit variance.
  - `VarianceThreshold`: For removing low-variance features.
  - `SelectPercentile`: For univariate feature selection based on statistical tests.
  - `StratifiedKFold`: For cross-validation with stratified sampling.
  - `LogisticRegression`: Used for L1-penalized feature selection.
  - `RFECV`: For recursive feature elimination with cross-validation.
  - `RandomForestClassifier`: Used for classification.
- **Lifelines**: A library for survival analysis, including tools like `KaplanMeierFitter` and `logrank_test` for survival curve estimation and statistical testing.
- **Torch (PyTorch)**: A deep learning framework used for building and training Graph Neural Networks (GNNs) with models such as GCN (Graph Convolutional Network), GAT (Graph Attention Network), and GraphSAGE (Graph Sample and Aggregation).
- **PyTorch Geometric**: A library built on top of PyTorch for deep learning on graph-structured data, including the use of `GCNConv`, `GATConv`, and `SAGEConv` layers for graph convolution operations.
- **KNN (k-Nearest Neighbors)**: A method used for constructing the graph edges based on feature similarity in the dataset.

Chapter 6

Closing Remarks and Future Research

6.1 Conclusion

This work underscores the pivotal role of using data-driven analysis for the stratification of bladder cancer patients according to their ages and provides critical prognostication values. By applying Kaplan-Meier survival analysis and log-rank testing, we concluded that 64 years was the most appropriate threshold. We found a 30.46% mortality rate for those 64 years and below compared with 51.74% for those older than 64 years, thus establishing the relevance of age as a prognostic factor.

In addition, we implemented a rigorous feature selection pipeline, reducing the high-dimensional datasets from tens of thousands of features to a manageable subset. Techniques such as variance thresholding, ANOVA F-scores, L1 regularization, and RFECV were used to improve computational efficiency and model interpretability.

Our work also employed graph neural networks (GNNs)—specifically GCNs, GATs, and GraphSAGE—to capture complex feature interactions. Among these, the GraphSAGE model outperformed others in accuracy, F1-score, and AUC, highlighting the power of graph-based learning strategies.

To improve model interpretability, we conducted SHAP analysis, which identified key biomarkers like *SNRPN*, *LINC01091*, and *DHX36* that influence survival based on age. These findings provide crucial insights into potential biological mechanisms and suggest promising directions for future studies.

The current research introduces a new methodology that merges multi-omics analysis with survival analysis to derive the optimal prognostic age cut-off for bladder cancer. This approach improves patient stratification and provides a platform for studying the interrelationship between molecular and clinical factors, enabling personalized therapy design. Importantly, our model's ability to integrate multi-omics data with advanced graph neural network techniques provides a powerful tool for clinical decision-making. By accurately stratifying bladder cancer patients according to age and molecular profiles, our approach facilitates early intervention and guides the selection of targeted therapies, including immunotherapy and chemotherapy. This precision oncology framework promises to enhance treatment efficacy, reduce adverse effects, and ultimately improve patient outcomes in clinical practice.

Furthermore, our study's biological insights talks about the molecular underpinnings of bladder cancer. Key biomarkers—including SNRPN, LINC01091, DHX36, KIF15, SRRM4, C8orf55, BPTF, and TMEM229B—have been identified as critical factors influencing patient survival. These markers are involved in essential processes such as RNA splicing, chromatin remodeling, and cell cycle regulation, thereby offering promising targets for future therapeutic interventions. Their integration into our analysis not only enhances our understanding of the disease's progression but also underscores the potential for molecular stratification in personalizing patient care.

In conclusion, this research defines 64 years as a significant prognostic factor for bladder cancer. Our findings—achieved through advanced feature selection methods, graph neural networks, and SHAP analysis—provide valuable insights into the molecular complexities of bladder cancer and offer a strong foundation for developing personalized therapeutic strategies.

6.2 Limitations

While this specific research gives us a vast number of important observations about the different biomarkers present in individuals afflicted with bladder cancer, and how these biomarkers relate to individuals' ages, it is both necessary and responsible that we spend a moment to recognize the limitations present in our research due to being inherent. The dataset we used for our research comes from a single point, and this creates serious questions as to whether or not it actually represents or can be extrapolated to the larger and more heterogeneous population of bladder cancer sufferers present in society in general. Additionally, the nature of our research places certain limitations on what conclusions we can draw regarding causal relationships between the biomarkers we have discovered and survival rates in such patients, specifically in light of their respective ages. This fact serves to emphasize, in a very vivid way, the serious need for further research to be undertaken in order to fully corroborate the results we have outlined, and to advance our knowledge in this important and influential field of research. Building upon what has already been established in the framework of this specific study, it is apparent that there is a vast and fertile ground for additional research to be conducted. Such research should specifically target the exploration and elucidation of the intricate and multifaceted relationship between age-related biomarkers that are implicated in the field with regard to patients afflicted with bladder cancer. To better illustrate, the particular genes that have been discovered throughout the research process can provide a foundation for additional investigations designed to create highly sophisticated and highly developed predictive models. Such models could contribute greatly to the early detection of bladder cancer, thereby potentially enhancing patient outcomes.

Through carefully studying how the discovered genes correlate to the different states of the disease, and to what outcomes are present with those states, researchers would be able to contribute greatly to a greater understanding of how bladder cancer progresses over time. Such information acquired can then be used to better develop treatments that are specifically designed to address the individual needs and demands of different ages afflicted with this specific disease. Additionally, by widening the dataset to include a greater number of patients, specifically one that is more heterogeneous and diverse, researchers would be well-suited to provide greater insight into the complex interplay between factors like patient age, genetic factors, and survival rates with bladder cancer.

6.3 Future Directions

The results derived in this study suggest many directions for future studies that could increase the understanding of bladder cancer and lead to better outcomes in patients.

Firstly, increasing the dataset to include a wide range of patient demographics across various geographical regions and healthcare settings would increase the study's overall generalizability. Having a diverse dataset would allow for the determination of the global applicability of the 64-year age cutoff, and if other parameters like ethnicity, genetic predisposition, or environmental factors could influence its predictive value. Additionally, longitudinal studies that follow up on survival and treatment outcomes of the patients after intervals of varying lengths of time may yield better data that can be used in model training, allowing for model adaptation as new patient data arrive.

In the field of modeling, the application of state-of-the-art graph-based methods, such as Temporal Graph Neural Networks (TGNNs) and heterogeneous graph models, holds promising potential in better depicting cancer progression-related dynamic processes. These methods are uniquely poised to tackle both the temporal aspects of cancer initiation and the complex interactions between different omics layers (like genetic, epigenetic, and transcriptomics data) and individual-specific traits over time.

Further studies of the biological roles of the identified biomarkers—especially SNRPN, LINC01091, and DHX36, among others—could allow for the identification of new therapeutic targets. Validation of the biomarkers by in vitro and in vivo studies would provide better insight into their functional significance in bladder cancer disease progression, which may eventually lead to the identification of specific treatment options.

Additionally, the addition of ancillary data types, e.g., imaging data (radiomics and CT/MRI scans), clinical intervention data, and patient-reported outcomes, can substantially benefit predictive modeling. Utilizing a multi-modal approach is expected to better address the heterogeneity of bladder cancer and possibly lead to better prognostication and treatment planning models.

In summary, as machine learning continues to evolve, there is a need to incorporate explainability and interpretability into predictive modeling systems. While models like GraphSAGE have shown promising results, more efforts to make such models more interpretable for clinicians may help their adoption in real clinical settings. This could involve developing user-friendly interfaces that allow clinicians to visualize the contribution of each feature to the model's predictions, thus empowering them to make better-informed decisions.

In conclusion, this study provides significant insight into how survival in bladder cancer relates to age. Future directions point to potential developments towards improving predictive models, discovering new biomarkers, and eventually designing individualized therapies for bladder cancer.

Bibliography

[1] U. Fakhar, B. Elkarami, and A. Alkhateeb, "Machine learning model to predict autism spectrum disorder using eye gaze tracking," *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 4002–4006, 2023.

[2] U. Fakhar, M. Alsmadi, and A. Alkhateeb, "Machine learning model for anxiety disorder diagnosis based on sensory time-series data," in *International Work-Conference on Bioinformatics and Biomedical Engineering*. Cham: Springer Nature Switzerland, 2024, pp. 241–249.

[3] D. Hanahan and R. A. Weinberg, "Hallmarks of cancer: The next generation," *Cell*, vol. 144, no. 5, pp. 646–674, 2011.

[4] B. Vogelstein, N. Papadopoulos, V. E. Velculescu, S. Zhou, L. A. Diaz Jr, and K. W. Kinzler, "Cancer genome landscapes," *Science*, vol. 339, no. 6127, pp. 1546–1558, 2013.

[5] M. R. Stratton, P. J. Campbell, and P. A. Futreal, "The cancer genome," *Nature*, vol. 458, no. 7239, pp. 719–724, 2009.

[6] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 71, no. 3, pp. 209–249, 2021.

[7] I. J. Fidler, "The pathogenesis of cancer metastasis: The 'seed and soil' hypothesis revisited," *Nature Reviews Cancer*, vol. 3, no. 6, pp. 453–458, 2003.

[8] R. Fisher, L. Pusztai, and C. Swanton, "Cancer heterogeneity: Implications for targeted therapeutics," *British Journal of Cancer*, vol. 108, no. 3, pp. 479–485, 2013.

[9] C. Meldrum, M. A. Doyle, and R. W. Tothill, "Next-generation sequencing for cancer diagnostics: A practical perspective," *The Clinical Biochemist Reviews*, vol. 32, no. 4, p. 177, 2011.

[10] V. Kulasingam and E. P. Diamandis, "Strategies for discovering novel cancer biomarkers through utilization of emerging technologies," *Nature Clinical Practice Oncology*, vol. 5, no. 10, pp. 588–599, 2008.

[11] R. A. Gatenby and J. S. Brown, "Integrating evolutionary dynamics into cancer therapy," *Nature Reviews Clinical Oncology*, vol. 17, no. 11, pp. 675–686, 2020.

[12] F. Pettini, A. Visibelli, V. Cicaloni, D. Iovinelli, and O. Spiga, "Multi-omics model applied to cancer genetics," *International Journal of Molecular Sciences*, vol. 22, no. 11, p. 5751, 2021.

[13] J. Mushtaq, R. Thurairaja, and R. Nair, "Bladder cancer," *Surgery (Oxford)*, vol. 37, no. 9, pp. 529–537, 2019.

[14] A. T. Lenis, P. M. Lec, K. Chamie, and M. D. Mshs, "Bladder cancer: A review," *JAMA*, vol. 324, no. 19, pp. 1980–1991, 2020.

[15] O. Sanli, J. Dobruch, M. A. Knowles, M. Burger, M. Alemozaffar, M. E. Nielsen, and Y. Lotan, "Bladder cancer," *Nature Reviews Disease Primers*, vol. 3, no. 1, pp. 1–19, 2017.

[16] K. J. Kiriluk, S. M. Prasad, A. R. Patel, G. D. Steinberg, and N. D. Smith, "Bladder cancer risk from occupational and environmental exposures," *Urologic Oncology: Seminars and Original Investigations*, vol. 30, no. 2, pp. 199–211, 2012.

[17] M. Burger, J. W. Catto, G. Dalbagni, H. B. Grossman, H. Herr, P. Karakiewicz, and Y. Lotan, "Epidemiology and risk factors of urothelial bladder cancer," *European Urology*, vol. 63, no. 2, pp. 234–241, 2013.

[18] S. F. Shariat, J. P. Sfakianos, M. J. Droller, P. I. Karakiewicz, S. Meryn, and B. H. Bochner, "The effect of age and gender on bladder cancer: A critical review of the literature," *BJU International*, vol. 105, no. 3, pp. 300–308, 2010.

[19] M. Horstmann, R. Witthuhn, M. Falk, and A. Stenzl, "Gender-specific differences in bladder cancer: A retrospective analysis," *Gender Medicine*, vol. 5, no. 4, pp. 385–394, 2008.

[20] C. G. A. R. Network, "Comprehensive molecular characterization of urothelial bladder carcinoma," *Nature*, vol. 507, no. 7492, p. 315, 2014.

[21] C. Cordon-Cardo, "Molecular alterations associated with bladder cancer initiation and progression," *Scandinavian Journal of Urology and Nephrology*, vol. 42, no. sup218, pp. 154–165, 2008.

[22] A. Lopez-Beltran, M. S. Cookson, B. J. Guercio, and L. Cheng, "Advances in diagnosis and treatment of bladder cancer," *BMJ*, vol. 384, 2024.

[23] S. Chakraborty, G. Sharma, S. Karmakar, and S. Banerjee, "Multi-omics approaches in cancer biology: New era in cancer therapy," *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, vol. 1870, no. 5, p. 167120, 2024.

[24] K. Meghani, Y. Yu, N. Frydenlund, E. Z. Li, B. Choy, S. A. Abdulkadir, and J. J. Meeks, "Genomic and transcriptomic profiling of high-risk bladder cancer reveals diverse molecular and microenvironment ecosystems," *bioRxiv*, 2024, 2024-12.

[25] Y. Hasin, M. Seldin, and A. Lusis, "Multi-omics approaches to disease," *Genome Medicine*, vol. 9, no. 1, p. 8, 2017.

[26] I. Subramanian, S. Verma, S. Kumar, A. Jere, and K. Anamika, "Multi-omics data integration, interpretation, and its application," *Bioinformatics and Biology Insights*, vol. 14, p. 1177932219899051, 2020.

[27] S. Huang, K. Chaudhary, and L. X. Garmire, "More is better: recent progress in multi-omics data integration methods," *Frontiers in Genetics*, vol. 8, p. 84, 2017.

[28] K. J. Karczewski and M. P. Snyder, "Integrative omics for health and disease," *Nature Reviews Genetics*, vol. 19, no. 5, pp. 299–310, 2018.

[29] S. Richardson, G. C. Tseng, and W. Sun, “Statistical methods in integrative genomics,” *Annual Review of Statistics and Its Application*, vol. 3, no. 1, pp. 181–209, 2016.

[30] T. Stuart and R. Satija, “Integrative single-cell analysis,” *Nature Reviews Genetics*, vol. 20, no. 5, pp. 257–272, 2019.

[31] R. Beroukhim, C. H. Mermel, D. Porter, G. Wei, S. Raychaudhuri, J. Donovan *et al.*, “The landscape of somatic copy-number alteration across human cancers,” *Nature*, vol. 463, no. 7283, pp. 899–905, 2010.

[32] T. I. Zack, S. E. Schumacher, S. L. Carter, A. D. Cherniack, G. Saksena, B. Tabak *et al.*, “Pan-cancer patterns of somatic copy number alteration,” *Nature Genetics*, vol. 45, no. 10, pp. 1134–1140, 2013.

[33] D. H. Kim and K. E. Lee, “Discovering breast cancer biomarkers candidates through mrna expression analysis based on the cancer genome atlas database,” *Journal of Personalized Medicine*, vol. 12, no. 10, p. 1753, 2022.

[34] A. Bhattacharya, R. D. Bense, C. G. Urzúa-Traslaviña *et al.*, “Transcriptional effects of copy number alterations in a large set of human cancers,” *Nature Communications*, vol. 11, p. 715, 2020.

[35] K. Holm, C. Hegardt, J. Staaf *et al.*, “Molecular subtypes of breast cancer are associated with characteristic dna methylation patterns,” *Breast Cancer Research*, vol. 12, p. 36, 2010.

[36] J. C. Wan, C. Massie, J. Garcia-Corbacho, F. Mouliere, J. D. Brenton, C. Caldas, and N. Rosenfeld, “Liquid biopsies come of age: towards implementation of circulating tumour dna,” *Nature Reviews Cancer*, vol. 17, no. 4, pp. 223–238, 2017.

[37] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, “Machine learning applications in cancer prognosis and prediction,” *Computational and Structural Biotechnology Journal*, vol. 13, pp. 8–17, 2015.

[38] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016. [Online]. Available: https://www.deeplearningbook.org

[39] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

[40] J. D. Kelleher, B. Mac Namee, and A. D'arcy, *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies*, 2nd ed. MIT Press, 2020.

[41] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, 2023.

[42] B. Zhang, H. Shi, and H. Wang, "Machine learning and ai in cancer prognosis, prediction, and treatment selection: a critical approach," *Journal of Multidisciplinary Healthcare*, pp. 1779–1791, 2023.

[43] D. Bertsimas and H. Wiberg, "Machine learning in oncology: methods, applications, and challenges," *JCO Clinical Cancer Informatics*, vol. 4, pp. CCI–20, 2020.

[44] K. Kourou, K. P. Exarchos, C. Papaloukas, P. Sakaloglou, T. Exarchos, and D. I. Fotiadis, "Applied machine learning in cancer research: A systematic review for patient diagnosis, classification and prognosis," *Computational and Structural Biotechnology Journal*, vol. 19, pp. 5546–5555, 2021.

[45] D. M. Koh, N. Papanikolaou, U. Bick, R. Illing, C. E. Kahn Jr, J. Kalpathi-Cramer, and F. Prior, "Artificial intelligence and machine learning in cancer imaging," *Communications Medicine*, vol. 2, no. 1, p. 133, 2022.

[46] N. Valous, F. Popp, I. Zörnig *et al.*, "Graph machine learning for integrated multi-omics analysis," *British Journal of Cancer*, vol. 131, pp. 205–211, 2024. [Online]. Available: https://doi.org/10.1038/s41416-024-02706-7

[47] S. Lee, H. Park, and J. Kim, "Prognostic biomarkers in bladder cancer: A survival analysis approach," *Cancer Research*, vol. 81, pp. 3035–3043, 2021.

[48] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," *arXiv preprint arXiv:1710.10903*, 2017. [Online]. Available: https://arxiv.org/abs/1710.10903

[49] J. Smith and M. Jones, "Multi-omics data integration for cancer biomarker discovery," *Journal of Clinical Bioinformatics*, vol. 10, pp. 15–27, 2019.

[50] X. Tang, W. L. Qian, W. F. Yan, T. Pang, Y. L. Gong, and Z. G. Yang, "Radiomic assessment as a method for predicting tumor mutation burden (tmb) of bladder cancer patients: a feasibility study," *BMC Cancer*, vol. 21, pp. 1–9, 2021.

[51] R. Cao, L. Yuan, B. Ma, G. Wang, and Y. Tian, "Tumour microenvironment (tme) characterization identified prognosis and immunotherapy response in muscle-invasive bladder cancer (mibc)," *Cancer Immunology, Immunotherapy*, vol. 70, no. 1, pp. 1–18, 2021.

[52] Z. Nan, W. Guoqing, Y. Xiaoxu, M. Yin, H. Xin, L. Xue, and W. Rong, "The predictive efficacy of tumor mutation burden (tmb) on nonsmall cell lung cancer treated by immune checkpoint inhibitors: a systematic review and meta-analysis," *BioMed Research International*, vol. 2021, no. 1, p. 1780860, 2021.

[53] I. Al-Ghafer, N. AlAfeshat, L. Alshomali *et al.*, "Nmf-guided feature selection and genetic algorithm-driven framework for tumor mutational burden classification in bladder cancer using multi-omics data," *Network Modeling Analysis in Health Informatics and Bioinformatics*, vol. 13, p. 26, 2024. [Online]. Available: https://doi.org/10.1007/s13721-024-00460-7

[54] L. Chen, Q. Zhang, and J. Liu, "Multi-omics integration for personalized therapy in bladder cancer," *Journal of Clinical Oncology*, vol. 39, pp. 450–460, 2021.

[55] Y. Wang, M. Li, and X. Zhou, "Graph neural networks in biomarker discovery: A case study in bladder cancer," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, pp. 1120–1130, 2022.

[56] D. Kim, Y. Choi, and H. Ryu, "Feature selection methods in multi-omics studies for cancer research," *Bioinformatics*, vol. 36, pp. 245–252, 2020.

[57] A. Garcia, P. Martinez, and T. Nguyen, "Deep learning and graph-based approaches in precision oncology," *IEEE Access*, vol. 8, pp. 130 345–130 356, 2020.

[58] X. Chen, L. Zhao, and F. Wang, "Graph neural networks in medical imaging and biomarker discovery," *Computers in Biology and Medicine*, vol. 134, p. 104381, 2021.

[59] R. Davis, K. Patel, and R. Singh, "Predicting immunotherapy response in bladder cancer using machine learning," *Oncology Letters*, vol. 22, pp. 1040–1048, 2022.

[60] H. Liu, Y. Sun, and Z. Li, "Radiogenomics: Bridging imaging and genomics in cancer research," *European Journal of Radiology*, vol. 134, pp. 109–115, 2020.

[61] T. Miller, R. Evans, and D. Johnson, "Advanced computational methods for multi-omics integration in cancer research," *BMC Bioinformatics*, vol. 21, pp. 1–12, 2020.

[62] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, January 1996. [Online]. Available: https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

[63] I. Guyon, J. Weston, S. Barnhill, and et al., "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, pp. 389–422, 2002. [Online]. Available: https://doi.org/10.1023/A:1012487302797

[64] E. L. Kaplan and P. Meier, "Nonparametric estimation from incomplete observations," *Journal of the American Statistical Association*, vol. 53, no. 282, pp. 457–481, 1958. [Online]. Available: https://doi.org/10.1080/01621459.1958.10501452

[65] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. [Online]. Available: https://doi.org/10.1023/A:1010933404324

[66] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," vol. 30, 2017.

[67] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. [Online]. Available: https://doi.org/10.48550/arXiv.1705.07874

[68] cBioPortal for Cancer Genomics, "Bladder urothelial carcinoma (tcga, firehose legacy)," https://www.cbioportal.org/study/summary?id=blca_tcga, n.d. [Online]. Available: https://www.cbioportal.org/study/summary?id=blca_tcga

[69] GeneCards, "Genecards: The human gene database," 2025, accessed: 21-Mar-2025. [Online]. Available: https://www.genecards.org

[70] K. J. Bussey, H. J. Lawce, E. Himoe, X. O. Shu, N. A. Heerema, E. J. Perlman, and R. E. Magenis, "Snrpn methylation patterns in germ cell tumors as a reflection of primordial germ cell development," *Genes, Chromosomes and Cancer*, vol. 32, no. 4, pp. 342–352, 2001.

[71] S. H. Lee, V. Appleby, J. N. Jeyapalan, R. D. Palmer, J. C. Nicholson, V. Sottile, and P. J. Scotting, "Variable methylation of the imprinted gene, snrpn, supports a relationship between intracranial germ cell tumours and neural stem cells," *Journal of Neuro-Oncology*, vol. 101, pp. 419–428, 2011.

[72] Q. Wang, C. Zhang, S. Cao, and et al., "Tumor-derived exosomes orchestrate the microrna-128-3p/elf4/cdx2 axis to facilitate the growth and metastasis of gastric cancer via delivery of linc01091," *Cell Biology and Toxicology*, vol. 39, pp. 519–536, 2023.

[73] Y. Zeng, T. Qin, V. Flamini, C. Tan, X. Zhang, Y. Cong, E. Birkin, W. G. Jiang, H. Yao, and Y. Cui, "Identification of dhx36 as a tumour suppressor through modulating the activities of the stress-associated proteins and cyclin-dependent kinases in breast cancer," *American Journal of Cancer Research*, vol. 10, no. 12, pp. 4211–4233, 2020.

[74] Y. Cui, Z. Li, J. Cao, J. Lane, E. Birkin, X. Dong, and W. G. Jiang, "The g4 resolvase dhx36 possesses a prognosis significance and exerts tumour suppressing function through multiple causal regulations in non-small cell lung cancer," *Frontiers in Oncology*, vol. 11, p. 655757, 2021.

[75] J. Wang, X. Guo, C. Xie, and et al., "Kif15 promotes pancreatic cancer proliferation via the mek–erk signalling pathway," *British Journal of Cancer*, vol. 117, pp. 245–255, 2017.

[76] X. Gao, L. Zhu, X. Lu, and et al., "Kif15 contributes to cell proliferation and migration in breast cancer," *Human Cell*, vol. 33, pp. 1218–1228, 2020.

[77] S. A. Head, X. Hernandez-Alias, J. S. Yang, L. Ciampi, V. Beltran-Sastre, A. Torres-Méndez, and L. Serrano, "Silencing of srrm4 suppresses microexon

inclusion and promotes tumor growth across cancers," *PLoS Biology*, vol. 19, no. 2, p. e3001138, 2021.

[78] Y. Li, Q. Zhang, J. Lovnicki, R. Chen, L. Fazli, Y. Wang, and X. Dong, "Srrm4 gene expression correlates with neuroendocrine prostate cancer," *The Prostate*, vol. 79, no. 1, pp. 96–104, 2019.

[79] H. Kume, S. Muraoka, T. Kuga, J. Adachi, R. Narumi, S. Watanabe, and T. Tomonaga, "Discovery of colorectal cancer biomarker candidates by membrane proteomic analysis and subsequent verification using selected reaction monitoring (srm) and tissue microarray (tma) analysis," *Molecular & Cellular Proteomics*, vol. 13, no. 6, pp. 1471–1484, 2014.

[80] S. Xu and et al., "Bptf as a potential biomarker for bladder cancer," *Journal of Cancer Research*, vol. 25, no. 4, pp. 123–130, 2020.

[81] L. Chen and et al., "Overexpression of bptf in lung cancer correlates with poor prognosis," *Oncology Letters*, vol. 42, no. 2, pp. 55–62, 2021.

[82] X. Li and et al., "Regulation of cell cycle progression by bptf in cancer cells," *Cancer Cell Biology*, vol. 19, no. 3, pp. 210–220, 2022.

[83] Y. Yu, Y. Huang, C. Li, S. Ou, C. Xu, and Z. Kang, "Clinical value of m1 macrophage-related genes identification in bladder urothelial carcinoma and in vitro validation," *Frontiers in Genetics*, vol. 13, p. 1047004, 2022.