Breast Cancer Survival Prediction Using Machine Learning Models Trained on Clinical Data and Gene Expression Profiles

A report submitted in fulfillment of the requirement for the degree of Master of Science

Adeola Odunewu

ABSTRACT

Breast cancer is a complex disease influenced by genetics and environmental factors, making accurate survival prediction essential for personalized treatment strategies. This study integrates clinical data with gene expression profiles to enhance survival prediction using machine learning techniques. Understanding the interplay between clinical attributes and gene expression profiles is a critical challenge in breast cancer research. By merging data from the METABRIC database, this study aims to improve prognostic accuracy. The integration of diverse data types allows for a comprehensive analysis of patient outcomes. The approach involves rigorous data preprocessing and exploratory analysis. Missing values are addressed, and outliers are managed to ensure data integrity.

Principal Component Analysis is utilized for dimensionality reduction, distilling essential features from the dataset. Unsupervised learning techniques, including cluster analysis, unveil underlying patterns within the data. The primary contribution lies in the development of predictive models with substantial accuracy. Random Forest, Artificial Neural Network and Gradient Boosting and XGBoost classifiers are employed, achieving high precision and recall scores. The models effectively stratify patients into survival outcome categories, aiding in treatment decisions. Key implications emerge from the identified influential biomarkers associated with breast surgery types. These biomarkers provide insights into treatment response and guide tailored interventions.

This integration of multi-omics data enhances the precision of survival prediction and strengthens the basis for personalized medicine. In conclusion, this study demonstrates the potential of integrating clinical and gene expression data for accurate breast cancer survival prediction. The machine learning models showcase the efficacy of this approach in improving prognostic accuracy. By using machine learning to uncover and identifying biomarkers and predicting patient survival this study advances the scope of personalized breast cancer treatment.

1. Introduction

Breast cancer is a convoluted and heterogeneous disease that affects a significant number of women worldwide. It involves the anomalous proliferation of cells within the breast tissue, which, if not detected and addressed on time, it can potentially metastasize to other body regions (Hill, H.E., Schiemann, W.P. et al., 2020). In recent years, the integration of machine learning techniques with clinical and genomic data has shown promise in improving breast cancer diagnosis, treatment, and prognosis Yadav et al. (2023). This research will explore the application of machine learning models trained on clinical data and gene expression profiles to predict survival in breast cancer patients and identify influential factors in treatment outcomes. However, for the sake of clarity, it is crucial to understand the concept of Breast Cancer: Breast cancer is a form of cancer that begins in the tissue of the breast (American Cancer Society (ACS) 2022). It occurs when cells in the breast begin to grow and divide abnormally, forming a malignant tumour. Breast cancer can affect both women and men, but it is much more common in women. It is the most common cancer among women, accounting for about 25% of all cancers in women Li, Y., et al. (2023). The causes of breast cancer are complex and multifactorial, involving a combination of genetic, hormonal, environmental, and lifestyle factors. While the precise cause of breast cancer is frequently elusive, various risk factors have been identified. However, it is thought to be caused by a combination of genetic and environmental factors. Some of the risk factors for breast cancer include:

Age and Gender: The risk of having breast cancer increases with age. The majority of breast cancer cases occur in women over the age of 50. Women are more susceptible to breast cancer than men, although men can also develop the disease.

Family History and Genetic Factors: Having a family history of breast cancer, especially in close relatives such as a mother, sister, or daughter, increases the risk. Certain gene mutations, such as mutations in the BRCA1 and BRCA2 genes, significantly raise the risk of developing breast cancer.

Hormonal Factors: Estrogen and progesterone, female sex hormones, play a role in breast cancer development. Factors that increase exposure to these hormones, such as early onset of menstruation (before age 12), late menopause (after age 55), and a history of hormone replacement therapy, can increase the risk.

Personal History of Breast Cancer: Individuals who have previously had breast cancer in one breast have an increased risk of developing cancer in the other breast or developing a new cancer.

Dense Breast Tissue: Women with heavy breast tissue, are known through mammography, to have a higher risk of developing breast cancer. Dense breast tissue makes it more challenging to detect tumours on mammograms.

Lifestyle and Environmental Factors: Certain lifestyle choices and environmental exposures may contribute to the risk of breast cancer. These include excessive alcohol

consumption, smoking, obesity, lack of physical activity, exposure to radiation, and long-term use of oral contraceptives.

In recent years, the scientist has evolved from the traditional tissue biopsy (A tissue biopsy is a medical procedure in which a doctor removes a small piece of tissue from the body for examination under a microscope. Biopsies serve the purpose of cancer diagnosis, assessing its stage, and providing treatment guidance. On the other hand, ctDNA, an abbreviation for circulating tumour DNA, pertains to small DNA fragments released into the bloodstream by tumour cells. As cancer cells continue to multiply and divide, they expel DNA into the bloodstream through processes like apoptosis or necrosis. This DNA holds distinctive genetic information specific to the tumour and can be detected and analysed through a non-invasive blood test. ctDNA contains significant genetic changes or mutations that are indicative of the tumour's origin. These mutations can include single nucleotide variants (SNVs), insertions/deletions (indels), copy number alterations (CNAs), and structural variations. By analysing ctDNA, researchers and clinicians can gain insights into the genetic makeup and evolution of tumours, as well as monitor treatment response and detect minimal residual disease (MRD) or recurrence Sant M., Bernat-Peguera A., Felip, E (2022).

1.2 Breast Cancer Survival Prediction Challenges

Breast cancer poses a critical challenge in healthcare, demanding accurate survival predictions to guide personalized treatment approaches. Conventional prognostic approaches primarily rely on clinical data, yet emerging studies indicate that gene expression profiles hold untapped potential for illuminating the disease's course Malik, V., et al. (2021). By scrutinizing a spectrum of machine learning algorithms, the study aims to assess their efficacy in precise patient survival prediction.

1.2.1 Investigating Breast Cancer Survival and Understanding Influential factors in the Treatment.

The aim of this study is to investigate survival in breast cancer patients using multi-omics data integration, specifically by integrating clinical data and gene expression profiles. Additionally, the study aims to investigate the factors that influence the response to treatment for breast cancer.

The Objectives is to conduct a literature review on machine learning techniques used to analyse genetic profiles and clinical data to predict survival outcomes in breast cancer patients, Address data preprocessing and handling missing data for multi-omics data integration using R and Python, Utilize machine learning to explore the interrelationship between clinical data and gene expression profiles, enabling a comprehensive analysis of breast cancer response to treatment, Identify novel predictive markers and understand the underlying biological mechanisms associated with survival after treatment, Evaluate the accuracy and reliability of the developed predictive models using appropriate validation techniques, Contribute to advancements in breast cancer research and precision medicine by facilitating personalized treatment decisions and improving patient outcomes. Document the study in the form of a master's dissertation.

1.2.2 Insights from Literature and Integrative Multi-Omics Analysis

In this study, existing literature will be reviewed, particularly research conducted by Sammut, SJ., et al. (2021), which revealed that chemotherapy response in breast cancer is influenced by genomic features and tumour ecosystem characteristics. Understanding and targeting these genomic factors, such as TP53 mutations, tumour mutation burden, BRCA mutations, HRD, APOBEC mutational signatures, and chromosomal instability, could improve treatment outcomes. Additionally, the assessment of survival and drug response is vital in determining prognosis. Malik, V. et al. (2021) introduced a multi-omics integrative framework to quantify survival and drug response accurately in breast cancer patients. They employed neighbourhood component analysis (NCA) for feature selection from multi-omics datasets and used a neural network framework to develop prediction models. The proposed strategy effectively combines diverse omics data types, providing valuable prognostic indicators for breast cancer patients. These findings demonstrate the potential of using machine learning with clinical and genetic profiles to predict survival outcomes.

1.2.3 A Machine Learning Approach Using Publicly Available Datasets

The primary objective of this study is to examine breast cancer patients' survival outcomes by integrating clinical data and gene expression profiles. No primary data collection were conducted; instead, a publicly available dataset from Kaggle is utilised. The study objectives include a literature review on machine learning's role in predicting survival outcomes, analysing the interrelationship between clinical data and gene expression profiles using machine learning algorithms, and addressing data preprocessing and missing data with R and Python. Novel predictive markers will be identified, enhancing our understanding of survival mechanisms. Rigorous evaluation will ensure the developed predictive models' accuracy and reliability. This will help to contribute valuable insights to breast cancer research and precision medicine advancement.

1.2.4 Summary

Breast cancer, a complex and widespread disease, demands innovative solutions. Integrating machine learning with clinical and genomic data shows promise in diagnosis, treatment, and prognosis (Yadav et al., 2023). This study employs machine learning to predict breast cancer patients' survival by fusing clinical and genomic data. Leveraging existing literature and diverse datasets, it identifies influential factors, advancing personalised treatment. Emphasizing the potential of multi-omics analysis and machine learning, the study aims to uncover complex relationships, improving patient outcomes and contributing to precision medicine.

2 Literature Review

2.1 Leveraging Genomic Characteristics and Data Analytics

Breast cancer poses a significant global health challenge, underscoring the need for robust prognostic approaches that leverage genomic traits and medical records for tailored treatments. According to Yadav et al. (2023), breast cancer maintains its prominence as a widespread concern affecting millions worldwide. Swift and accurate detection remains pivotal for enhancing patient prognoses. Recent years have witnessed the integration of machine learning algorithms to aid breast cancer detection and diagnosis.

2.2 Conduct an Assessment in Alignment with the Recommended System

In pursuit of the established framework, a diverse range of methodologies has been employed to execute a comprehensive assessment. This assessment encompasses crucial stages including precise data preprocessing, effective dimensionality reduction, and the implementation of advanced algorithms. These endeavours have yielded commendable accuracy rates, particularly when distinguishing between distinct categories such as treatment outcomes and survival predictions Sammut SJ. et al. (2021). However, it is imperative to acknowledge the existing limitations, that is the scarcity of available datasets and the potential pitfalls associated with overfitting. Addressing these challenges requires a concerted focus on acquiring comprehensive and diverse datasets, meticulous validation protocols, and a deliberate approach to surmount the intricacies posed by deep learning techniques Deeba Khan et al. (2022). By navigating these intricacies, we can lay the groundwork for further advancements in this domain, culminating in more robust and reliable outcomes.

2.2.1 Machine Learning Advancements in Breast Cancer Detection

In recent years, machine learning (ML) algorithms has aid in breast cancer detection and diagnosis. Yadav et al. (2023) evolve in a study that employed diverse methodologies, encompassing data preprocessing and dimensionality reduction techniques to prepare the datasets. Feature selection and extraction have been utilized to identify relevant biomarkers and reduce data complexity. ML algorithms such as Support Vector Machines (SVM), Logistic Regression (LR), Decision Trees (DT), Naive Bayes (NB), K Nearest Neighbors (KNN), and neural networks (NN) were harnessed for classification tasks, while ensemble methods were explored to enhance predictive performance. Deep learning techniques, particularly Convolutional Neural Networks (CNN), have been applied for feature extraction and image-based analysis using mammography datasets.

The findings of the studies, show high accuracy rates, ranging from 90% to 99%, in distinguishing between benign and malignant tumours. SVM and CNN emerged as particularly effective classifiers across various datasets. Moreover, the integration of multi-omics data and advanced feature extraction methods further bolstered the predictive capabilities of the models. Ensemble techniques demonstrated superior performance compared to individual models in several instances. Deep learning approaches, like CNNs, exhibited high efficiency in analysing mammography images, contributing to accurate cancer detection.

However, the studies also underscored certain limitations in the application of ML for breast cancer detection. Foremost among them was the restricted availability and size of datasets, potentially leading to overfitting and limited generalizability. Acquiring labelled and annotated medical data proved challenging due to privacy concerns and ethical considerations. Nevertheless, the limited availability of large and diverse datasets remains a key challenge for further enhancing model performance. Future research should prioritize acquiring comprehensive and representative datasets, conducting rigorous external validation, and addressing the computational demands of deep learning models.

2.2.2 Application of Machine Learning Algorithms to Multi-omics Data

Breast cancer represents a significant health threat, particularly in regions like the Indian subcontinent. This pressing factor poses higher risks to human health, demanding urgent attention and effective strategies to combat the disease. Malik et al. (2021) explored the application of machine learning algorithms to multi-omics data to predict survival outcomes in breast cancer patients, using data primarily from TCGA and GDSC databases. Their approach integrated diverse omics datasets, including gene expression, somatic mutation, CNV, protein expression, and microRNA profiles. Feature selection techniques like Neighbourhood Component Analysis (NCA) were employed to identify relevant features for constructing predictive models.

The study highlighted the potency of neural network models in accurately predicting clinical outcomes and drug responses based on multi-omics data. Malik et al.'s neural network-based classifier achieved an impressive 94% accuracy in distinguishing breast cancer patients into high-risk and low-risk categories, with an excellent AUROC value of 0.98. The drug response prediction model also showed promise, outperforming some traditional methods. Integrating multi-omics data allows for comprehensive analysis of the human genome at various complexity levels, but the high dimensionality and heterogeneity of omics datasets present challenges for traditional linear prediction models. Neural networks, known for capturing complex relationships and mitigating overfitting risks, were employed to overcome these challenges.

Moreover, the study identified biologically significant gene sets as potential prognostic biomarkers for breast cancer, such as EFHD1(EF-hand domain-containing protein D1: A protein encoded by the EFHD1 gene in humans), CDH1 (E-cadherin (epithelial cadherin), is a protein encoded by the CDH1 gene in humans), PIK3CA ((Phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha) is a gene that encodes the catalytic subunit alpha of the phosphatidylinositol 3-kinase (PI3K) enzyme in humans), and TP53 (Tumour Protein 53 commonly referred to as the "guardian of the genome," is a crucial tumour suppressor gene located on chromosome 17 in humans). These findings underscore their relevance in patient stratification and personalized therapeutic decision-making.

2.2.3 Interpretable Meta-Learning of Multi-Omics Data

The article "Interpretable Meta-Learning of Multi-Omics Data for Survival Analysis and Pathway Enrichment," authored by Hyun Jae Cho et al. (2022), explores the integration of multi-omics data for cancer survival analysis using machine learning

techniques. Traditional survival analysis methods encounter challenges in predicting patient outcomes due to the complexity of cancer. To address this, the authors propose advanced machine learning algorithms capable of handling high-dimensional feature spaces and missing survival information. Their solution involves a meta-learning approach, training a model on related tasks using multi-omics datasets from The Cancer Genome Atlas (TCGA). This approach leverage on the unique characteristics of each omics dataset to enhance survival predictions. The authors also utilize DeepLIFT, a variable importance analysis method, to interpret deep neural network models and identify enriched pathways and gene functional relationships.

The authors demonstrate the effectiveness of their meta-learning approach by training on related tasks and fine-tuning on target tasks. Leveraging information from diverse datasets, including transcriptomics, proteomics, and clinical data, their model achieves superior predictive performance. DeepLIFT aids in interpreting deep neural networks, providing valuable insights into gene functional relationships and enriched pathways, further advancing researchers' knowledge of cancer's molecular mechanisms. However, the review acknowledges several limitations. The 'big P, small N' problem, where the number of features exceeds available data samples, which can lead to overfitting and reduced generalization capabilities of machine learning models. Despite these challenges, the integration of multi-omics data with advanced machine learning techniques shows promise for advancing cancer research and personalized medicine.

2.2.4 Utilizing Deep Learning Methods and Integrated Omics Data for Personalized Breast Cancer Treatment

The article titled "Leveraging Deep Learning Techniques and Integrated Omics Data for Tailored Treatment of Breast Cancer" by Deeba Khan et al (2022) delves into the critical importance of enhancing the diagnosis and treatment of breast cancer, which stands as the most prevalent cancer among women worldwide. Breast cancer's incidence is surging at a concerning rate of 14% per year, contributing significantly to the overall cancer burden globally. Personalized therapies have emerged as a promising avenue for improving cancer prognosis, capitalizing on individual patient genotypes and their unique responsiveness to specific treatment approaches. However, it is noteworthy that only a small proportion of genomic-based studies have received regulatory consent, underscoring the necessity for advanced intelligent systems to bolster the effectiveness of personalized treatment strategies.

To address the complexity feature engineering were conducted using nearest component analysis (NCA) is employed to reduce dimensionality while preserving the integrity of biological processes in the face of the 'big p, small n' problem in omics data. Deep learning techniques, such as Deep Neural Networks (DNNs), Convolutional Neural Networks (CNNs), and autoencoders, are leveraged in the proposed methodology. The book presents impressive findings from its proposed deep learning models. The multiomics-based classifier achieves high accuracy in classifying breast cancer subtypes, specifically HER2+, basal-like, and unknown subtypes. However, there are some challenges in accurately distinguishing luminal A and luminal B

subtypes due to their molecular similarities. The performance of the proposed model is competitive compared to existing methods for breast cancer subtype classification.

The study acknowledges certain limitations in the research. Misclassifications between luminal A and luminal B subtypes need further investigation to improve the distinction of these molecularly similar subtypes. The predictive power of some drugs may be affected by outliers in the data, influencing the model's confidence in drug response predictions. The study offers a promising direction in breast cancer research and personalized medicine. The proposed models show potential in accurately classifying breast cancer subtypes and predicting drug responses tailored to specific subtypes. It is important to note, addressing the identified limitations is crucial to fully realize the clinical impact of these models.

2.2.5 Analysis of ctDNA Relevance

The study by Sant, M. et al. (2022) presents an analysis of the relevance of ctDNA in breast cancer, using the PubMed database from July 2014 to July 2021, employing key search terms, including breast cancer, ctDNA, liquid biopsy, and precision oncology. The study reveals ctDNA's potential as a diagnostic tool for breast cancer. It has shown promise in situations where tumour biopsy is not feasible and may complement conventional screening methods. Elevated cfDNA levels in breast cancer patients and their correlation with tumour characteristics offer opportunities for early diagnosis and prognostication. Studies have shown that ctDNA analysis can serve as an early predictor of relapse, outperforming conventional monitoring methods. Serial monitoring of ctDNA during follow-up visits aids in personalized treatment decisions and improves patient outcomes by detecting metastatic progression sooner. Furthermore, the study highlights ctDNA's significance in evaluating treatment response and guiding therapeutic decisions. Pre- and post-neoadjuvant therapy ctDNA analysis has been associated with treatment response and pCR rates. Furthermore, clearance of ctDNA after treatment indicates better survival outcomes.

The study acknowledges certain limitations in ctDNA research for breast cancer. Low ctDNA fraction in cfDNA presents challenges in detection and quantification, especially in early-stage breast cancer. Tumour heterogeneity may cause the loss of specific mutations during targeted ctDNA analysis, and false positives may occur due to clonal haematopoiesis. In conclusion, the literature review demonstrates the potential of ctDNA analysis in breast cancer diagnosis, follow-up assessments, and MRD detection. Liquid biopsy, specifically ctDNA analysis, holds promise for early relapse detection and personalized treatment decisions. However, addressing existing limitations and further research is essential to fully integrate ctDNA analysis into routine clinical practice for breast cancer management.

2.2.6 Application of Machine Learning in Metastatic Cancer Research

According to Olutomilayo O. P et al. (2023) in their comprehensive study review, they explore the application of machine learning in metastatic cancer research, covering various aspects and highlighting significant findings. The researchers delve into the use of machine learning algorithms, including supervised and unsupervised techniques, for analysing complex datasets encompassing genomic, transcriptomic,

and imaging data. They emphasize the importance of public data repositories such as TCGA, GEO, and HCMBD, which provide valuable resources for accessing and sharing metastatic cancer data. Additionally, the study discusses the utilization of machine learning in drug discovery, virtual screening, and prediction analysis, contributing to advancements in precision medicine for cancer treatment.

Among the key findings, the study reveals promising results in the early detection and prognosis of metastatic cancer, along with the prediction of patient survival outcomes through the application of machine learning models. Deep learning techniques have proven effective in identifying and characterizing metastatic lesions from imaging data, aiding in improved diagnosis and treatment planning. Moreover, machine learning has facilitated drug discovery, toxicity prediction, drug repositioning, and the assessment of molecular bioactivity, leading to more efficient drug design and development processes.

However, the researchers acknowledge certain limitations in the application of machine learning in metastatic cancer research. One major constraint is the lack of diversity and representation in clinical trial data, which may result in disparities in treatment and outcomes for different populations. Missing data and data imbalances can also impact the performance and generalizability of machine learning models, requiring careful handling and preprocessing. Furthermore, the complexity of tumour heterogeneity presents challenges in accurately predicting treatment responses and devising effective treatment strategies. The study emphasizes the importance of promoting diversity and inclusivity in clinical trials and data collection, ensuring a comprehensive understanding of metastatic cancer across various patient populations.

2.2.7 Predicting Disease-Related lncRNAs:

Lin Yuan et al. (2021) their study proposes a novel machine learning framework called LGDLDA for predicting disease-related lncRNAs based on multi-omics data and neural network neighbourhood information aggregation. The methodology involves data collection from various databases containing lncRNA-disease, lncRNA-gene, and gene-disease associations, followed by the calculation of similarity matrices for lncRNAs, genes, and diseases using various omics data. A neural network is then utilized to aggregate neighbourhood information from these matrices, capturing complex features in the data. Low-dimensional spatial node representations are generated to capture relationships between lncRNAs, genes, and diseases in a reduced-dimensional space. Projection matrices are used to approximate observed matrices and learn as much information as possible from the original data. Candidate lncRNA-disease pairs are ranked based on the learned association matrix to identify potential disease-related lncRNAs. The LGDLDA method is applied to real cancer data, resulting in the identification of several potential cancer-related lncRNAs supported by recent literature.

However, the study has limitations, including reliance on small datasets that may affect the method's performance and generalizability. Additionally, the generalization of LGDLDA to other diseases remains unclear, and further validation on larger

datasets is necessary. The lack of detailed biological validation experiments for the predicted lncRNAs is another limitation. LGDLDA shows promise as a machine learning framework for predicting disease-related lncRNAs, outperforming existing methods in predicting lncRNA-disease associations on a small simulation network. The identification of potential cancer-related lncRNAs is valuable, but further validation on larger datasets and biological experiments are needed to assess its practical applicability in cancer diagnosis and prognosis. Despite the limitations, LGDLDA represents a significant step towards understanding the complex relationships between lncRNAs and diseases, providing valuable insights into potential biomarkers and therapeutic targets.

2.2.8 Predicting Neoadjuvant Therapy Response in Breast Cancer

Sammut SJ. et al (2021) conducted a study involving 180 women with early and locally advanced breast cancer undergoing neoadjuvant treatment. They collected pre-treatment core tumour biopsies and employed various techniques, such as shallow whole-genome sequencing, whole-exome sequencing, and RNA sequencing, to profile the samples. Their analysis revealed that individual clinical features, including tumour grade, ER status, and lymph node involvement, were associated with treatment response, but their predictive performance was limited. On the other hand, tumour mutation burden, HRD and APOBEC mutational signatures, and chromosomal instability were found to be correlated with response, particularly in HER2- tumours. The study also identified immune signatures, such as proliferation and immune activation, as strong indicators of treatment response, with tumours exhibiting an active tumour immune microenvironment showing better response.

Conversely, tumours displaying T cell dysfunction and exclusion tended to be resistant to therapy, suggesting a role of immune escape in treatment resistance. Additionally, loss of heterozygosity over the HLA class I locus was linked to therapy resistance. Although the study provided valuable insights, it acknowledged certain limitations. The sample size of 168 patients might be considered relatively small, potentially affecting the generalizability of the findings. While the researchers integrated diverse data sources, the addition of other omics data, such as proteomics or metabolomics, could enhance the predictive models further. Although the models demonstrated good performance in an external validation cohort, larger validation studies with diverse patient populations are necessary to confirm their robustness.

The findings underscore the significance of considering the tumour ecosystem, encompassing genomic landscapes and the tumour immune microenvironment, in predicting treatment response. The study identified potential biomarkers, such as tumour mutation burden, HRD signatures, and immune activation, which could aid in treatment decision-making and personalized therapy. The predictive models hold promise for clinical application, guiding treatment decisions, and identifying patients who might benefit from novel therapies in clinical trials. To enhance the predictive models, future research should explore additional data types, validate the models in larger cohorts, and investigate the interplay between different features within the tumour ecosystem. The challenge faced is size of the data in term of generalization, the machine learning model faces difficulties in attaining a high level of accuracy.

Presently, it depends on having a sufficient amount of data to effectively unravel the connections among variables.

2.2.9 Machine Learning Approach for Predicting Engineered System Behaviour

According to Zak Costello et al. (2018) In their study the proposed machine learning approach is grounded in supervised learning methods, utilizing time-series data of proteomics and metabolomics to predict the behaviour of engineered biological systems. The training set is thoughtfully constructed, comprising sets of proteomics and metabolomics data along with their corresponding derivatives. To ensure precise derivative estimation, a Savitzky–Golay filter is applied to the noisy time-series data, followed by central difference scheme computation.

Model selection is performed using a tree-based pipeline optimization tool (TPOT), effectively combining various machine-learning regressors and preprocessing algorithms. The best-performing model is selected based on tenfold cross-validated performance on the training set. The machine learning approach demonstrates significant promise in qualitatively predicting pathway dynamics. Remarkably, with just two time-series strains as training data, the approach outperforms classical Michaelis–Menten kinetic models in predictive power. The predictions offered by the machine learning model are sufficiently accurate to drive design decisions, rank production levels, and enhance the understanding of pathway dynamics for metabolic engineers and synthetic biologists.

Despite its success, the machine learning approach faces certain limitations that necessitate careful consideration. Chiefly, the available data sets for training the model are often limited in size, posing challenges in achieving high accuracy. The machine learning model currently relies on enough data to accurately disentangle relationships between variables. Additionally, the model's performance may be constrained when applied to pathways involving changes in catalytic rates (kcat), as its primary reliance lies on protein abundance changes.

2.2.10 Unlocking Cancer Insights through Multi-Omics Approaches

Yong Jin Heo et al. (2012). In their study provide an extensive overview of the significance of multi-omics approaches in comprehending cancer biology and pathophysiological features during oncogenesis and tumour progression. The researchers emphasize the need for a comprehensive perspective by exploring genomic and epigenetic aberrations and introduce multi-omics studies as data-driven investigations that analyse high-dimensional datasets. The study highlights the recent advancements in high-throughput technologies and proteomics, which have led to a transformative shift in cancer research. The methodology of the study outlines the computational frameworks employed for multi-omics investigations. It involves the integration of various omics datasets, including genomics, transcriptomics, epigenomics, and proteomics, to holistically characterize the molecular and clinical structures of cancer patients. Moreover, the study delves into the development of data-driven mathematical and computational methods to effectively analyse the complex datasets obtained from multiple analysis platforms. Additionally, the

researchers mention specific tools, such as iCluster and iOmicsPASS, which have been effectively applied in cancer research.

The findings present the latest discoveries obtained through multi-omics approaches in cancer research. It covers the identification of coherent subtypes, the association between molecular profiles and clinical features, and the potential to reveal new subtypes of breast cancer. The section discusses the use of proteogenomic approaches and large-scale proteomic research to uncover new biological mechanisms in cancers and provide fundamental information for integration strategies and computational algorithms. The limitations section addresses the challenges associated with multi-omics approaches. It mentions the need for high-quality and unbiased datasets, proper integration methods, and computational algorithms for robust and systematic assessments. The section also emphasizes the computational and biological challenges in acquiring biological insights from multi-omics data, which requires researchers to select appropriate multi-omics tools.

2.3.1 Theoretical Framework

The reviews are guided by an integrated theoretical framework encompassing machine learning principles, data-driven analysis, clinical research considerations, ethical and regulatory aspects, translational research concepts, and the complexities of biology. This comprehensive framework serves as a lens to evaluate the literature. Machine learning and predictive modelling principles underpin the assessment of algorithms' accuracy and generalizability. Data-driven analysis and validation methodologies are employed to gauge the robustness of the models. Clinical trial and real-world application perspectives gauge the clinical utility of the models. Ethical and regulatory considerations are explored to address potential ethical challenges. Translational research principles highlight the transformation of computational accuracy into meaningful clinical impact. Lastly, the complex and uncertain nature of cancer biology informs discussions about model limitations. By integrating these frameworks, the reviews systematically assess the literature, identify gaps, and unearth challenges and opportunities at the intersection of machine learning, clinical research, and biology.

2.4.1 Summary and Recurring Limitation

Several of the examined papers highlight the challenge of limited dataset availability and size, a common hurdle in both machine learning and multi-omics research within the realm of cancer detection and prediction. The constraints posed by small datasets can result in overfitting issues and hamper the ability of models to generalize effectively. Researchers are urged to prioritize endeavors aimed at procuring comprehensive and representative datasets, a critical step toward enhancing model performance. This issue underscores the imperative for collaborative data-sharing initiatives and the expansion of data collection efforts on a larger scale within the field.

Within the literature, there is conspicuous attention paid to privacy concerns and ethical considerations associated with medical data. The acquisition of labeled and annotated medical data for research purposes proves challenging, primarily due to

stringent privacy regulations and the paramount importance of safeguarding patient information. Striking the delicate balance between the imperative of data accessibility and the preservation of patient privacy emerges as a fundamental ethical quandary in the domain of cancer research. Consequently, there arises a pressing need for the development of robust frameworks for data anonymization and secure data sharing. Another recurrent challenge that surfaces in select papers revolves around tumour heterogeneity. This phenomenon can significantly impact the accuracy of predictive models, particularly within the context of personalized medicine approaches. Tumours frequently exhibit intricate molecular profiles, rendering precise predictions of treatment responses a formidable task Sant, M. et al. (2022). To address this issue comprehensively, researchers must harness more extensive and detailed omics data that can capture the multifaceted nature of tumor heterogeneity Yadav et al. (2023).

The contrast in methodologies across the literature review is noteworthy. On one hand, several papers delve into the application of machine learning algorithms for cancer research. These approaches primarily lean on algorithms and statistical models to analyze data, offering the potential to make accurate predictions. On the other hand, some papers pivot toward the integration of multi-omics data. This approach involves the amalgamation of data from diverse omics sources, encompassing genomics, transcriptomics, and proteomics, in order to acquire a holistic comprehension of cancer biology. The selection between these approaches hinges on the specific research objectives and data accessibility, with each methodology offering distinct advantages.

Furthermore, the variance in emphasis across the reviewed papers is apparent, particularly concerning clinical application and biological insight. Yong Jin Heo et al. (2012) papers place a strong emphasis on the clinical utility of their models, striving to enhance patient outcomes through precise diagnoses and predictions of treatment responses. In contrast, other papers are geared towards furnishing a deeper understanding of cancer biology. They aim to unveil biological mechanisms and identify potential biomarkers and therapeutic targets. This distinction highlights the multifaceted nature of cancer research, where some endeavors are geared towards immediate clinical impact, while others contribute to the broader body of knowledge surrounding the disease.

2.4.2 Table 1 Review Summary

Table 1.1

S/N	Literature	Review Element	Methodology
1	Yadav et al. (2023)#	Diagnosis of Breast Cancer using Machine Learning Techniques -A Survey (detection and prediction of breast cancer,)	Deep learning techniques
2	Malik et al. (2021)#	Deep learning assisted multi-omics integration for survival and drug-response prediction in breast cancer	TCGA and GDSC to generate robust survival and drug response prediction models
3	Hyun Jae Cho et al. (2022)#	Interpretable meta-learning of multi-omics data for survival analysis and pathway enrichment.	DeepLIFT (a variable importance analysis method), Deep Neural Network.
4	Deeba Khan et al. (2022)#	Utilizing Deep Learning Techniques and Integrated Omics Data for Tailored Treatment.	Convolutional Neural Networks (CNNs), and autoencoder
5	Sant, M. et al (2022)#	ctDNA's potential as a diagnostic tool for breast cancer	ctDNA tumor biopsy detection
6	Olutomilayo O. P et al. (2023)#	Application of machine learning in metastatic cancer research.	Early detection and prognosis of metastatic cancer
7	Lin Yuan et al (2021)#	Identification of several potential cancer-related lncRNAs	LGDLDA for predicting disease-related lncRNAs
8	Sammut SJ. et al (2021) #	Predicting Neoadjuvant Therapy Response in Breast Cancer	Ensemble machine learning approach
9	Zak Costello and Hector Garcia et al (2018)#	Predict metabolic pathway dynamics from time-series multiomics data	Employing supervised learning techniques, using time-series data from proteomics and metabolomics to forecast the performance of engineered biological systems
10	Yong Jin Heo et al (2021)#	The association between molecular profiles and clinical features	Computational frameworks employed for multi-omics investigations

Methodology

3.1 Research Design

This study adopts a mixed-methods research design, combining both quantitative and qualitative data analysis methods. The use of a mixed-methods design allows for a comprehensive understanding of the research problem and enables triangulation of data, enhancing the validity and reliability of the findings Hyun Jae Cho et al. (2022).

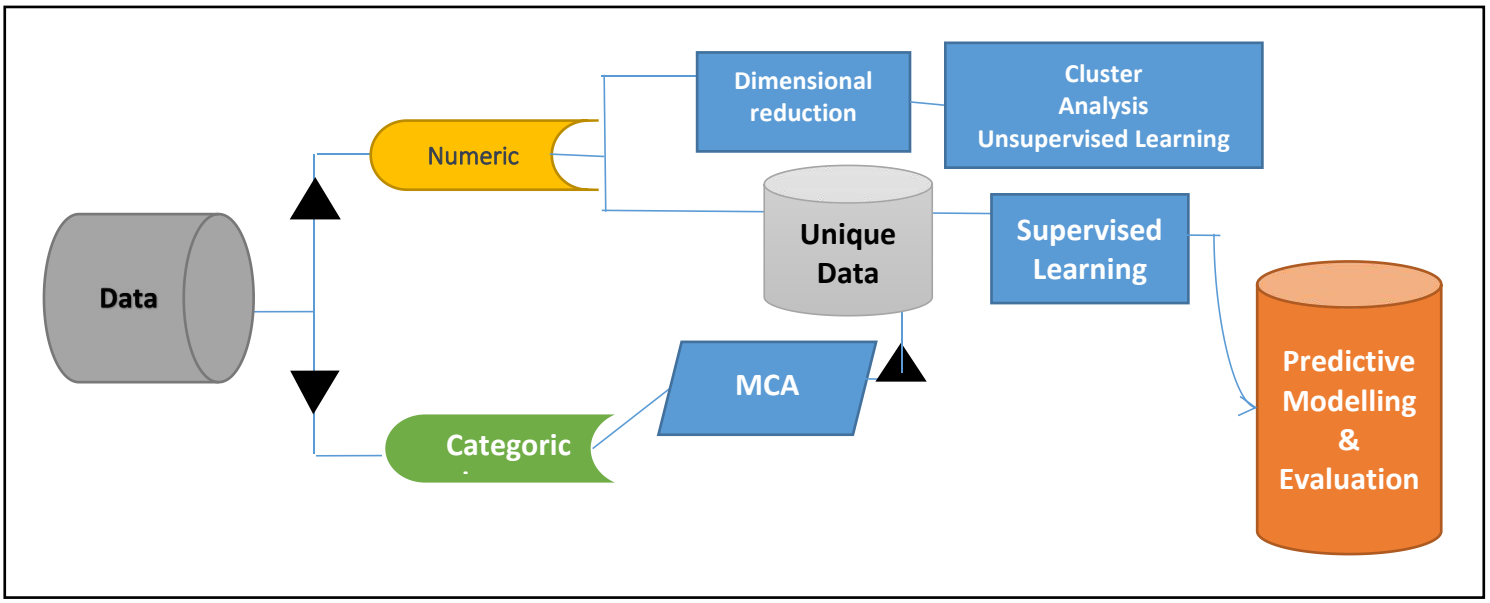


Figure 3.1 Research approach

Data Collection

The data used in this study is from the METABRIC database, which is an open-source database. It consists of targeted sequencing data obtained from 1,980 primary breast cancer samples. To complement the genomic information, clinical data pertaining to these samples was also downloaded from cBioPortal. The integration of these datasets from METABRIC can be accessed from https://www.kaggle.com/datasets/raghadalharbi/breast-cancer-gene-expression-profiles-metabric

The dataset comprises a blend of clinical attributes, m-RNA levels z-score, and gene mutations data for a total of 1904 patients. Within the dataset. Among these, 190 instances are categorized as being of a categorical nature, while 503 instances are characterized as numerical. The dataset encompasses both independent/categorical and numerical columns. Additionally, a dependent/predictor variable is present. The survival status information exists in two dimensions: "overall_survival" represented as a binary value, and "overall_survival_months" represented as discrete values. All variables/parameters, inclusive of their corresponding descriptions, values, and the respective proportions of each value, are comprehensively outlined at List of Parameters in the appendix.

3.2 Implementation

3.2.1 Data Preparation and Feature Engineering

The initial phase of the methodology revolves around data acquisition and preparation, the breast cancer dataset, derived from "METABRIC_RNA_Mutation.csv," is loaded into the analytical environment. Relevant libraries and packages in the R programming language are imported to facilitate data manipulation, visualization, and analysis. The predictive analysis involves data preprocessing, feature engineering, numerical and categorical feature selection, and merging datasets. The goal is to generate a merged dataset that captures crucial variables for constructing predictive models Malik et al. (2021).

3.2.2 *Exploratory Data Analysis (EDA)*

The aim of Exploratory data analysis in this study is to provide insights into the dataset's structure and characteristics Malik et al. (2021). These steps visualized Numerical variable distributions using histograms, to understand the data's range and distribution. Summary statistics are computed to gain an overview of the dataset's central tendencies and dispersions. Detection of missing data points was employed to help assess data completeness. Correlations is computed using Plotly (Python package via Anaconda environment) Express offers interactive box plots for detailed exploration, while Matplotlib and Seaborn to provide static bar plots with enhanced customization options. The top 30 attributes with the highest absolute correlations were selected, creating a subset DataFrame, and calculating a correlation matrix.

3.2.3 *Numeric Analysis*

Numeric data were extracted from the dataset using R, employing an iterative approach through the dataset. Subsequently, the identification of missing values within the numeric variables led to the utilization of mean values for imputation, thereby mitigating potential data loss. To address issues of multicollinearity and enhance model performance, a process of numerical feature selection was undertaken. The evaluation of multicollinearity was accomplished through a Variance Inflation Factor (VIF) analysis. Variables with elevated VIF values were selectively removed to alleviate the adverse impacts of multicollinearity. Furthermore, the exploration of relationships was facilitated through the application of correlation algorithms. Also, outliers present within the numerical features were pinpointed and subsequently treated using Z-scores.

3.2.4 *Dimensionality Reduction*

PCA is applied to the numerical data to reduce dimensionality while preserving key patterns Malik, V et al. (2021). Principal components, standard deviations, and variance proportions are computed to quantify dimensionality reduction Humaira, H (2020). Loading of original variables on principal components are examined to interpret their influence. Variable selection based on PCA loading is conducted to retain critical features.

3.2.5 *Categorical Data Analysis*

Categorical data is analysed to uncover patterns and distributions. Bar plots visualize the distribution of categorical variables, facilitating insights into their prevalence. Assessment of missing data in categorical features is performed to ensure data completeness. Categorical variables are converted into factors, enabling subsequent categorical analyses also transformation of select ordinal variables.

Multiple Correspondence Analysis is employed (Using R package. Available at appendix), MCA reveals relationships between categorical variables, enhancing the understanding of underlying patterns Ferraro, F. R. (2013). MCA offer Eigenvalues and variance proportions the selection of important categorical variables is conducted based on their associations with dimensions from MCA. A subset of significant categorical variables is chosen for further analysis, focusing on informative features. Also feature engineering involved the transformation of categorical data into standardized factor variables using one-hot encoding techniques was conducted. This preprocessing step was performed to prepare nominal categorical variables for future analysis, converting them into factors for improved handling.

3.2.6 Unsupervised Learning (Cluster Analysis)

In adherence to the stipulated framework, the assessment process was executed employing a multifaceted approach. Initially, Principal Component Analysis (PCA) Utilize PCA to reduce the dimensionality of the numerical data Humaira, H et al. (2020). A Shiny app to facilitate the exploration of principal component contributions. Allow users to view top contributing variables to principal components interactively.

To determine the Optimal Clusters this study employs the elbow method to identify the ideal number of clusters for K-Means Humaira, H. et al. (2020). Plot the within-cluster sum of squares (WCSS) against different cluster counts. Identify the point of inflexion as the optimal number of clusters. K-Means Clustering Iterate through possible cluster counts to find the most suitable one, while evaluation of the cluster is done using silhouette scores to quantify clustering quality for each configuration. Lastly each cluster mean values for the principal components are profile. This discuss the characteristics of the clusters, emphasizing their differences. Analyse the relevance of higher-order principal components in cluster separation.

Hierarchical Clustering and Visualization were conducted, involving the computation of the distance matrix among data points for subsequent hierarchical clustering. The chosen method for hierarchical clustering was Ward's method, known for its efficacy in forming coherent clusters Ganggayah, M.D. et al. (2019). To visualize the hierarchy of these clusters, the resulting dendrogram was generated and analysed. The clusters, as derived from the dendrogram partitioning, were employed as the basis for colour assignments in this scatter plot.

3.2.7 Regression Modelling

Random Forest

Random Forest is a powerful ensemble learning algorithm used in machine learning for both classification and regression tasks Sammut, S.J. et al. (2022). It's an extension

of decision tree algorithms that aims to improve their performance, robustness, and generalization ability Ganggayah, M.D et al. (2019). Having obtain the cleaned dataset "original_datasetCatNum.csv" containing multi-omic data. The data was read into the variable `survival_data`, The target Column Selection define the target column name as "overall_survival_months." Feature and Target Isolate features (`X`) and the target (`y`) from the dataset.Transform categorical variables into integer-encoded factors to prepare them for modelling. (Kindly refer to the code snip at appendix)

Hyperparameter Tuning and Model Selection

Cross-Validation: Set the number of folds for cross-validation (`num_folds = 5`). And the Hyperparameter define a grid of hyperparameter combinations to explore:

- Number of trees (`ntree_values`= (50, 100, 150))
- Maximum tree depth (`maxdepth_values`= (10, 20, 30))
- Minimum samples per leaf node (`nodesize_values`= (1, 5, 10)).

Then grid search was performed follow by finding hyperparameters with lowest RMSE, then predictions on the test data using the final model.

3.2.8 Classification Modelling

Random Forest Classifier

Using Random Forest Classifier, the dataset from the specified file path using pandas. Split the dataset into features (X) and target (y). The code snips perform One-hot encoding on categorical variables to convert them into a suitable format, prepare the data for training and evaluation. Model Initialization and Cross-Validation Initialize a Random Forest Classifier with a specified random seed. Then perform k-fold cross-validation using KFold to calculate cross-validation accuracy. Make predictions using cross-validation and calculate overall accuracy.

Artificial Neural Network (ANN) Classifier

ANN is a type of machine learning model inspired by the structure and functioning of biological neural networks in the human brain. It is used for classification tasks, where the goal is to categorize input data points into different classes Malik, V. et al. (2021). These research load data and preparation using pandas. Dataset is split into features (X) and target (y). Feature Engineering and Normalization is perform using one-hot encoding on categorical variables to normalize the features. A Sequential model using tf.keras.models.Sequential() define the architecture with input, hidden, and output layers. The framework include dropout layers to mitigate over-fitting. Model Compilation and Training Compile the ANN model with specified optimizer and loss function. Then train the model on the training data with a specified number of epochs and batch size. Lastly, display the confusion matrix and classification report by calculating the ROC curve, AUC score and plotting the ROC curve.

3.2.9 Model Evaluation and Selection

By adhering to this structured approach, a comprehensive exploration of hyperparameter space is carried out, culminating in the identification of an optimal configuration Cho, H.J. et. al. (2023). The undertaken process involves systematically exploring various hyperparameter combinations through nested loops. Within this

configuration, each combination is sequentially evaluated using k-fold cross-validation Root Mean Squared Error (RMSE) values is computed and logged for every fold during cross-validation. The hyperparameter configuration resulting in the lowest calculated RMSE is pinpointed as the most optimal

Model Evaluation Metrics: The model's predictions are evaluated using a predefined set of metrics:

- Root Mean Squared Error (RMSE) (Evaluate the accuracy of a predictive model)
- Mean Absolute Error (MAE) (Measuring the error between the predicted values and the actual (observed) values.)
- Mean Squared Error (MSE) (It quantifies the average of the squared differences between the predicted values and the actual (observed) values)
- R-squared (R2) score. (It measures the proportion of the variance in the dependent variable that is explained by the independent variables in the model)
- Adjusted R-squared score. (Measures the goodness of fit by quantifying the proportion of the variance in the dependent variable that is explained by the independent variables)

3.2.10 Identifying breast cancer survival and treatment strategies objective.

Plotly Express:

This approach employs Plotly Express (Python package), a library for interactive visualization. It focuses on creating interactive box plots to showcase the distribution of survival months across various treatment factors. The code iterates through specified treatment columns (Chemotherapy, Hormone therapy, Radio therapy), crafting individual box plots with treatment on the x-axis, survival months on the y-axis, and color-coded by overall survival. The layout is customized with labels and titles, offering an interactive exploration of the data.

Matplotlib & Seaborn:

This technique utilizes Matplotlib and Seaborn (Python package), emphasizing bar plots to visualize survival percentages across different treatments. The script computes survival percentages by grouping data according to treatment factors and calculating mean percentages. The bar plot is generated with treatments on the x-axis, survival percentages on the y-axis, and hue-coded by treatment type. The next step visualizes correlations between all the attributes and "overall_survival" using Seaborn and Matplotlib.

3.2.11 Gradient Boosting and XGBoost Modelling (Classifier)

Gradient Boosting and XGBoost focuses on developing accurate predictive models for classifying treatment. Ganggayah, M.D. et al. (2019). Categorical attributes like age groups and surgical approaches are identified and numerically encoded(Using Python package). Features (patient attributes) and a target variable (surgery type) are defined, as it was do for other treatment, followed by data splitting for training and evaluation. Two ensemble techniques, Gradient Boosting and XGBoost, are employed to build robust models that iteratively combine weaker models for accurate surgery type classification. Feature importance are calculated, revealing attributes with the greatest impact on predictions, aiding medical insights. Following model training, the

names of influential features are extracted and organized, simplifying analysis. A concise presentation of significant features is provided, highlighting patient attributes essential for surgery type prediction. This methodology's systematic approach ensures effective predictive modelling by comprehensively covering preprocessing, model training, feature importance, and potential evaluation.

3.3 Research Ethics

The study adhere to ethical principles and obtain informed consent from the University. As the data is sourced from the Kaggle repository, there is no need for ethical approval from the Institutional Review Board (BREO) since participants' rights and well-being are not at risk.

3.4 Limitations

The research acknowledges several limitations, including potential response biases due to self-reported data and limitations in making causal inferences with the cross-sectional design. Generalizability of the findings may also be restricted to the specific target population. However, efforts will be made to enhance the validity and reliability of the study through triangulation and careful data analysis.

3.5 Summary

In the practice of investigating, both Multiple Correspondence Analysis (MCA) and Principal Component Analysis (PCA) have successfully identified several crucial features. However, a compelling rationale for excluding any variables has not been established. Consequently, all 488 variables deemed clean have been retained for both training and testing across 1902 observations. The dataset underscores the significant relevance of all 488 variables in predicting survival, as evidenced by the model's performance (Refer to the findings chapter).

In this chapter, the methodology used to conduct the study is outlined. This study employs a mixed-methods approach to predict breast cancer survival using clinical and gene expression data integration. The research aims to evaluate machine learning models' accuracy and performance in prognosticating breast cancer survival. By merging datasets from the METABRIC database, the study explores the potential of multi-omics data to enhance predictive models.

The methodology involves data preprocessing, dimensionality reduction using PCA, cluster analysis, and regression and classification modeling. Through Random Forest, Artificial Neural Network models and Gradient Boosting and XGBoost model Classifier, the study evaluates accuracy, confusion matrices, classification reports, ROC curves, and AUC scores. Ethical considerations ensure participant rights. Despite potential limitations, this research contributes to improved prognostic tools and personalized breast cancer treatment strategies, aligning with precision medicine objectives.

4.1 Findings

Analytical Approach for Investigating Results

- Exploratory Data Analysis (EDA):
- Clustered Analysis:
- Graphical Illustration:
- Table Illustration:
- Biomarker Analysis:
- Model Analysis:
- Clinical Relevance:
- Future Directions:

Exploratory Data Analysis Findings (EDA):

During the Exploratory Data Analysis phase, several steps were taken to prepare the dataset for further analysis. These included effective preprocessing of both numerical and categorical data, utilization of visualizations like box plots and scatter plots to uncover data patterns. Addressing missing data there are 638 instances of missing data these data are replace with the mean, duplicates are removed to ensure data integrity, pruning non-informative features like the patient_id. Verifying variable relationships using correlation as shown in Figure 4.1 & Table 4.1 identifying especially age at diagnosis, lymph_nodes_examined_positive, and tumor_size exhibit a negative correlation with breast cancer patients' overall survival. It mean as the variable increases, (As a patient's age increases before diagnosis, and increase in tumor size.) overall survival in breast cancer patients tends to decline. While, variables such as "radiotherapy," "PMS2," and "CCND2" positively correlate with overall survival, to indicate the association with a patient's survival. The aim of correlation analysis in this study is to measure and understand the statistical relationship between survival and other variable as identify in Table 4.1 .

Multicollinearity is the statistical examination of intercorrelations among independent variables, which has the potential to undermine the model's precision. In order to detect multicollinearity, the Variance Inflation Factor (VIF), is utilised focusing on cases where VIF values surpassed the threshold of 10. These elevated VIF values could potentially impair the predictive prowess of the model when multicollinearity is present CFI Team (2020). Variable with VIF greater than 10 are removed reducing the dimension of the data from [1] 1904 502 to [1] 1904 488 (The number of variable in the dataset reduce from 502 to 488). Z-score-based outlier detection aims to identify extreme values within multi-omic data. Given the data's normal distribution (refer to the datapreparation code at appendix), setting the threshold at 20 is an attempt to identify only the most severe outliers, avoiding excessive labeling of data points as such. Subsequently, outliers were removed, leading to a change in data observations from [1] 1904 488 to [1] 1902 488 (signifies a change in the data value from 1904 to 1902) resulting in improved data quality.

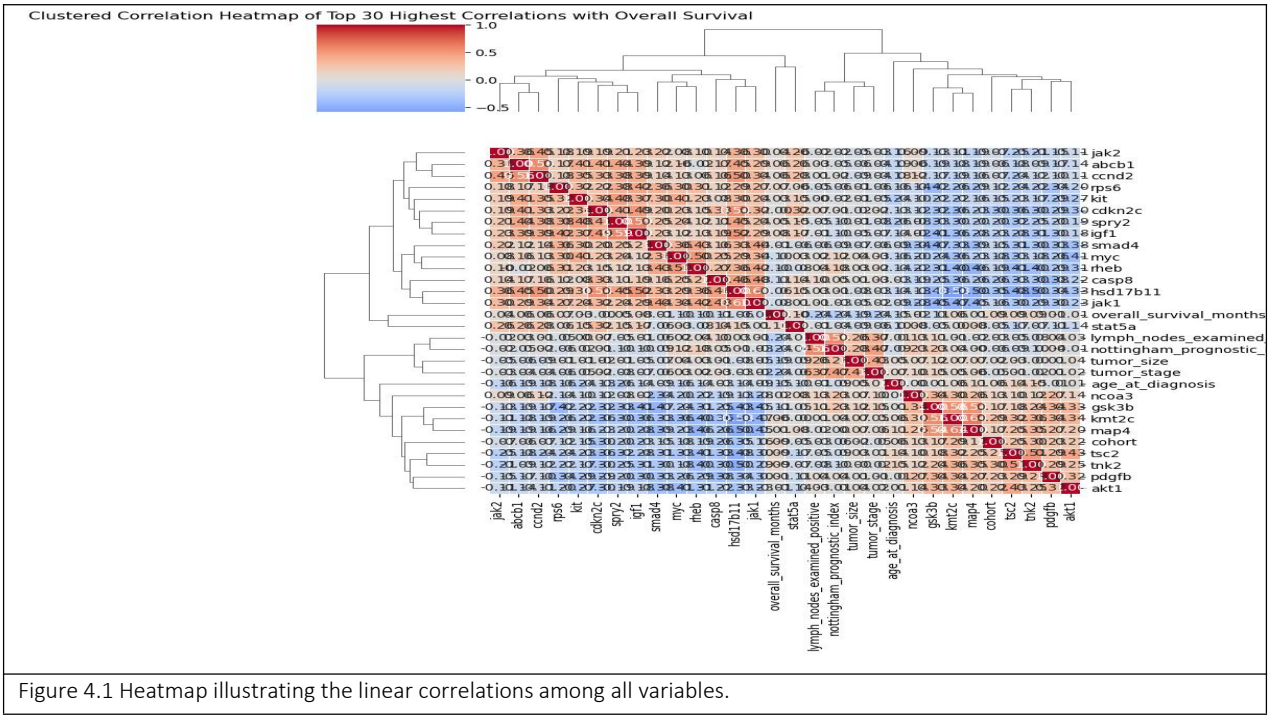


Figure 4.1 Heatmap illustrating the linear correlations among all variables.

Table 4.1 A table displaying and explaining the correlations within the dataset.

Correlation sign	Variable	Descriptions	Interpretation
Negative Correlation Variable	age_at_diagnosis, cohort, lymph_nodes_examined_positive, nottingham_prognostic_index, tumor_size, tumor_stage, nf1, bard1, e2f7, dtx3, jag2, maml1, notch3, gsk3b, map2k2, mmp11, mmp15, mlst8, smad3, tsc2, vegfa, wwox, kmt2c, map4, tubb4b, afdn, bap1, fancd2, rps6kb2, rptor, slc19a1, nr3c1, npnt, rbpjl, tubb4b, vegfa, wwox	Mix of clinical and genetic factors that could potentially be associated with a medical condition, such as cancer. These variables may be used in medical research or clinical practice to assess and understand various aspects of the condition and its prognosis.	Negative correlations are inversely related to overall survival. In the context of breast cancer survival prediction, this means that as these variables increase, the chances of a negative outcome (lower survival) tend to increase. For example, variables like "age_at_diagnosis," "lymph_nodes_examined_positive," and "tumor_size" negatively correlate with overall survival. This implies that older age at diagnosis, more positive lymph nodes, and larger tumor sizes are associated with poorer survival outcomes in breast cancer patients.
Positive Correlation Variable	overall_survival_months, radio_therapy, pms2, rb1, ccnd2, myc, jak1, jak2, stat5a, adam10, ccnd1, casp8, fas, fgf1, folr2, bmp6, mapk14, mmp10, mmp19, mmp7, psen1, akt1, akt1s1, bcl2l1, rheb, rps6, smad4, arid1a, cbfb, abcb1, setd1a, sf3b1, sik1, smarcc2, smarcd1, syne1, ubr5, akr1c3, ar, cdkn2c, hsd17b11, ran, sdc4, spry2, tnk2	This represent a comprehensive set of factors that can be considered in medical research and clinical practice to better understand the genetic and molecular aspects of various diseases, including cancer, and their impact on patient outcomes and treatment responses.	Variables with positive correlations are directly related to overall survival. In the context of breast cancer survival prediction, this means that as these variables increase, the chances of a positive outcome (higher survival) tend to increase. For example, variables like "radio_therapy," "pms2," and "ccnd2" positively correlate with overall survival. This suggests that receiving radiotherapy as part of treatment, higher levels of pms2, and the presence of ccnd2 may be associated with improved survival rates in breast cancer patients.

Cluster Analysis Findings:

Principal Component Analysis (PCA) was utilized to reduce dimensionality, with a particular focus on the most influential variables represented by the first and second principal components (PC1 and PC2). The relationship between clusters and principal components often involves the use of principal components (PC) to reduce data dimensionality, followed by an examination of data point grouping or clustering. The refined dataset was seamlessly integrated into the Shinny framework, enabling

advanced analytical techniques to identify variables associated with multiple principal components (PCs).

However, Principal Component Analysis was executed twice—first on the METABRIC dataset and subsequently on the reported clean data within the original_datasetCatNum. 2 (two) principle cluster were identified using to Elbow method and Silhouette score. During the cluster analysis for the multi-omic data, four principal components (PC1 to PC4) were examined as shown in Table 4.2. PC1 and PC2 showed strong separation between two clusters, indicating distinct breast cancer subgroups. Variables in PC1 (hsd17b11, kmt2c, setd1a, rheb, tnk2) and PC2 (ccne1, e2f3, aph1b, chek1, cdc25a) played pivotal roles in distinguishing these subgroups, likely representing unique molecular profiles with implications for survival prediction. However, as shown in Figure 4.2A & Figure 4.2B respectively PC2 vs. PC3 and PC2 vs. PC4, overlaps were observed between clusters, suggesting shared characteristics or transitional subgroups. Variables in PC3 (adgra2, acvrl1, dab2, foxo1, rps6ka2) and PC4 (aph1a, pik3ca, smad5, cir1, ctcf) may contribute to this overlap, indicating the complexity of breast cancer heterogeneity.

In Principal Component Analysis (PCA), certain variables had the highest variance in PC1 and PC2, and these specific variables were notably important for predicting survival, as shown in Figure 4.2A. The variable includes, hsd17b11, kmt2c, setd1a, rheb, tnk2, ccne1, e2f3, aph1b, chek1, and cdc25a. This observation suggests the algorithm effectively grouped data points into distinct clusters, driven by their similarities.

The overlaps in PC2 vs. PC3 and PC2 vs. PC4 highlight the need to consider multiple dimensions when assessing subgroups. Variables in PC2 appear particularly relevant for different between clusters as shown in Figure 4.2A & 4.2B. Predictive models for breast cancer survival should account for this complexity of multiple gene expression and mutation variable, potentially incorporating interactions between variables from different dimensions to improve accuracy and capture nuances in subtypes. In summary, multi-omic data analysis reveals distinct breast cancer subgroups in PC1 and PC2, with shared characteristics differentiating in PC2 vs. PC3 and PC2 vs. PC4, necessitating comprehensive modeling approaches for survival prediction.

Cluster 1 represents a group with more extreme values in certain features, while Cluster 2 represents a group with more moderate values as shown in Figure 4.2C. the clusters are well-separated and distinct from each other, it indicates that the clustering algorithm has successfully identified different groups in the data. note larger convex hull indicates that the cluster covers a broader region.

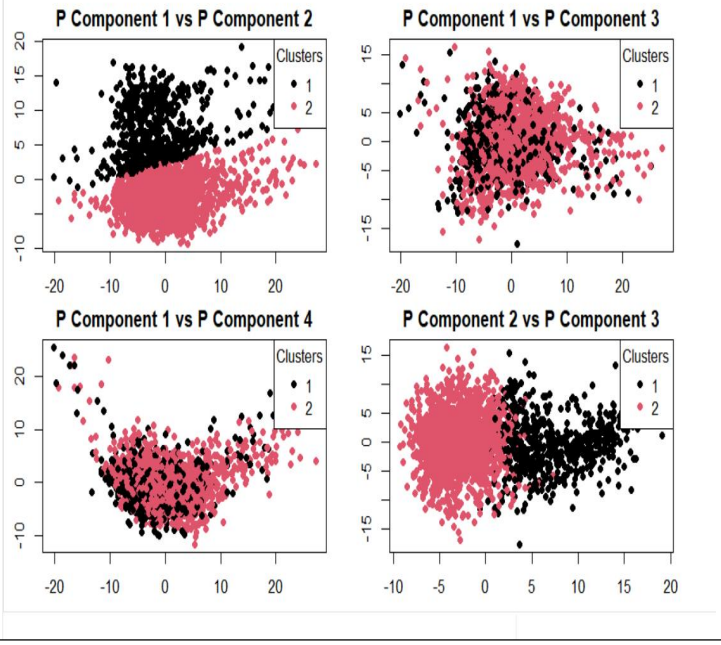


Figure 4.2A Displaying PCA Cluster plot

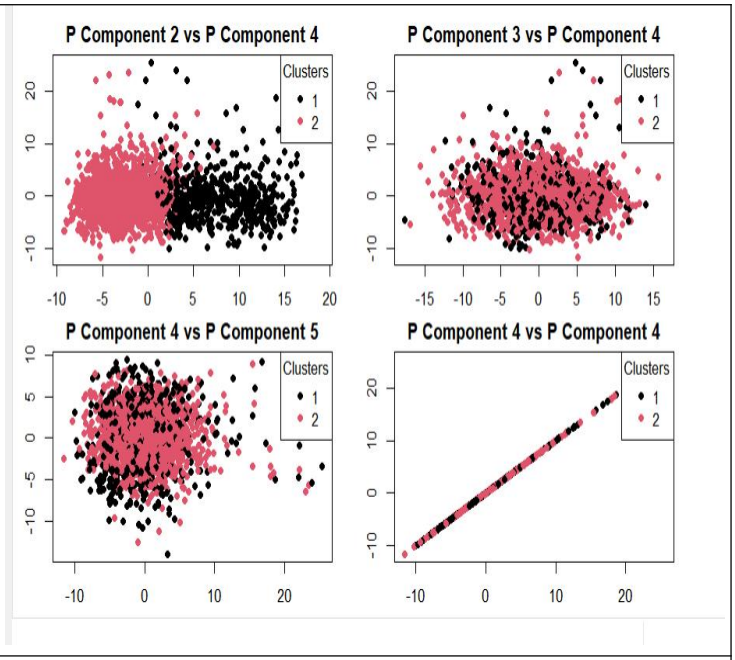


Figure 4.2B Displaying PCA Cluster plot

*Note P Component 1 is Y axis and P Component 2 is X axis, P Component 1 is Y axis and P Component 3 is X axis, P Component 2 is Y axis and P Component 3 is X axis, respectively.

Figure 4.2C

Table 4.2 Displaying Principal Components based on their grouping.

PC	Variable	Class
PC1	hsd17b11, kmt2c, setd1a, rheb, tnk2	genes/proteins
PC2	ccne1, e2f3, aph1b, chek1, cdc25a	Cell cycle regulation, DNA replication, and signal transduction.
PC3	adgra2, acvrl1, dab2, foxo1, rps6ka2	Cell signaling, gene expression regulation, and cellular responses to various signals and stimuli.
PC4	aph1a, pik3ca, smad5, cir1, ctcf	Molecular components and proteins involved in important cellular processes, including signal transduction, gene expression regulation, and genome organization.

Gradient Boosting Classifier and XGBoost Classifier:

To discover a novel predictive biomarker for objective assessment of a biological or medical condition, the code underwent data preprocessing, trained distinct machine learning classifiers, conducted feature importance analysis, and assessed model performance using a testing dataset. The focus is on identifying important biomakers and assessing how well the classifiers perform in predicting breast survival rate (In this context, the survival rate is represented as follows: a value of 1 corresponds to patients who have survived for 10 or more years after treatment, with 10 years being the average post-treatment survival time, while a value of 0 indicates patients who have survived for less than 10 years after treatment.).

Biomakers

To identify the variables that influences the outcome (survival) after breast cancer treatment, Gradient Boosting Classifier and XGBoost Classifier is used. The variables identified are mostly gene expression changes and mutations as shown in **Table 4.4**.

In the context of predicting breast cancer treatment outcomes across four distinct treatment categories, both the Gradient Classifier and XGBoost Classifier reveal a multitude of significant biomarkers in four treatment types, (as indicated in **Table 4.4**.) Among these biomarkers is Cadherin 1 (cdh1), specifically in terms of its gene expression, Checkpoint Kinase 2 (gene expression), BRCA1-Associated RING Domain 1, MutS Homolog 6 (msh6), PIK3CA Mutation (pik3ca_mut), GATA3 Mutation (gata3_mut), BRCA1 Mutation (brca1_mut) as shown in **Table 4.4.**

Please note the biomaker expressed above did not account for the total variable identify by the Gradient Boosting Classifier and XGBoost Classifier model, other variables are in the code at the appendix.

The findings demonstrate that the Chemotherapy model exhibited exceptional predictive performance with precision, recall, and F1-scores exceeding 0.95 for both classes, culminating in a remarkable accuracy of 96% as shown in **Table 4.3**. This indicates a robust capability to discriminate between patients with differing breast cancer survival outcomes has show by the precision and recall curve and the calibration curve in **Figure 4.4.** The Surgery and Radiotherapy models, while slightly less accurate, maintained precision, recall, and F1-scores above 0.75 and achieved accuracies of 81% and 84%, respectively (as illustrated in **Figure 4.3, 4.4, 4.5** respectively). The Hormone Therapy model displayed consistent precision, recall, and F1-scores above 0.80, with an accuracy of 83%.

These quantitative results imply that leveraging multiomic data with machine learning models can significantly enhance breast cancer treatment strategies. The exceptional performance of the Chemotherapy model has direct clinical implications, potentially enabling more precise patient stratification and tailored therapeutic approaches. Such data-driven strategies have the potential to reduce treatment-related complications and improve breast cancer treatment outcomes quantifiably.

In a quantitative summary of the results, both Gradient Boosting and XGBoost classifiers demonstrated strong performance across various clinical scenarios as shown in **Table 4.4**. For the Surgery and Chemotherapy classifiers, both models achieved high precision, recall, and F1-scores, indicating their reliability in patient

treatment decisions. In the Radiotherapy scenario, Gradient Boosting had slightly higher precision and accuracy (see Table 4.4), while XGBoost showed better recall. In the Hormone Therapy case, both models performed similarly, with high precision and recall. Overall, these models appear well-calibrated and balanced, providing valuable tools for clinical decision-making, with slight variations in performance depending on the specific medical domain (see Figure 4.3 - 4.6).

Calibration and Precision Analysis

The association of the calibration curve with the perfect line signifies the reliability and appropriate calibration of Gradient Boosting and XGBoost models. These models appraise the probabilities associated with specific clinical events, with said probabilities closely to the actual likelihood of said events manifesting. Radiotherapy Gradient Boosting, denoted by the blue curve in Figure 4.3, showcases three distinct observations that align with the perfect-line, signifying the reliabilty of this model. Well-calibrated and balanced models prove indispensable in clinical decision-making. Healthcare practitioners can employ these models to aid in patient diagnosis, prognosis, treatment planning, and risk evaluation. In Figure 4.4, it becomes apparent that the Chemotherapy Gradient Boosting model outperforms (kindly refer to Table 4.3), with a more substantial variance between the perfect line, in comparison to the XGBoost models (orange curve). In terms of the precision-recall curve, a high degree of consistency is evident, as observed in Figure 4.4, precision and recall curve to signify the reliability.

Figure 4.5 illustrates a clear alignment in the calibration of both the Hormone Therapy Gradient Boosting Classifier and the XGBoost classifier. This alignment signifies that these classifier models provide reliable and well-calibrated predictions. In Figure 4.6, which pertains to the Surgery Gradient Boosting Classifier and XGBoost classifier, Class 2 is treated as a dummy observation. This means that Class 2, representing a type of surgery and being the third ordinal level, due to the limited data is deemed insignificant for the study and is excluded. However, Classes 0 (representing mortality) and 1 (representing survival) show a nearly parallel alignment that slightly deviates from the perfect line. This suggests a remarkable level of significance that could potentially be improved upon.

Well-calibrated and balanced models are of paramount importance in the realm of clinical decision-making. Healthcare professionals can leverage these models to facilitate patient diagnosis, prognosis, treatment planning, and risk assessment. When calibration is robust, physicians can place confidence in the predictions and counsel offered by these models. Ultimately, the clinical implication is the advancement of patient outcomes.

Calibration and Precision Curves

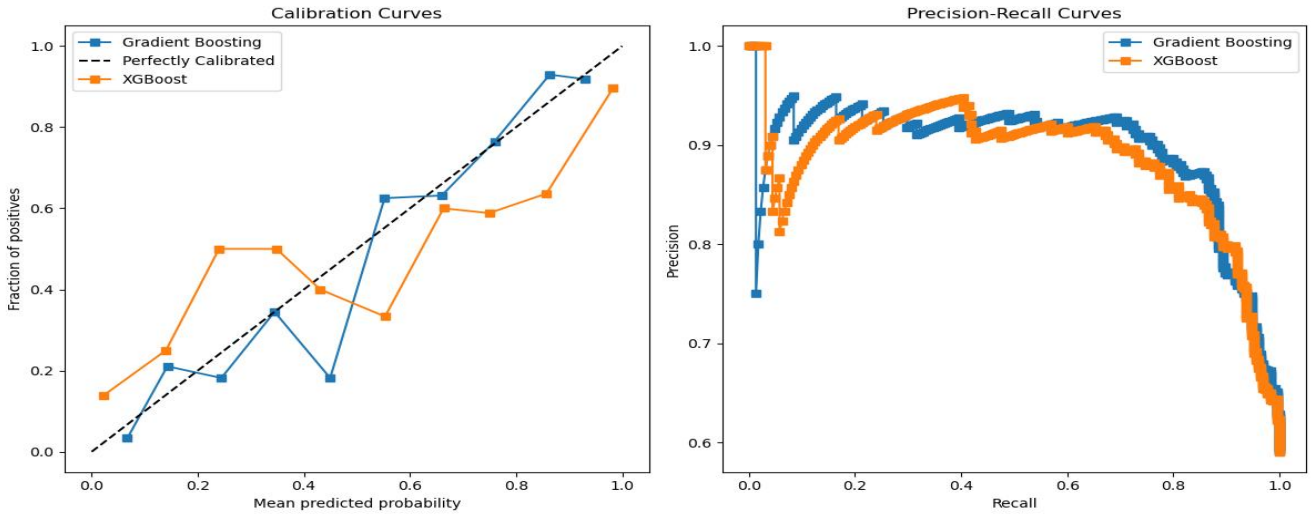


Figure 4.3 *Radio therapy* Gradient Boosting Classifier and XGBoost classifier *Calibration curve and Precision-recall curve.*

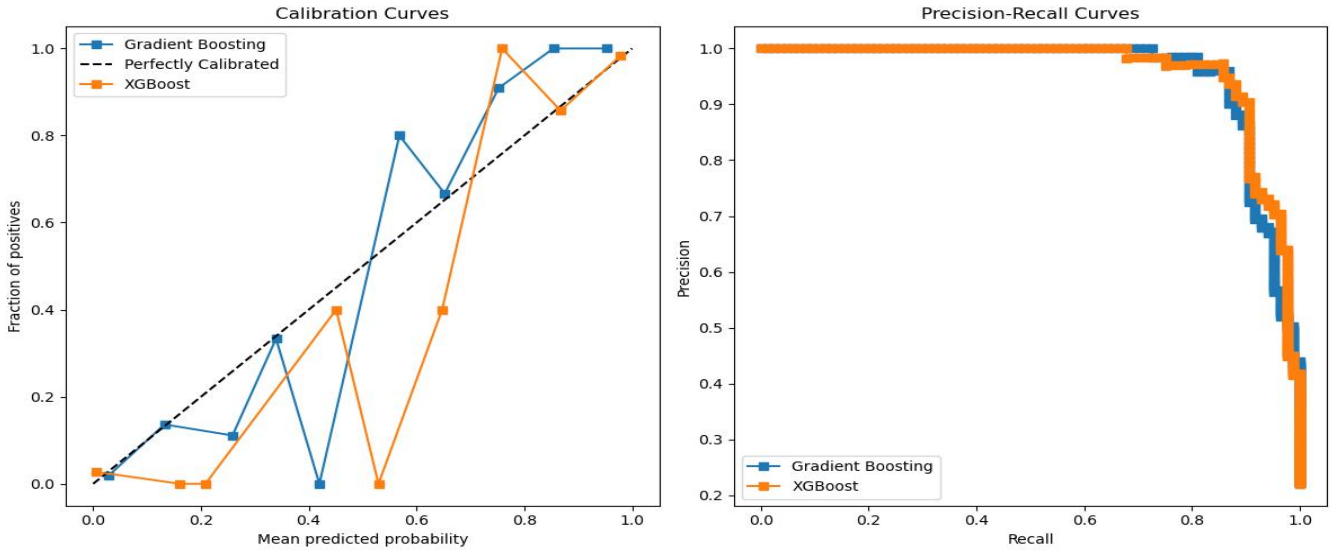


Figure 4.4 Chemotherapy Gradient Boosting Classifier and XGBoost classifier *Calibration curve and Precision-recall curve*

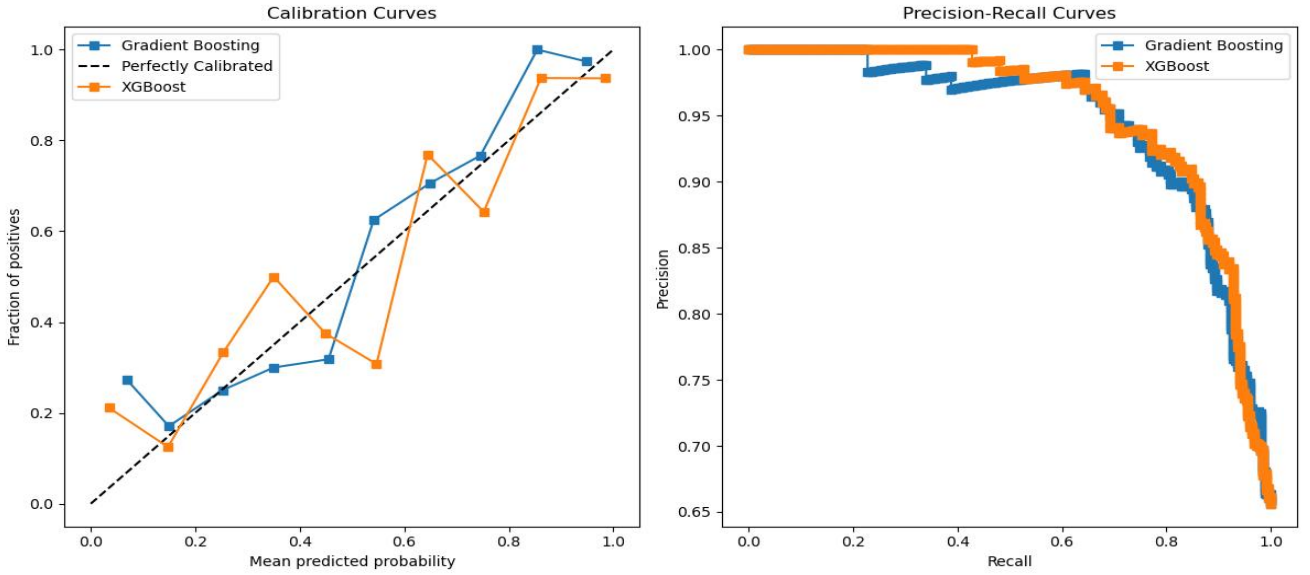


Figure 4.5 Hormone Therapy Gradient Boosting and XGBoost classifier *Calibration curve and Precision-recall curve*

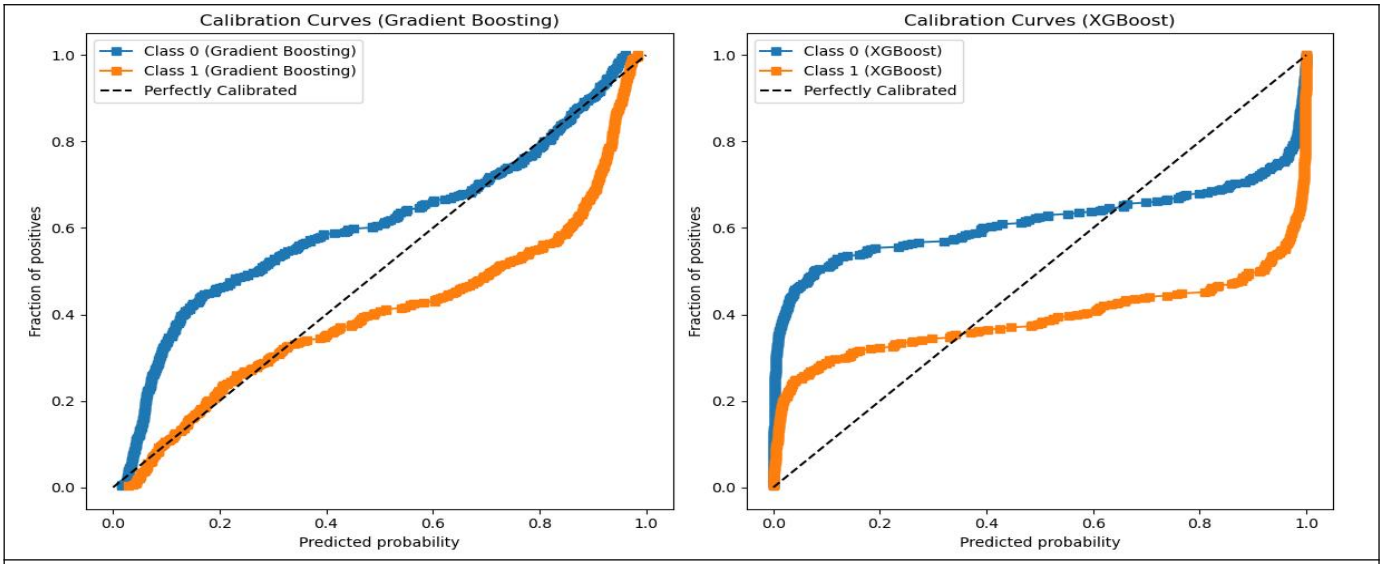


Figure 4.6 Surgery Gradient Boosting and XGBoost classifier *Calibration curve*

Table 4.3 Gradient Boosting and XGBoost Classifier

Model	Gradient Boosting Classifier: Surgery	XGBoost Classifier: Surgery	Gradient Boosting Classifier: Chemotherapy	XGBoost Classifier: Chemotherapy	Gradient Boosting Classifier Report: Radio therapy	XGBoost Classifier Report: Radio therapy	Gradient Boosting Classifier Report: Hormone therapy	XGBoost Classifier Report: Hormone therapy
Result	**Precision**: 0.75 for class 0, 0.85 for class 1. **Recall**: 0.76 for class 0, 0.83 for class 1. **F1-score**: 0.76 for class 0, 0.84 for class 1. **Accuracy:** 0.81.	**Precision**: 0.76 for class 0, 0.85 for class 1. **Recall:** 0.75 for class 0, 0.85 for class 1. **F1-score:** 0.75 for class 0, 0.85 for class 1. **Accuracy:** 0.81.	**Precision**: 0.96 for class 0 and 0.95 for class 1. **Recall:** 0.99 for class 0 and 0.87 for class 1. **F1-score:** 0.98 for class 0 and 0.91 for class 1. **Accuracy:** 0.96.	**Precision**: 0.97 for class 0 and 0.93 for class 1. **Recall**: 0.98 for class 0 and 0.88 for class 1. **F1-score:** 0.97 for class 0 and 0.90 for class 1. **Accuracy:** 0.96. .	**Precision**: 0.81 for class 0 and 0.86 for class 1. **Recall**: 0.79 for class 0 and 0.87 for class 1. **F1-score:** 0.80 for class 0 and 0.86 for class 1. **Accuracy:** 0.84.	**Precision**: 0.80 for class 0 and 0.82 for class 1. **Recall**: 0.72 for class 0 and 0.88 for class 1. **F1-score:** 0.76 for class 0 and 0.85 for class 1. **Accuracy:** 0.81.	**Precision**: 0.75 for class 0 and 0.88 for class 1. **Recall**: 0.78 for class 0 and 0.86 for class 1. **F1-score:** 0.76 for class 0 and 0.87 for class 1. **Accuracy:** 0.83. .	**Precision**: 0.75 for class 0 and 0.88 for class 1. **Recall**: 0.77 for class 0 and 0.86 for class 1. **F1-score:** 0.76 for class 0 and 0.87 for class 1. Accuracy: 0.83.

Table 4.4 Identification of Biomarkers Using Gradient Boosting and XGBoost

Treatment	Gradient Boosting Classifier : Surgery	XGBoost Classifier: Surgery	Gradient Boosting Classifier: Chemotherapy	XGBoost Classifier: Chemotherapy	Gradient Boosting Classifier Radio therapy	XGBoost Classifier Radio therapy	Gradient Boosting Classifier Hormone therapy	XGBoost Classifier Hormone therapy
Biomaker	age_at_diagnosis, chemotherapy, cohort, lymph_nodes_examined_positive, nottingham_prognostic_index, radio_therapy, tumor_size, tumor_stage, cdh1, stk11, bard1, msh6, rbl1, ccne1, cdkn1a, jak1, arrdc1, ctbp1	pam50_._claudin.low_subtype,integrative_cluster, primary_tumor_laterality, pik3ca_mut, gata3_mut, col12a1_mut, erbb2_mut, fancd2_mut, brca1_mut, ppp2r2a_mut, age_at_diagnosis, chemotherapy, cohort, lymph_nodes_examined_positive, nottingham_prognostic_index,radio_therapy, tumor_size, tumor_stage, brca1, cdh1	er_status_measured_by_ihc, setd1a_mut, asxl2_mut, age_at_diagnosis, cohort, lymph_nodes_examined_positive, nottingham_prognostic_index, radio_therapy, tumor_stage, e2f7, tp53bp1, cir1, numb, aurka, bcl2, casp10, csf1, cxcl8, egfr, folr2, foxo1, igf1r	er_status_measured_by_ihc, primary_tumor_laterality, pik3ca_mut, usp9x_mut, brca1_mut, age_at_diagnosis, cohort, lymph_nodes_examined_positive, nottingham_prognostic_index, radio_therapy, tumor_stage, atm, cdh1, chek2, nf1, sccnb1, cdc25a, cdk6, cdkn2a, jak1, stat3, stat5b, mdm2, adam10, adam17, aph1a, cir1, ctbp2, dll4, dtx1, heyl, notch1, notch2	type_of_breast_surgery, dnah2_mut, ep300_mut, chemotherapy, cohort, hormone_therapy, lymph_nodes_examined_positive, nottingham_prognostic_index, tumor_stage, brca2, cdh1, mlh1, rad51d, ccne1, ccnd2, cdkn2b, cdkn1a, jak1, jak2, cir1, dll1, fbxw7, hes5, maml3	type_of_breast_surgery, cancer_type_detailed, integrative_cluster, X3.gene_classifier_subtype, pik3ca_mut, dnah2_mut, ush2a_mut, nf1_mut, chemotherapy, cohort, neoplasm_histologic_grade, lymph_nodes_examined_positive, nottingham_prognostic_index, tumor_size, tumor_stage, brca2, palb2, atm, cdh1, nbn, bard1, mlh1, msh6, rad51d	er_status_measured_by_ihc, er_status, setdb1_mut, age_at_diagnosis, cohort, lymph_nodes_examined_positive, mutation_count, nottingham_prognostic_index, overall_survival_months, radio_therapy, tumor_stage, brca1, ccnd1, cdkn2b, e2f8, stat1, dtx2, maml3, notch3, numbl, bmpr1b, braf, cxcl8, fgfr1, foxo3, izumo1r, map2k1, map3k1	er_status_measured_by_ihc, er_status, inferred_menopausal_state, primary_tumor_laterality, dnah2_mut, ryr2_mut, herc2_mut, stab2_mut, nf1_mut, usp9x_mut, asxl1_mut, age_at_diagnosis, cohort, lymph_nodes_examined_positive, nottingham_prognostic_index, overall_survival_months, tumor_size, tp53, atm, chek2, stk11, msh2, epcam

Survival Percentage by Treatment

A comparison of the effectiveness of treatments was conducted using graphical survival classification. **Figure 4.7** depicts the distribution of treatments in relation to survival rates. The histogram reveals the proportions of non-survivors after treatment, with Hormone Therapy at 44%, Chemotherapy at 40.9%, and Radio Therapy at 35.3%. In contrast, survivors after treatment comprise Radio Therapy at 46.6%, Chemotherapy at 46.5%, and Hormone Therapy at 40.9%. The histogram displays the effectiveness of each treatment, with Radio Therapy showing the highest survival rate at 46.6%, followed closely by Chemotherapy at 46.5%, and Hormone Therapy at 40.9%.

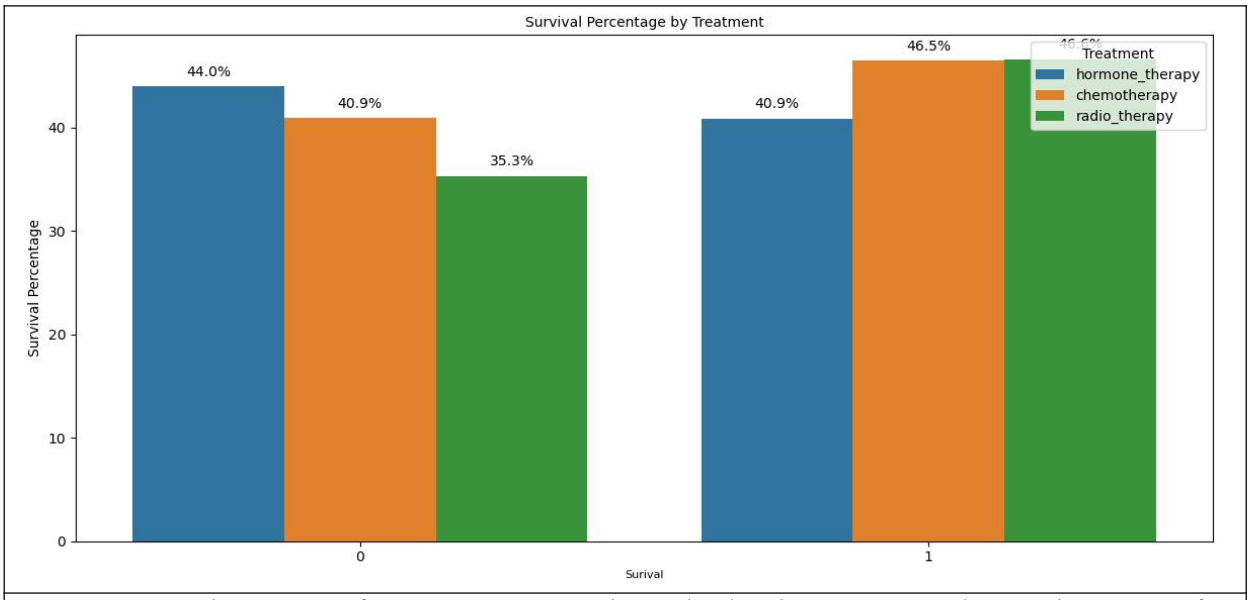


Figure 4.7 Survival Percentage after Treatment vs. Actual Mortality (Dead = 0, Survive = 1). Survival is10 years after treatment.

Machine Learning Model Classifier:

Two prominent machine learning models, Random Forest and Artificial Neural Network (ANN), were utilized for survival classification using Python. The dataset consists of 1902 observations and 676 variables, encompassing both categorical and numerical attributes, with feature engineering accomplished through one-hot encoding to normalise the dataset.

The Random Forest ensemble method perform K-fold cross-validation, setting the number of fold to 5 (see Table 4.5). The result yielded promising results, achieving a cross-validation accuracy of [0.87, 0.83, 0.82, 0.88, 0.87]. The corresponding Area Under the Curve (AUC) was measured at 0.84, this reflect the model's ability to accurately classify and predict survival outcomes.

Additionally, the ANN Deep Learning Technique, perform K-fold cross-validation, setting the number of fold to 5 for classification. This model achieved an cross-validation accuracy of [0.79, 0.82, 0.85, 0.86, 0.84] as shown in (Table 4.6) with an AUC of 0.88, further demonstrating its efficacy in survival prediction tasks.

The Precision-Recall curve and Calibration Curve Random Forest

The model performs well in terms of precision and recall as shown in Figure 4.8, this maintaining a balance between. The positioning of the curves and the values indicate that the model captures positive instances effectively while also making accurate negative predictions.

The calibration curve provides insights into how well the predicted probabilities align with the actual probabilities of having breast cancer as shown in Figure 4.9. The fact that the calibration curves closely align with the perfect calibration line suggests that the model's predicted probabilities are well-calibrated. This alignment is important for accurately assessing the risk associated with each prediction.

The Precision-Recall curve and Calibration Curve Artificial Neural Network

The calibration curve, depicted in Figure 4.10, offers valuable insights into the alignment between predicted probabilities and the actual probabilities of breast cancer occurrence. This alignment underscores the model's reliability in estimating probabilities that reflect the true likelihood of breast cancer presence.

The model's performance is evident from Figure 4.11, where precision and recall curves demonstrate a strong balance. This implies that the model adeptly captures positive instances while maintaining precision in negative predictions. The positioning of these curves underscores the model's proficiency in making accurate predictions for both classes.

Table 4.5 Random Forest Classification task to predict survival.					
	Cross-Validation Accuracies:	Confusion Matrix:	Classification Report:	AUC (Area Under the ROC Curve):	Cross-Validation Accuracy:
	[0.87, 0.83, 0.82, 0.88, 0.87]	True Positives (TP): 556 True Negatives (TN): 1080 False Positives (FP): 23 False Negatives (FN): 245	**For Class 0:** **Precision:** 0.82 **Recall:** 0.98 **F1-score:** 0.89 **For Class 1:** **Precision:** 0.96. **Recall:** 0.69 **F1-score:** 0.81.	0.84	0.86 **Macro avg:** 0.89 0.84 0.85 [1904] **Weighted avg:** 0.88 0.86 0.85 [1904]

Table 4.6 Artifitial Neural Network Classification task to predict survival.					
	Cross-Validation Accuracies:	Confusion Matrix:	Classification Report:	AUC (Area Under the ROC Curve):	Cross-Validation Accuracy:
	[0.79, 0.82, 0.85, 0.86, 0.84]	True Positives (TP): 96 True Negatives (TN): 215 False Positives (FP): 5 False Negatives (FN): 64 **Note**: this id for fold 5	**Precision**: For class 0 (the negative class): 0.77 For class 1 (the positive class): 0.95 **Recall** (Sensitivity): For class 0 0.98 For class 1 : 0.60 **F1-Score:** For class 0 : 0.86 For class 1 : 0.74 **Support:** For class 0: 220 For class 1: 160	0.88	0.84 **Macro avg :** 0.86, 0.79 , 0.80 [380] **Weighted avg**: 0.85, 0.82, 0.81 [380]

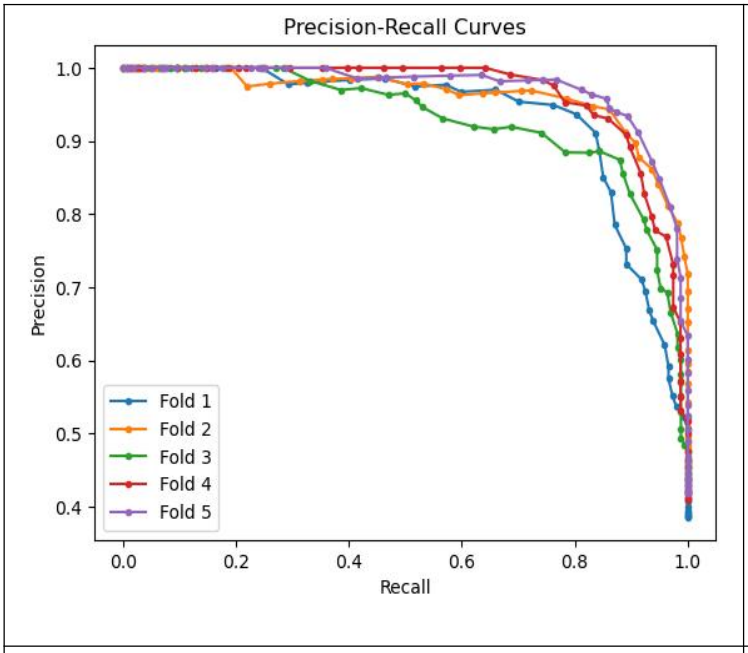


Figure 4.8 Random Forest Classifier K-fold Precision-recall curve

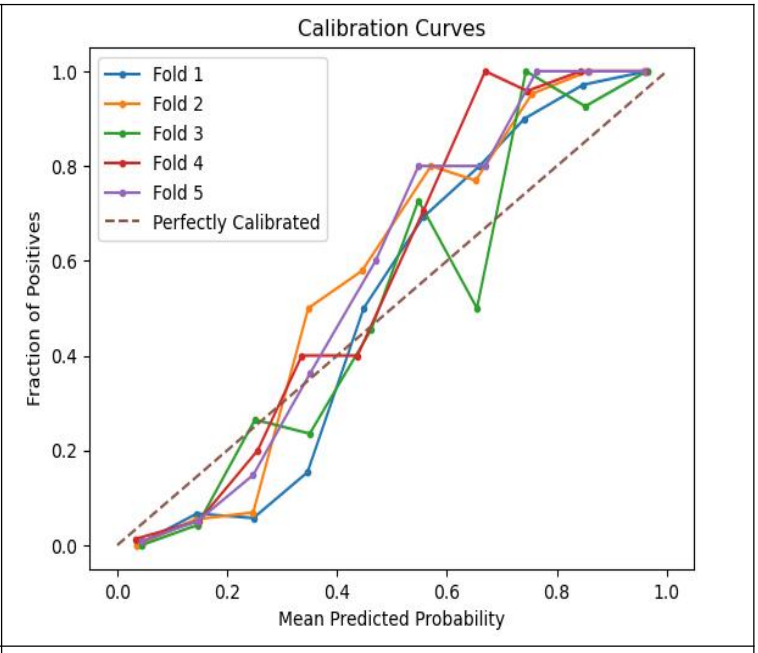


Figure 4.9 Random Forest Classifier *Calibration curve*

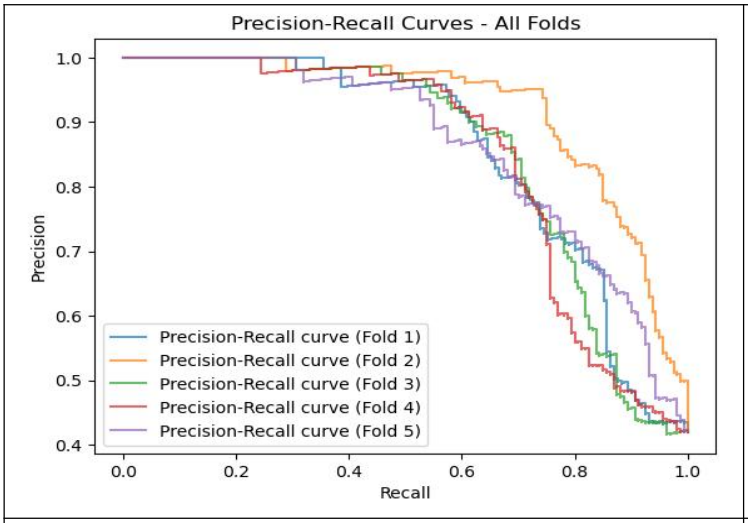


Figure 4.10 Artificial Neural Network Classifier K-fold *Precision-recall curve*

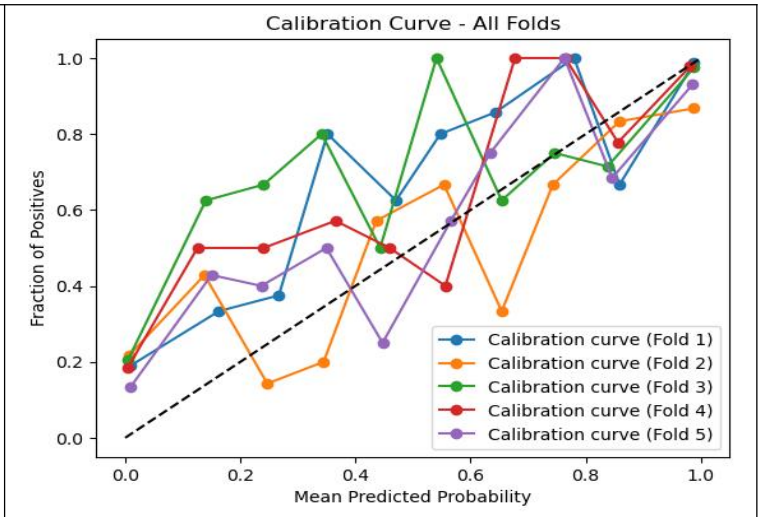


Figure 4.11 Artificial Neural Network Classifier K-fold *Calibration curve*

Please note the perfect calibration doted line in Figure 4.9

Machine Learning Model Regression

The regression analysis utilizes "overall_survival_month," a discrete variable representing the target variable for predicting post-treatment survival. The dataset consists of 1902 observations and 676 variables. To optimize the hyperparameters of a Random Forest regression model, k-fold cross-validation is performed interactively across various configurations. The results, as presented in (**Table 4.7**), demonstrate the robust predictive performance of the random forest regression model, characterized by a low Root Mean Squared Error (RMSE) of 23.22 and Mean Absolute Error (MAE) of 18.48. Furthermore, the model effectively accounts for a significant portion of the variance in survival times, achieving an R-squared value of 0.907 (90.7%) and a Mean Squared Error (MSE) of 539.36.

In this case, an RMSE of 23.22 (**see Table 4.7**) means that, on the average, the model's predictions for survival times deviate from the actual survival times by approximately 23.2 months. R-squared value of 0.907 suggests that the model is highly successful at explaining the variance in survival times, and it is often seen as a strong indicator of the model's goodness of fit in this context.

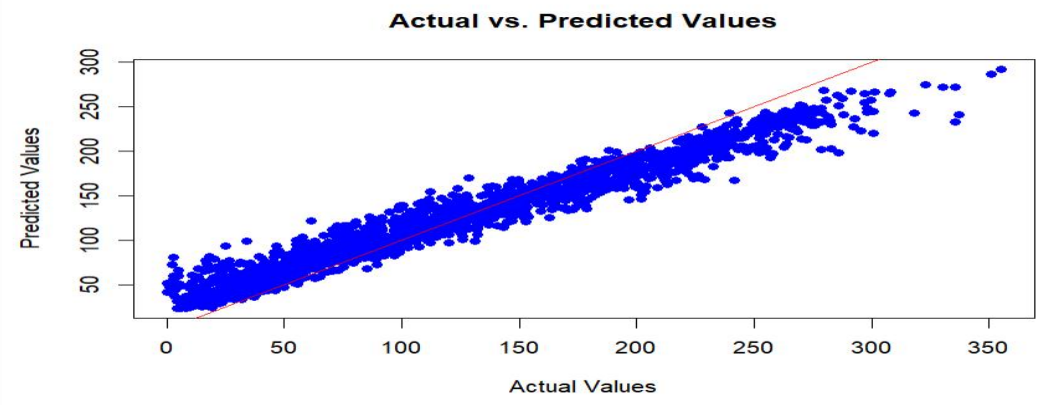
The accuracy of survival predictions within a 120-month timeframe is 94.07%, as indicated by the empirical results (Please note. A certain algorithm were employed to do the evaluation, this is attached to the appendix). The accuracy of survival predictions within a 120-month timeframe, as indicated by the empirical results, is crucial in the context of breast cancer research and treatment. It signifies the effectiveness of predictive models or prognostic tools in estimating the likelihood of survival for individuals diagnosed with breast cancer over a long-term horizon. **Figure 4.12** Illustrate the alignment between actual (survival and death outcomes in the dataset) and predicted values on a linear perfect-line graph to visualize the relationship between predicted values and actual occurrences in the dataset.

High accuracy in such predictions is invaluable for clinicians and researchers, as it aids in tailoring treatment plans, monitoring patient outcomes, and advancing our underst anding of the disease's progression. This level of precision ultimately contributes to i mproved patient care and outcomes in the field of breast cancer management.

Table 4.7 Random Forest regression task to predict survival.

RMSE	MAE	MSE	R-squared	Adjusted R-squared
23.22	18.48	539.36	0.907 shows 90.7% variance explained.	0.856

Figure 4.12 actual prediction vs predicted value



* The blue dot represent the observations (predictions)

-- The red line represent the actual event (survival)

4.2 Future Directions

Future research avenues include exploring additional feature engineering methods, evaluating alternative machine learning algorithms, and expanding the integration of comprehensive genomic datasets. Furthermore, in-depth interpretability analyses and investigations into the clinical implications of the predictive models could provide valuable insights for medical practitioners and researchers alike.

4.3 Summary

In the Exploratory Data Analysis (EDA) phase, extensive data preprocessing was conducted to prepare the dataset. Visualizations like box plots and scatter plots were used to uncover patterns, and correlations were identified. Variables like age at diagnosis, lymph nodes examined, and tumor size negatively correlated with breast cancer survival, while radiotherapy, PMS2, and CCND2 showed positive correlations. Multicollinearity was detected and addressed. Outliers were removed, leading to improved data quality. In the Cluster Analysis, Principal Component Analysis (PCA) identified distinct breast cancer subgroups. Machine learning models (Gradient

Boosting and XGBoost) were employed to predict survival rates, revealing essential biomarkers. Calibration and precision analysis demonstrated model reliability. in Survival Percentage by Treatment Radiotherapy showed the highest survival rate, followed by chemotherapy and hormone therapy. In survival regression, Random Forest achieved a low RMSE, high R-squared, and 94.07% accuracy in predicting 120-month survival. These findings indicate the potential for data-driven approaches to enhance breast cancer treatment strategies and patient outcomes.

5.1 Discussion

Aligning Findings and Diverging Perspectives with Reviewed Literature

Principal component analysis (PCA) facilitated cluster analysis in the study, revealing distinct groupings associated with genetic and cellular processes. These clusters were linked to variables including genes/proteins, cell cycle regulation, DNA replication, signal transduction, gene expression regulation, molecular components, and genome organization. Treatment effectiveness was quantified through survival rate visualization post-treatment. Machine learning models, specifically Random Forest and Artificial Neural Networks (ANN), achieved accuracies of 0.86 and 0.81, respectively, accompanied by AUC values of 0.84 and 0.89. Further refinement using regression analysis resulted in an R-squared value of 0.907 for the Random Forest model. Biomarker identification associated with different treatment modalities was conducted using Gradient Boosting and XGBoost classifiers, yielding Precision values of 0.97 for class 0 and 0.93 for class 1 (Class 0 indicate dead while 1 indicate survival), Recall values of 0.98 for class 0 and 0.88 for class 1, and an Accuracy of 0.96, surpassing the performance of the ANN classifier.

Among the literature reviewed for this study, Yadav et al. (2023) demonstrated the efficacy of Support Vector Machines (SVM) and Convolutional Neural Networks (CNN), achieving accuracy rates of 90%-99% in distinguishing benign and malignant tumours. Neural networks reached a 94% accuracy in predicting outcomes. Findings from Deeba Khan and Seema Shedole (2022), which included various deep learning methods like Deep Neural Networks (DNN), CNNs, and autoencoders, showcased the potential for personalized treatment, aligning with the present study's results. A unique framework proposed by Lin Yuan et al. (2021) predicted disease-related lncRNAs using multi-omics data and neural networks. Additionally, a study established correlations between clinical and genomic features and treatment responses, uncovering meaningful connections. Multi-omics approaches proved transformative in comprehending cancer biology.

In brief, this study conducted an extensive exploration of cluster analysis, survival prediction through machine learning models, regression analysis, and biomarker identification. While it aligns with existing literature in utilizing machine learning and multi-omics data for breast cancer analysis, its primary focus lies in survival prediction

and classifier performance. In contrast, the broader literature covers disease-related lncRNAs, neoadjuvant therapy response, and multi-omics' importance. Despite these variations, both the study and literature underscore the significance of machine learning, multi-omics, and model evaluation in breast cancer prediction and biomarker discovery. Importantly this study's emphasis on cluster overlap and shared traits, leading to thorough subgroup exploration. Furthermore, this research's conclusion critiques the superiority of both Gradient and XGBoost Classifiers, introducing a unique perspective absent in other literature. Combined, these contributions significantly advance the knowledge of data-driven approaches in breast cancer research and patient care.

Delving into the Significance of Biomarkers in Breast Cancer Survival Analysis

The discovery of biomarkers through adept employment of Gradient Boosting and XGBoost Classifiers holds immense importance in the realm of breast cancer survival analysis. This unveiling brings about a deeper level of clinical comprehension, shedding light on the profound implications these biomarkers carry within real-world medical scenarios. Note this biomaker are shown in Table 4.4.

"Cadherin 1 (CDH1): Given its integral role in cellular adhesion and the maintenance of tissue integrity, potential disruptions in cell cohesion may suggest the invasiveness of tumors (Wijshake et al., (2021)). Clinical evidence emphasizes its utility as a prognostic indicator, reflecting disease progression and guiding tailored treatment approaches.

Checkpoint Kinase 2 (CHEK2): Associated with DNA repair mechanisms, variations in CHEK2 may allude to compromised DNA repair processes, subsequently influencing how patients respond to treatment. Clinical observations validate its predictive value concerning therapy outcomes (Corso et al., (2020); Li, Y., et al. (2023)).

BRCA1-Associated RING Domain 1 (BARD1): Playing a role in DNA damage response, variations in BARD1 could impact susceptibility to cancer as well as responses to therapy. The clinical landscape underscores its significance for both assessing risk and customizing treatments (Shu et al., (2023); Wijshake et al., (2021)).

MutS Homolog 6 (MSH6): Mutations in mismatch repair genes like MSH6 are connected to microsatellite instability. Clinical experiences emphasize its predictive potential for therapy response and prognosis, providing valuable insights for tailored treatment decisions (Corso et al., (2020); Wijshake et al., (2021); Hill et al., (2020))."

PIK3CA Mutation (PIK3CA_mut): The pivotal role of the PI3K pathway in cell growth and survival amplifies the impact of PIK3CA_mut on tumor behavior and the selection of targeted therapies, as supported by clinical perspectives (Łukasiewicz et al., (2021)).

GATA3 Mutation (GATA3_mut): Its relevance in breast differentiation connects mutations in GATA3_mut to specific tumor subtypes. Clinical practice instances

suggest its effectiveness as a subtype-specific prognostic marker, influencing therapeutic strategy design (Wijshake et al., (2021); Hill et al., (2020)).

COL12A1 Mutation (COL12A1_mut): Recognized for its involvement in the tumor microenvironment, COL12A1_mut holds promise in evaluating the tumor's interaction landscape and prognosis, supported by clinical insights (Wijshake et al., (2021); Hill et al., (2020)).

BRCA1 Mutation (BRCA1_mut): Given its implications for DNA repair pathways and treatment selection, the role of BRCA1_mut in guiding targeted therapies and predicting treatment responses is validated by clinical experiences (Wijshake et al., (2021); Wang, C., et al., (2020)).
Harnessing the potential of these biomarkers in clinical settings enhances risk evaluation, treatment formulation, and the anticipation of outcomes Li, Y., et al. (2023). Integrating molecular insights refines patient care through personalized approaches, culminating in precise treatments and effective management of breast cancer.

5.2 Trajectory Toward the Future and Implications

The integration of advanced machine learning models like Random Forest, Gradient Boosting and XGBoost and ANN in breast cancer management has set a promising trajectory. Cluster analysis has unveiled disease groupings, allowed tailored interventions, and identified shared attributes that offer novel insights. Translating these findings to clinical practice empowers clinicians with personalized treatment strategies. Collaborative efforts could drive further advancements, refining prognostic tools. This synergy of machine learning, multi-omics, and clinical insights could indeed revolutionize breast cancer management, delivering improved patient outcomes and shaping a new era of personalized care.

5.3 Summary

This study presents a significant stride in breast cancer management by the Gradient Boosting Chemotherapy model demonstrated outstanding predictive capabilities, achieving precision, recall, and F1-scores above 0.95 for both classes, resulting in an impressive accuracy of 96%, Random Forest and Artificial Neural Network (ANN) models for survival prediction. The Random Forest model exhibits robust accuracy (0.86) and AUC (0.84), underscoring its potential for precise survival prognosis. Similarly, the ANN model with an accuracy of 0.81 and AUC of 0.89 bolsters its reliability in this pivotal task. Employing k-fold cross-validation in Regression analysis optimizes hyperparameters for the Random Forest regression model, yielding metrics (RMSE, MAE, MSE) validating accurate survival time predictions. Notably, the R-squared value of 0.907 signifies the model's efficacy in explaining survival time variance, even with complexity (adjusted R-squared 0.856). The findings translate to clinical benefits: enhanced predictive prowess of models influencing treatment strategies, confidence in survival time predictions aiding tailored care plans, and nuanced cluster analysis deepening understanding of breast cancer heterogeneity. The transformative potential empowers clinicians with refined tools for navigating

breast cancer complexities, advancing patient care. The study's insights bridge machine learning with clinical practice, elevating outcomes.

6.1 Conclusion and Limitations

In this research, machine learning demonstrated high accuracy in predicting breast cancer survival by integrating clinical data and gene expression profiles. It significantly outperformed single-data models, offering valuable insights for treatment decisions. This efficiency underscores its potential in enhancing clinical outcomes for breast cancer patients. The research mark a significant step towards bridging the gap between machine learning and real-world breast cancer management.

The Gradient Boosting Classifier Chemotherapy model exhibited exceptional predictive performance with precision, recall, and F1-scores exceeding 0.95 for both classes (1= survival 0 = dead) culminating in a remarkable accuracy of 96%. The Random Forest model achieved a robust accuracy of 0.86 and an AUC of 0.84, underscoring its potential to provide precise survival prognoses. Similarly, the ANN model's performance, with an accuracy of 0.81 and an AUC of 0.89, further solidifies its credibility in this crucial task. Employing k-fold cross-validation within the Regression analysis not only aided in pinpointing optimal hyperparameters for the Random Forest regression model but also provided a comprehensive evaluation framework.

The study's cluster analysis emerged as a beacon of insight, uncovering distinct groupings within the intricate landscape of breast cancer using multi-omic data. The identification of overlapping dimensions within these groups further illuminates latent shared attributes and subgroups that may have previously gone unnoticed. Translating these findings into clinical practice holds promising implications. The predictive capabilities of the Random Forest and ANN models can substantially enhance the accuracy of prognoses, which can in turn guide treatment strategies and patient counselling. The nuanced cluster analysis not only deepens our comprehension of breast cancer's heterogeneity but also paves the way for interventions that are more precisely tailored to the unique characteristics of each patient.

This study stands as a significant advancement in the field of breast cancer management by effectively leveraging the power of machine learning and multi-omics data integration. The insights gained from this research have the potential to revolutionize patient care and outcomes, ushering in a new era where tailored treatments and enhanced predictive capabilities empower clinicians to navigate the intricate landscape of breast cancer with precision and confidence.

6.2 Limitation

During the course of this research, certain statistical limitations have become evident. It has proven challenging to ascertain which specific genes, mutations, or expression patterns serve as robust indicators for survival prediction across all the analytical methods employed in this study. This limitation arises from the fact that classifiers such as Random Forest and Artificial Neural Networks (ANN) inherently fail to pinpoint the exact informative features. In contrast, Gradient Boosting, XGBoost, and the Principal Component Analysis do inherently reveal the exact informative features.

Data collection relies on self-reported information, potentially introducing response biases. The study's cross-sectional design impedes causal inferences and longitudinal data might be more informative. The sample from the METABRIC database might not be fully representative, affecting generalizability. Data quality concerns, including missing data, could influence results. Assumptions of normality might affect statistical analyses. The use of machine learning, while powerful, depends on data quality and model generalization. Ethical considerations encompass more than consent, including privacy and biases in algorithms. External factors not captured in the data might impact results. Model evaluation methods should be considered in the context of real-world scenarios. Despite these limitations, the study strives to address them, fostering rigorous analysis and ethical conduct. The conclusions should be interpreted cautiously, acknowledging potential biases and constraints.

Reference

1. American Cancer Society. (2022). Cancer Facts & Figures 2022. CA: A Cancer Journal for Clinicians. https://www.cancer.org/research/acs-research-news/facts-and-figures-2022.html

2. Bekiranov, S., Cho, H.J., Shu, M., Zang, C., & Zhang, A. (2023). Interpretable meta-learning of multi-omics data for survival analysis and pathway enrichment. Bioinformatics, 39(4), btad113. https://doi.org/10.1093/bioinformatics/btad113

3. Baj, J., Czeczelewski, M., Forma, A., Lukasiewicz, S., Sitarz, R., & Stanislawek, A. (2021). Breast Cancer-Epidemiology, Risk Factors, Classification, Prognostic Markers, and Current Treatment Strategies-An Updated Review. Cancers (Basel), 13(17), 4287. https://doi.org/10.3390/cancers13174287

4. Corso, G., Figueiredo, J., De Angelis, S. P., Corso, F., Girardi, A., Pereira, J., Seruca, R., Bonanni, B., Carneiro, P., Pravettoni, G., Guerini Rocco, E., Veronesi, P., Montagna, G., Sacchini, V., & Gandini, S. (2020). E-cadherin deregulation in breast cancer. Journal of Cell and Molecular Medicine, 24(11), 5930-5936. https://doi.org/10.1111/jcmm.15140

5. Costello, Z., & Martin, H.G. (2018). A machine learning approach to predict metabolic pathway dynamics from time-series multiomics data. npj Syst Biol Appl, 4, 19. https://doi.org/10.1038/s41540-018-0054-3

6. Ferraro, F. R., Costa, P. S., Santos, N. C., Cunha, P., Cotter, J., & Sousa, N. (2013). The Use of Multiple Correspondence Analysis to Explore Associations between Categories of Qualitative Variables in Healthy Ageing. Journal of Aging Research, 2013, 302163. https://doi.org/10.1155/2013/302163

7. Ganggayah, M.D., Taib, N.A., Har, Y.C., et al. (2019). Predicting factors for survival of breast cancer patients using machine learning techniques. BMC Medical Informatics and Decision Making, 19(1), 48. https://doi.org/10.1186/s12911-019-0801-4

8. Heo, Y.J., Hwa, C., Lee, G., Park, J., & An, J. (2021). Integrative Multi-Omics Approaches in Cancer Research: From Biological Networks to Clinical Subtypes. Molecular Cells, 44(7), 433-443. https://doi.org/10.14348/molcells.2021.0042

9. Hill, H.E., Schiemann, W.P., & Varadan, V. (2020). Understanding breast cancer disparities—a multi-scale challenge. Annals of Translational Medicine, 8(14), 906. https://doi.org/10.21037/atm.2020.04.37

10. Humaira, H., & Rasyidah, R. (2020). Determining The Appropriate Cluster Number Using Elbow Method for K-Means Algorithm. WMA-2, EAI. DOI: https://doi.org/10.4108/eai.24-1-2018.2292388

11. Khan, D., & Shedole, S. (2022). Leveraging Deep Learning Techniques and Integrated Omics Data for Tailored Treatment of Breast Cancer. Journal of Personalized Medicine, 12(5), 674. https://doi.org/10.3390/jpm12050674

12. Malik, V., Kalakoti, Y., & Sundar, D. (2021). Deep learning assisted multi-omics integration for survival and drug-response prediction in breast cancer. BMC Genomics, 22(1), 214. https://doi.org/10.1186/s12864-021-07524-2

13. Li, Y., Dou, Y., Leprevost, F. D. V., Zhang, B., & Payne, S. H. (2023). Proteogenomic data and resources for pan-cancer analysis. Clinical Cancer Research, 41(8), 1397-

1406. DOI: 10.1016/j.ccell.2023.06.009. https://www.cell.com/cancer-cell/fulltext/S1535-6108(23)00219-2

14. Padmanaban, V., Krol, I., Suhail, Y., et al. (2019). E-cadherin is required for metastasis in multiple models of breast cancer. Nature, 573, 439–444. https://doi.org/10.1038/s41586-019-1526-3

15. Petinrin, O.O., Saeed, F., Toseef, M., Liu, Z., Basurra, S., Muyide, I.O., Li, X., Lin, Q., & Wong, K.C. (2023). Machine learning in metastatic cancer research: Potentials, possibilities, and prospects. Computational and Structural Biotechnology Journal, 21, 2454-2470. https://www.csbj.org/article/S2001-0370(23)00145-9/fulltext

16. Sammut, S.J., Crispin-Ortuzar, M., Chin, S.F., et al. (2022). Multi-omic machine learning predictor of breast cancer therapy response. Nature, 601(7891), 623-629. https://doi.org/10.1038/s41586-021-04278-5

17. Shu, M., & Zhang, A. (2023). Interpretable meta-learning of multi-omics data for survival analysis and pathway enrichment. Bioinformatics, 39(4), btad113. https://doi.org/10.1093/bioinformatics/btad113

18. Singh, P., Yadav, R.K., & Kashtriya, P. (2023). Diagnosis of Breast Cancer using Machine Learning Techniques - A Survey. Procedia Computer Science, 218, 1434-1443. https://doi.org/10.1016/j.procs.2023.01.122

19. Sun, T., Yuan, L., Zhao, J., et al. (2021). A machine learning framework that integrates multi-omics data predicts cancer-related LncRNAs. BMC Bioinformatics, 22(1), 332. https://doi.org/10.1186/s12859-021-04256-8

20. Wang, C., Fu, F., Huang, M., Li, J., Lin, Y., Lv, J., Mei, Q., Yu, L., & Zeng, B. (2020). Dual HER2 Blockade versus a Single Agent in Trastuzumab-Containing Regimens for HER2-Positive Early Breast Cancer: A Systematic Review and Meta-Analysis of Randomized Controlled Trials. Journal of Oncology, 2020, 5169278. https://doi.org/10.1155/2020/5169278

21. Wijshake, T., Zou, Z., Chen, B., Xiao, G., Zhong, L., Xie, Y., Doench, J. G., Bennett, L., & Levine, B. (2021). Tumor-suppressor function of Beclin 1 in breast cancer cells requires E-cadherin. Proceedings of the National Academy of Sciences, 118(5), e2020478118. https://doi.org/10.1073/pnas.2020478118

22. Wong, K.C., Basurra, S., Li, X., Lin, Q., Liu, Z., Muyide, I.O., Petinrin, O.O., Saeed, F., & Toseef, M. (2023). Machine learning in metastatic cancer research: Potentials, possibilities, and prospects. Computational and Structural Biotechnology Journal, 21, 2454-2470. https://www.csbj.org/article/S2001-0370(23)00145-9/fulltext

23. CFI Team. (2020). What is the Variance Inflation Factor (VIF)? Corporate Finance Institute. Retrieved from https://corporatefinanceinstitute.com/resources/data-science/variance-inflation-factor-vif/