Computational Intelligent Framework for Biomarker Identification in Multi-Omics Data

A Thesis

Submitted for the award of degree of

DOCTOR OF PHILOSOPHY

Submitted By

Arwinder Dhillon

(901903008)

Under the supervision of

Dr. Ashima Singh

Associate Professor, CSED

TIET, Patiala

Dr. Vinod Kumar Bhalla

Associate Professor, CSED

TIET, Patiala

Computer Science and Engineering Department

Thapar Institute of Engineering and Technology, Patiala

(Deemed to be University)

April 2024

Candidate Declaration

I hereby certify that the work, which is being presented in the thesis, entitled **Computational Intelligent Framework for Biomarker Identification in Multi-Omics Data**, in partial fulfillment of the requirements for the award of the degree of **Doctor of Philosophy** and submitted in Computer Science and Engineering Department, Thapar Institute of Engineering and Technology (Deemed University), Patiala, India is an authentic record of my own work carried out under the supervision of **Dr. Ashima Singh** and **Dr. Vinod Kumar Bhalla**. I have also cited the reference about the text(s)/figure(s)/table(s) from where they have been taken.

The matter presented in this thesis has not been submitted elsewhere for the award of any other degree or diploma from any institution.

Arwinder Dhillon

Date:

Arwinder Dhillon
Regd. No: 901903008

This is to certify that the above statement made by the candidate is correct to the best of our knowledge.

Ashima

Dr. Ashima Singh
Associate Professor
Computer Science and Engineering Department
Thapar Institute of Engineering and Technology
Patiala

Vinod K Bhalla

Dr. Vinod Kumar Bhalla
Associate Professor
Computer Science and Engineering Department
Thapar Institute of Engineering and Technology
Patiala

Acknowledgement

I am grateful to the Almighty for granting countless blessings, opportunities, knowledge, and strength, without which I could not have accomplished this work.

I am indebted to my supervisors, Dr. Ashima Singh (Associate Professor, TIET) and Dr. Vinod Kumar Bhalla (Associate Professor, TIET), for their continuous guidance and support. I have benefitted greatly from their wealth of knowledge and meticulous editing. The time and energy they devoted to this work have immensely improved the quality of this research. Their faith in me and encouragement have helped me stay focused over the years. Their honest approach towards research and life is a source of inspiration, which I hope to carry throughout my career.

I offer my sincere gratitude to our director Prof. Padmakumar Nair, Dr. N. Tejo Prakash (Dean, Research Sponsored Projects) and Dr. Shalini Batra (Professor and Head, CSED) for providing the necessary academic and administrative assistance in the completion of this work. I am thankful to my Doctoral committee members- Dr. Anju Bala (Associate Professor, CSED), Dr. Prashant Singh Rana (Associated Professor, CSED), and Dr. Vikas Handa (Assistant Professor, Department of Biotechnology) for ensuring the progress of my research work. I am also thankful to Ph.D. Coordinator Dr. Sushma Jain (Associate Professor, CSED) and all the faculty and staff members of CSED for always helping me.

It would not have been possible without my mother's unconditional love and belief in me. My deepest thanks to my dear father, mother, and brother for encouraging me to start this journey and providing me with every possible support. The love showered by my sister-in-law, and cousins have also been a great motivation in this journey. I will always be grateful to them for their patience and compassion in every sphere of my life.

I wish to extend my thanks to all the lab mates and peers for their constant support and for making this journey memorable. I treasure Arun Rana, Govind Ram Chhimpa, Komal Singh Gill, and Parampreet Kaur for their support and encouragement. I gratefully acknowledge their cooperation and blessings, which motivated me throughout this degree.

Arwinder Dhillon

Arwinder Dhillon

April, 2024

Abstract

Omics data, encompassing genomics, proteomics, transcriptomics, and metabolomics, is generated through cutting-edge sequencing and mass spectrometry technologies. Biomarker identification, crucial in omics data analysis, relies on Deoxyribonucleic Acid (DNA), Ribonucleic Acid (RNA), and protein indicators to reveal physiological processes and disease symptoms. Leveraging machine learning and deep learning in computational bioinformatics enables the identification of biomarkers across single and multi-omics datasets, offering groundbreaking potential for early disease prediction. Integration of computational technologies with multi-omics data revolutionizes healthcare by facilitating advanced insights, aiding in disease diagnosis, prognosis, and targeted therapy development, thus advancing human health outcomes.

This research aims to utilize computational technologies like machine learning, deep learning, and statistical methods for effective biomarker identification using multi-omics data, targeting disease survival prediction, subtype classification, and disease prediction. Beginning with a comprehensive review, the study explores intelligent computational approaches for biomarker identification across single and multi-omics datasets. It identifies a significant demand for a tailored framework specifically designed for biomarker identification using multi-omics data, highlighting shortcomings in existing tools related to data pre-processing, feature selection, biomarker validation, and prediction model creation. To bridge these gaps, the research proposes a novel framework for biomarker identification in multi-omics analysis, aiming to empower researchers with accessible and comprehensive options for conducting biomarker identification effectively.

The framework is proposed for biomarker identification in multi-omics data which consists of six phases, i.e., data acquisition, data preprocessing, feature/ biomarker identification, biological interpretation, modeling, and performance evaluation. Through the data acquisition phase, omics data is collected from public repositories, i.e., The Cancer Genome Atlas (TCGA), Molecular Taxonomy of Breast Cancer International Consortium (METABRIC), and Religious Orders Study and Rush Memory and Aging Project (ROSMAP). The data preprocessing is performed by removal and imputation of null values, data normalization, and removal of duplicate samples. Additionally, feature/ biomarker identification is done using three approaches, comprising, 1. statistical methods and Random Spatial Local Best Cat Swarm Optimization (RSLBCSO), 2. Multimodal Varia-

tional autoencoder (MVAE), 3. CpG site Aggregation, statistical methods, and Light Gradient Boosting Machine Recursive Feature Elimination (LGBMRFE). The extracted biomarkers are validated using DAVID analysis and Kalpan Meier (KM) plots in the biological interpretation phase. In the development of modeling phase, the features from different omics are integrated and three models have been developed comprising Bayesian optimized Deep Neural Network (DNN) model, Simplified Graph Convolutional Networks (SGC), and stacked ensemble model for survival prediction, subtype classification and disease prediction, respectively. The three feature/ biomarker selection techniques and models are combined named as BioSurv, iMVAN, and HBS-STACK which are designed for biomarker identification in multi-omics for survival prediction, subtype classification, and disease prediction respectively. The performance of proposed framework is evaluated using various performance parameters in the performance evaluation phase.

The integration of computational techniques such as ML, DL, and statistical methods has significantly improved the precise identification of biomarkers using multi-omics data. The proposed approaches including BioSurv, iMVAN, and HBS-STACK exhibit high accuracies in survival prediction, disease subtype classification, and disease prognosis on multi-omics datasets. These approaches yield critical biomarker insights crucial for early disease detection, customized treatment strategies, and informed clinical decisions.

Table of Contents

List of Figures

List of Tables

List of Algorithms

List of Abbreviations

AE	Autoencoders
AI	Artificial Intelligence
AMP-AD	Accelerating Medicine Partnership Alzheimer's Disease
ANN	Artificial Neural Network
ANOVA	Analysis of Variance
AUC	Area Under Curve
AUPR	Area Under Precision Recall
BRCA	Breast Carcinoma
CAE	Cycle Autoencoder
CCA	Canonical Correlation Coefficient
CESC	Cervival and Endocervical Cancer
CHI2	Chi-squared
CI	Concordance Index
circ RNAs	Circular Ribonucleic acid
CNN	Convolutional Neural Networks
CNV	Copy Number Variation
CoCA	Cluster of Cluster Analysis
coxPH	Cox Proportional Hazard
CPR	Constrained Page Rank
CSCC	Cervical Squamous Cell Carcinoma
CSO	Cat Swarm Optimization
CVAE	Conditional Variational Autoencoder
DAE	Denoising Autoencoder
DAVID	The Database for Annotation, Visualization and Integrated Discovery
DBN	Deep Belief Network
DCGs	Differentially Coexpressed Genes
DEGs	Differentially Expressed Genes
DL	Deep Learning
DM	DNA Methylation
DMGs	Differentially Methylated Genes
DMSs	Differentially Methylated Sites
DNA	Deoxyribonucleic acid
DNN	Deep Neural Network

DSLF	Deep Space Latent Fusion
DT	Decision Tree
ECA	Endo Cervical Adenocarcinoma
EI	Expected Improvement
ER	Estrogen Receptor
ESCC	Esophageal Squamous Cell Carcinoma
FA	Factor Analysis
FC	Fold Change
FDR	False Discovery Rate
FLOPs	Floating Point Operations
FP	False Positive
FN	False Negative
GAN	Generative Adversarial Networks
GBM	Gradient Boosting Machine
GCN	Graph Convolutional Neural Network
GDAC	Genomics Data Analysis Center
GDC	Genomics Data Common Data
GEO	Gene Expression Omnibus
GO	Gene Ontology
GNU	General Public Licence
HCC	Hepatocellular Carcinoma
HER2	human epidermal growth factor receptor 2
HMIM	Human Molecular Interaction Network
HNSCC	Head and Neck Squamous Cell Carcinoma
HR	Hazard Ratio
HTH	Hubei Taihe Hospital
ICA	Independent Component analysis
IDE	Integrated Development Environment
IG	Information Gain
JNMF	Joint Non-negative Matrix Factorization
KEGG	Kyoto Encyclopedia of Genes and Genomes
KIHC	Kidney Chromophob
KIPAN	Pan Kidney Cohorent
KIRC	Kidney Clear Cell Carcinoma
KIRP	Kidney Papillary Cell Carcinoma
KL	Kullback Leibler
KM	Kaplen Meier

KNN	K Nearest Neighbor
KPCA	Kernel Principal Component Analysis
LASSO	Least Absolute Shrinkage and Selection Operator
LDL	Low Density Lipoproteins
LGBMRFE	Light Gradient Boosting Machine with Recursive Feature Elimination
LIHC	Liver Hepatocellular Carcinoma
LR	Logistic Regression
lSVR	Linear Support Vector Regression
LUAD	Lung Adenocarcinoma
METABRIC	Molecular Taxonomy of Breast Cancer International Consortium
MCC	Mathews Correlation Coefficient
MCMC	Markov Chain Monte Carlo
MIBC	Muscle Invasive Bladder Cancer
miRNA	micro Ribonucleic Acid
mRMR	Minimum Redundancy Maximum Relevance
mRNA	messenger Ribonucleic Acid
ML	Machine Learning
MS	Mass Spectrometry
MSE	Mean Square Error
MVAE	Multi-modal Variational Autoencoder
NB	Naive Bayes
NCA	Neighborhood Component Analysis
NCBI	National Center for Biotechnology Information
NN	Neural Network
NSCLC	Non small cell lung cancer
OMIM	Online Mendelian Inheritance in Man
OV	Ovarian Cancer
PANDA	Prioritization of Autism using Network based Deep Learning Approach
PCA	Principal Component Analysis
PCC	Pearson Correlation Coefficient
PDAC	Pancreatic Adenocarcinoma
PI	Probability of Improvement
PPI	Protein Protein Interaction
PR	Progesterone Receptor
PSN	Patient Similarity Network
RBF	Radial Basis Function
RDFS	Random Forest Feature Selection

ReLU	Rectified Linear Unit
RF	Random Forest
RFE	Recursive Feature Elimination
RNN	Recurrent Neural Networks
rppa	Reverse Phase Protein Array
RRA	Reboust Rank Aggregiation
RSF	Random Survival Forest
RSLBCSO	Random Spatial Local Best Cat Swarm Optimization
SCC	Spearman Correlation Coefficient
SCR-SNN	Sparse Crossmodel Superlayered Neural Network
SFARI	Simons Foundation Autism Research Institute
SGC	Simplified Graph Convolutional Network
SM	Somatic Mutation
SNF	Similarity Network Fusion
SVM	Support Vector Machine
TCGA	The Cancer Genome Atlas Portal
THCA	Thyroid Carcinoma
TNBC	Tripple Negative Breast Cancer
TP	True Positive
TN	True Negative
UCB	Upper Confidence Bound
VAE	Variational Autoencoders
WGCNA	Weighted Correlation Network Analysis

Chapter 1
Introduction

Omics data refers to high-throughput genetic data generated by various "omics" technologies, which are used to study different aspects of organisms. Omics data includes genomics, transcriptomics, proteomics, and metabolomics. Biomarker Identification is one of the significant areas in omics data analysis. Biomarkers having Deoxyribonucleic Acids (DNAs) at the genomics level, Ribonucleic Acids (RNAs) at transcriptomics, and Reverse Phase Protein Arrays (rppas) at the proteomics level serve as vital indicators of physiological processes, symptoms of diseases, normalcy, and irregularities in humans. In the complex field of medicine, biomarkers are indispensable tools for understanding, diagnosing, and managing various health conditions. It aids healthcare professionals and researchers to dig into the actual root cause and navigating unexplored areas related to human conditions, directing the diagnosis, prognosis, and monitoring of various medical ailments as part of treatment protocols.

Multi-omics data or integrated omics refers to the simultaneous study and analysis of different types of omics data within a biological system. Biomarker identification can be done in single omics and multi-omics. Computational bioinformatics represents an interdisciplinary domain that devises and employs computational methods for the analysis of extensive multi-omics datasets encompassing genomics, transcriptomics, and protein samples to make disease predictions. Computational bioinformatics plays a crucial role in dealing with multi-omics data for biomarker identification. Various computational approaches such as machine learning (ML), deep learning (DL), and statistical methods have gained attention to analyze multi-omics data for biomarker discovery. The development of standardized predictive frameworks using computational approaches can help bioinformatics analysts in suggesting the right treatment for a patient at the right time by focusing on the identified biomarkers.

This chapter explains the basic description of biomarkers, their types, and biomarkers for disease characterization. This is followed by a detailed explanation of multi-omics data and its integrative analysis. Also, various computational techniques for

the identification of biomarkers on multi-omics data are discussed. Next, the research motivation has been discussed. In addition, the objectives are drafted based on research motivation. At the end of this chapter, the thesis contribution and organization are provided.

1.1 Overview of Biomarkers

Biomarkers are the molecules like genes, DNA, proteins, and metabolites that signify whether a process going on in the body is normal or irregular, and it can be used as a symptom of any disease or disorder. Within the complex realm of medicine, where the relentless pursuit of knowledge, identification, and management of diseases persists, biomarkers arise as prominent indicators that assist doctors and researchers in navigating unexplored domains. The molecular markers, which are frequently concealed within our biological systems, possess the capacity to provide significant revelations regarding the overall health and state of being. Biomarkers can be detected in many bodily components such as blood, tissues, or genetic material and can play a pivotal role in early disease identification, and tailoring medical treatments, and therapies. Biomarkers are found in every disease, including cancer, multiple sclerosis, diabetes, and heart diseases [5]. In particular, biomarkers have been commonly used for precise diagnosis or prognosis with the release of precision medicine [6]. Different biomarkers have different roles in various diseases, for instance, the presence of mutations in the SAMHD1 gene is significantly linked to the development of malignancies such as T-cell lymphoma of the skin, chronic lymphatic leukemia, and colon cancer [7]. The SAMHD1 gene has recently been employed as both a biomarker and a therapeutic target in the context of acute myeloid leukemia [8]. Except for protein-coding genes, non-coding genes are emerging as potential biomarkers for illness detection, including circular ribonucleic acid (RNAs) (circRNAs). Recent research has revealed F-circEA, a fusion circRNA, as a newly discovered biomarker for non-small cell lung cancer (NSCLC) in liquid biopsy [9]. Figure 1.1 shows examples of some common cancer biomarkers. Biomarkers are systematically categorized as scientifically reliable indicators of pathological processes, typical biological activities, or reactions to therapeutic interventions [10]. Individuals who are involved in the study of diseases have a crucial function in differentiating between the status of a disease and normal status, as well as distinguishing between various stages of a disease. This serves as a significant connection between the processes of diagnosing a disease, predicting its outcome, and developing specific treatment strategies. A variety

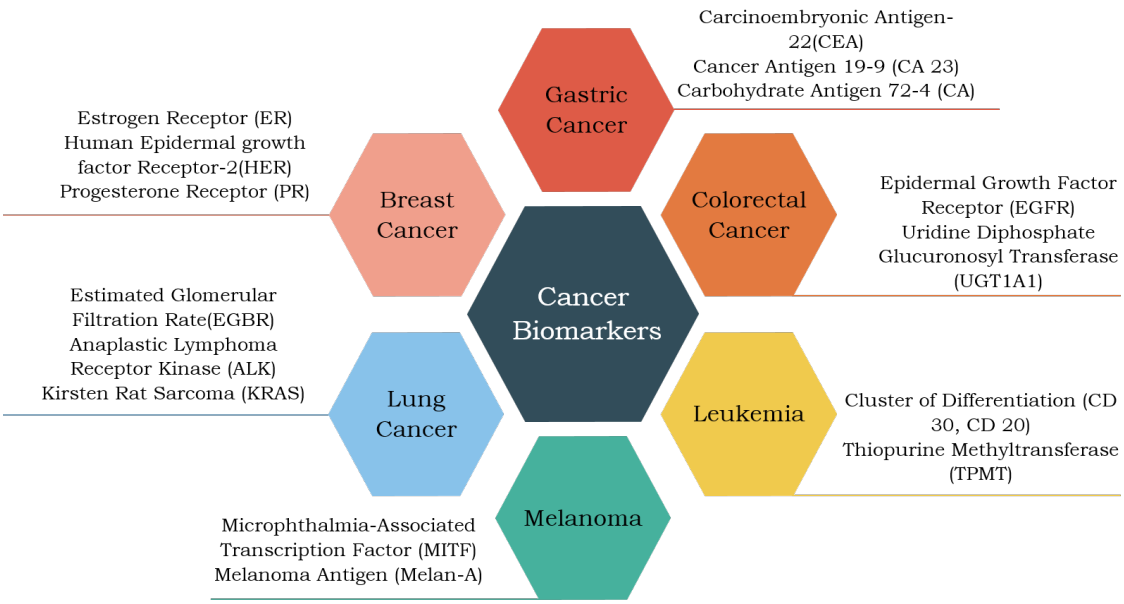


Figure 1.1: Examples of Cancer Biomarkers [1]

of biomarkers are used to accomplish this goal, each with specific properties and functions. There are seven types of biomarkers which are discussed in the following subsection.

1.1.1 Biomarkers Types

Biomarkers are classified into seven types including; 1. risk biomarkers (markers showing a risk of getting a disease); 2. diagnostic biomarkers (markers confirming the existence of disease); 3. prognostic biomarkers (markers predicting the recurrence of disease); 4. predictive biomarkers (marker used to detect the reaction of the patient to specific therapy); 5. monitoring biomarkers (markers that are monitored periodically); 6. safety biomarkers (markers used to measure the toxicity before and after treatment) and 7. response biomarkers (markers use to measure the response) [11]. The different categories of biomarkers are explained below along with Figure 1.2.

- **Risk Biomarkers:** A risk biomarker indicates a likelihood of developing a disease or health condition in individuals who currently do not exhibit the said ailment. An instance of such a biomarker is the genetic marker identifying BRCA1/2 mutations, specifically utilized to ascertain the increased susceptibility to future onset of breast cancer. Risk biomarkers hold significant value in clinical settings by directing preventive strategies. [11].

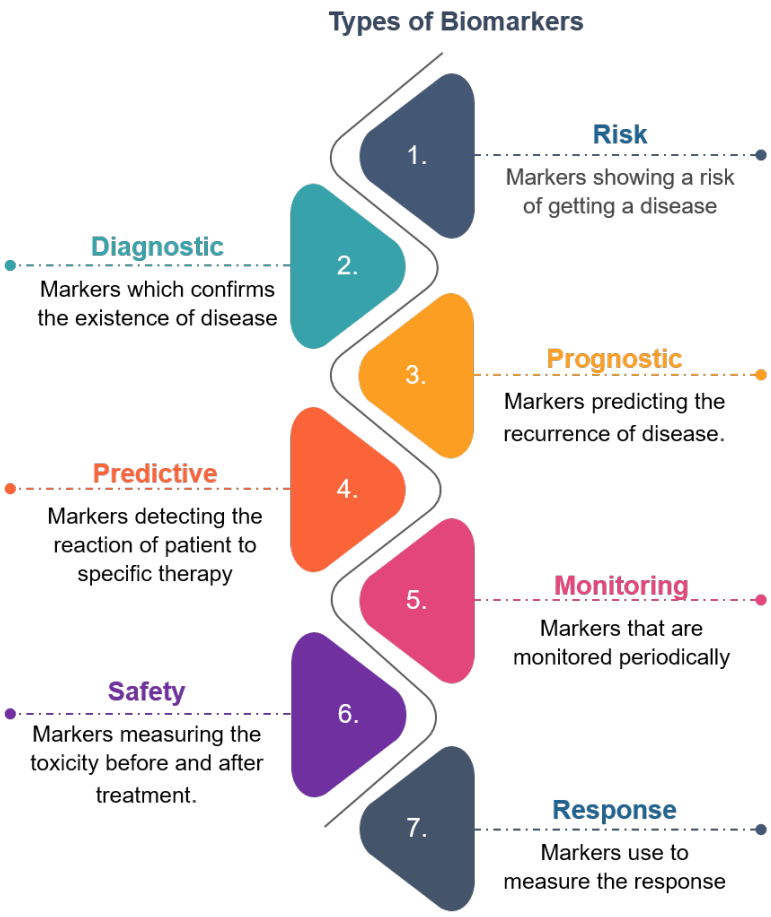


Figure 1.2: Types of Biomarkers [1]

- **Diagnostic Biomarkers:** A diagnostic biomarker refers to a molecular entity employed to anticipate or verify the presence of a specific disease or disorder or to categorize individuals exhibiting a particular subtype of the condition. Utilizing identified diagnostic biomarkers assists in steering treatment approaches. For instance, in patients diagnosed with diffuse large B-cell lymphoma, gene expression profiling serves as a diagnostic biomarker to segregate them into distinct subgroups based on unique tumor cell origin signatures [12].

- **Prognostic Biomarkers:** A prognostic biomarker anticipates the likelihood of a future clinical condition, disease recurrence, or relapse within a specified sample [13]. Biomarkers such as tumor size, the proportion of lymph nodes affected by tumor cells, and the presence of malignancy have been applied to forecast future prognoses. An example of a prognostic marker is elevated low-density lipoproteins (LDL) cholesterol levels, which

serve as an indicator for predicting the prognosis of individuals who have recently experienced a heart attack. [11].

- **Predictive Biomarkers:** A predictive biomarker serves as a diagnostic tool utilized for stratifying individuals who exhibit a higher probability of responding to a specific medication or chemical compound. This classification based on the biomarker may lead to symptomatic improvements, extended lifespan, or adverse reactions [11]. Within the realm of predictive biomarkers, there exists a gene prioritization challenge, wherein the identified gene signifies the potential onset of a specific disease by its association with known disease-related genes.

- **Monitoring Biomarkers:** A monitoring biomarker is assessed at regular intervals to track disease occurrences, including the onset of new symptoms, the progression of existing anomalies, or alterations in clinical outcomes or specific abnormalities. An instance of a monitoring biomarker is CA 125, used in ovarian cancer patients to gauge disease activity or effect both pre- and post-surgery. This biomarker aids in monitoring changes in the disease's status or progression over time [11].

- **Safety Biomarkers:** A safety biomarker is evaluated prior to or following exposure to a therapeutic drug or an environmental substance to ascertain the likelihood, incidence, and intensity of toxicity as an adverse outcome. An illustration of a safety biomarker is serum creatinine, utilized in patients receiving medications that may compromise kidney function. This biomarker aids in assessing and monitoring potential adverse effects on the kidneys resulting from medication usage [11].

- **Response Biomarkers:** A response biomarker signifies a patient's biological response to a pharmaceutical compound or an environmental substance. For instance, plasma microRNA serves as a response biomarker in Hodgkin lymphoma, indicating the biological reaction elicited in patients due to the disease or its treatment [11].

The biomarkers play a crucial role in the classification and characterization of diseases, facilitating the implementation of customized therapy strategies. These tools assist in the identification of disease risk, confirmation of diagnoses, prediction of prognosis, guidance in therapy selection, monitoring of disease development, assurance of patient safety, and evaluation of treatment outcomes. The biomarkers for disease categorization are discussed in the following section.

1.1.2 Biomarkers for disease characterization

Disease characterization encompasses various aspects and entails acquiring a full comprehension of the inherent characteristics and attributes of a particular disease. Biomarkers play a pivotal role in disease characterization by offering crucial insights into the molecular, genetic, and clinical dimensions of the disease. Biomarkers for diseases are characterized into the following categories and are shown in Figure 1.3.

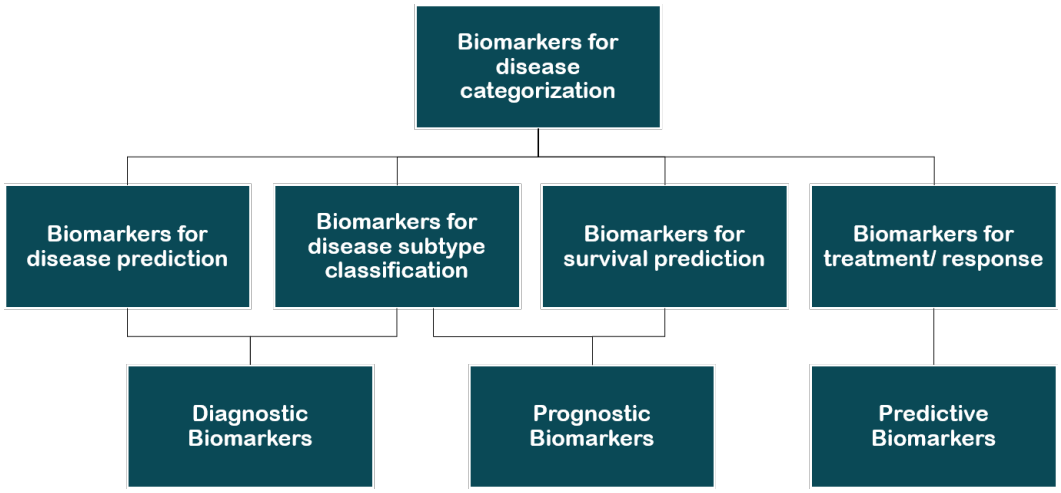


Figure 1.3: Biomarkers for disease categorization

1.1.2.1 Biomarkers for disease prediction

Prediction of chronic diseases like cancer, and Alzheimer's is important in the healthcare industry. Biomarkers for disease prediction are measured biological signs or traits that can help predict how likely it is that a person will get a certain disease in the future. These biomarkers can be identified through blood tests, imaging, genetic analysis, and other computational methods. Identifying and validating disease prediction biomarkers is important for early detection, prevention, and personalized treatment. Diagnostic biomarkers can be used to predict the disease as they are used as a tool to confirm the existence of a particular disease.

1.1.2.2 Biomarker Identification for disease subtype classification

The process of disease subtype classification entails the systematic categorization of a certain disease into discrete subgroups or subtypes, utilizing a range of criteria including clinical manifestations, molecular attributes, and responses to various

treatments. Biomarkers hold substantial significance in categorizing disease subtypes by facilitating the differentiation among various variants or subtypes of a particular illness. Biomarkers refer to particular biological signs or traits that are closely linked to distinct subtypes of diseases. The utilization of biomarkers in disease subtype classification holds significant value in the customization of treatment strategies, and the enhancement of patient outcomes. Diagnostic and prognostic biomarkers can be used to classify patients into subtypes.

1.1.2.3 Biomarker Identification for disease survival prediction

Biomarkers utilized for the prediction of survival encompass molecular, genetic, or clinical indicators that exhibit an association with the probability of a patient's survival or prognosis subsequent to a certain medical ailment or sickness. Biomarkers play a critical role in informing therapy choices, evaluating the extent of disease, and providing patients and healthcare practitioners with significant prognostic insights. The identification of prognostic biomarkers is closely associated with the ability to predict a patient's survival.

1.1.2.4 Biomarker Identification for treatment/response

The identification of biomarkers for the purpose of assessing therapy response is a fundamental component within the field of personalized medicine. Biomarkers play a crucial role in prognosticating an individual's response to a specific treatment, enabling healthcare professionals to customize medicines to achieve optimal efficacy while mitigating potential adverse effects. Predictive biomarkers fall under the treatment/ response category which can be identified to classify people who are reacting to a specific treatment.

1.2 Multi-omics Data

In recent years, multi-omics data has been used as molecular biomarkers using the integration of omics data types including genomic, transcriptomic, proteomic, metabolites, and interatomic for the prognosis and diagnosis of some specific diseases. The discovery of disease biomarkers with multi-omics data would not only aid in the stratification of various patient cohorts, but it would also include early diagnosis knowledge that may enhance patient care and possibly mitigate negative outcomes. There are different tools and techniques available for multi-omics data integration, which can be further used for biomarker identification, disease

diagnosis, and progression [14]. The types of omics data are discussed below and are shown in Figure 1.4.

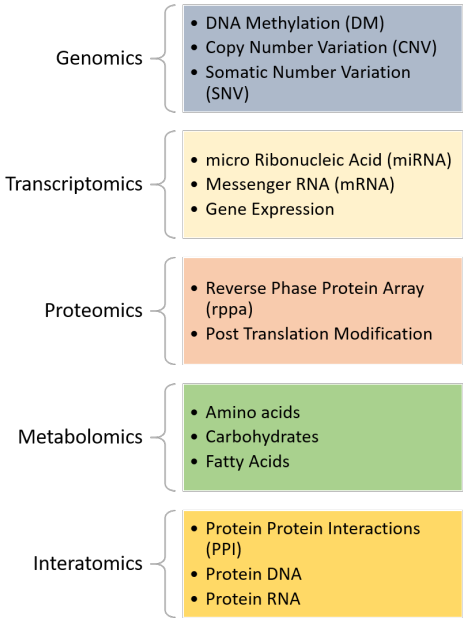


Figure 1.4: Types of Multi-omics

1.2.1 Types of omics data

- **Genomics:** The whole sequence of DNA in an organism, including all of its chromosomes, is referred to as a genome. Genomics seeks to characterize and quantify all of the genes of an organism, as well as their interrelationships and effects on the organism. The primary goal of genomics research in medicine is to find genetic variants that are linked to disease, therapeutic response, and patient prognosis [15].

- **Proteomics:** The entire universe of proteins in the cell is called proteome. Proteomics is a technique for detecting protein expression variations in response to a particular stimulus at a specific time, as well as determining protein structure networks at the tissue, organism, or cell level [16]. Proteomics is based on three technical key elements comprising a tool for fractionating complicated protein or peptide combinations, mass spectrometry (MS) for acquiring the data needed to classify specific proteins, and computational biology for analyzing and assembling the MS data. Classic unbiased approaches such as yeast two-hybrid assays and phage shows are used to identify protein interactions.

- **Transcriptomics:** A transcriptome is an organism's complete set of messenger RNA, including messenger RNA (mRNA), micro RNA (miRNA), and circRNA molecules. The sequence of mRNA transcripts generated in a specific cell or tissue type is referred to as the "transcriptome." RNA lies in between proteins and DNA and acts as a main function of DNA readouts [17]. To profile the transcripts or raw data, a technique called RNA-seq is used.

- **Metabolomics:** The metabolome contains a complete collection of small-molecule groups called metabolites, including carbohydrates, amino acids, sugars, and fatty acids. Similarly, like proteins, quantitative measurements of metabolites are performed using the MS technique. Metabolomics tasks are executed at different metabolite levels, and any relative distributions and disturbances signify the disease when occurs outside of the normal range [18].

- **Interatomics:** An interatomic is a multi-dimensional description of functional associations between molecules inside a cell or throughout the whole organism. A protein-protein interaction comes under this category of omics data [19].

The omics data types are integrated together for better biomarker identification using various techniques which are discussed in the following section.

1.2.1.1 Integrative analysis

Integrative analysis involves leveraging diverse data sources to gain a deeper understanding of complex systems. While many studies rely on single-source omics data, they often fall short in explaining the underlying causes of complex traits. Researchers have recognized that the analysis of a single data type lacks the explanatory power required to comprehend complex biological systems, as these systems are regulated by multiple levels [20]. To gain a more comprehensive understanding of disease biology, it is imperative to consider the behavior of molecules and the intricate interactions across various biological levels. In recent years, there has been a proliferation of multi-omics datasets, leading to the development of numerous computational models and applications designed to integrate data from multiple levels [21]. In the field of computational biology, integrative analysis plays a pivotal role. There are three different ways to integrate multi-omics data which are shown in Figure. 1.5 and are described as follows:

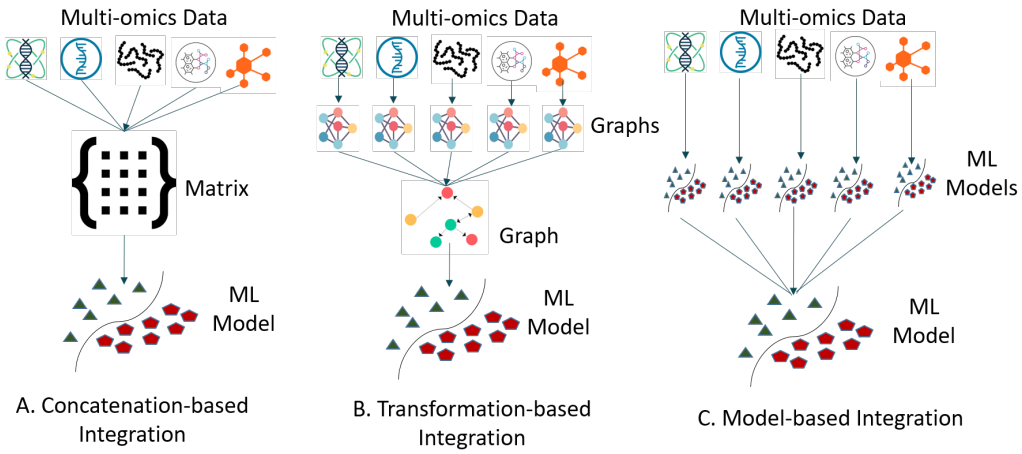


Figure 1.5: Type of multi-omics integration

- **Concatenation-based Integration:** In concatenation-based integration, multi-omics data are merged into a single combined matrix, which is subsequently employed for analysis. This method capitalizes on the compatibility of existing analytical techniques designed for single omics data, as they can be effectively applied to the consolidated matrix for comprehensive analysis.

- **Transformation-based Integration:** In transformation-based integration, the initial data types are first converted into matrices with a graph or kernel structure. These transformed representations of the data are then merged to create an integrated representation. This approach exhibits greater robustness compared to concatenation-based integration because it accommodates a wider range of data types, including categorical, continuous, or sequence data.

- **Model Based Integration:** In model-based integration, datasets are initially analyzed independently, and subsequently, the results are amalgamated to derive a unified outcome. This model-based approach exhibits exceptional flexibility, allowing for the application of distinct models tailored to various data types during the analysis process. In the realm of bioinformatics, model-based integration finds extensive application. This model-based approach can be broadly categorized into supervised and unsupervised methods, depending on the modeling techniques applied to individual data types. In the supervised category, diverse data types are utilized as training sets to construct multiple models. These models are then combined through techniques such as bagging or voting to yield a consolidated result. In the unsupervised category, clustering outcomes are derived from different

data types. Subsequently, these clustering results are aggregated, guided by specific optimization criteria, to facilitate integration.

The integration and analysis of multi-omics data are significantly influencing the comprehension of living organisms' biology. As multi-omics data becomes more affordable and accessible, it is likely to play a greater role in the identification of biomarkers, diagnosis and treatment of diseases, the development of new drugs, and the improvement of human health. Multi-omics patient datasets can be accessed through publicly available repositories such as The Cancer Genome Atlas (TCGA) [22], Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) [23], and Religious Orders Study and Rush Memory and Aging Project (ROSMAP) [24]. The repositories contribute to a huge production of multi-omics data which can be used by researchers for biomarker identification using computational intelligent techniques as discussed further.

1.3 Computational Intelligent techniques for Biomarker Identification

Computational intelligence is a group of techniques comprising ML [25], Deep Learning (DL), and statistical tests for building machine intelligence and decision-making. These techniques find extensive application in the identification of biomarkers for disease diagnosis and prognosis. These techniques serve to enhance the precision of medical prognoses, tailor treatments to individual patients by focusing on the identified markers, and lower the overall treatment expenditure per patient. In this regard, a spectrum of computational approaches, including statistical methods, ML, and DL has garnered significant attention. These methods are instrumental in the preprocessing, normalization, feature or biomarker identification, integration, and analysis of multi-omics data, which is pivotal in advancing our understanding of disease mechanisms and facilitating more effective clinical decision-making. Further, elaboration on these techniques is provided in the subsequent sections.

1.3.1 Statistical Methods

In the field of multi-omics data analysis for biomarker identification, two primary statistical approaches are prominent: descriptive methods and inferential methods. Both methodologies hold significant importance in research endeavors. Descriptive statistics encompass fundamental metrics such as the mean, median, and standard

deviation, which provide a concise summary of data characteristics. On the other hand, inferential methods, including Student's t-test and F-test, are employed for drawing inferences and making statistical comparisons. Descriptive statistics offer insights into data distributions and central tendencies, while inferential methods are used for hypothesis testing and making inferences about populations based on sample data. It's worth noting that statistical methods that involve comparing means are categorized as parametric, while others fall into the non-parametric category [26]. Parametric tests include various forms of t-tests and F-tests. For instance, when comparing the means of two groups, one would typically employ the t-test, which can take different forms like the one-sample t-test, independent samples t-test, or paired samples t-test. In scenarios involving the comparison of means across three or more groups, the F-test, often referred to as one-way Analysis of Variance (ANOVA) or repeated measures ANOVA, serves as an extension of the t-test.

Furthermore, parametric methods, such as the Pearson Correlation Coefficient (PCC) and linear regression, are used to establish relationships and associations between variables, particularly in the context of quantitative data analysis. For the identification of prognostic markers required for survival prediction, a specialized branch of statistics often applied in medical research and event-time analysis, distinct statistical methods come into play. These include the Kaplan-Meier estimator for survival probability, the Log-rank test for comparing survival curves, and the Cox regression model for assessing the influence of multiple variables on survival outcomes. These methodologies are tailored to address time-to-event data, where the primary goal is to analyze the duration until an event of interest occurs.

1.3.2 Machine Learning and Deep Learning Algorithms

ML consists of various algorithms required for analysis, which leads to an effective identification of biomarkers for disease prognosis and diagnosis. ML analytics is used in healthcare to deal with complex multi-omics data and its integration which is required for biomarker identification. The pipeline for ML analytics for multi-omics biomarker identification for disease diagnosis and prognosis is given in Figure 1.6. It consists of data preprocessing, feature/ biomarker identification, and modeling [17] which are discussed as follows:

- **Data Preprocessing:** To handle the data effectively, data cleaning involving the removal and imputation of remaining missing values is done. Further,

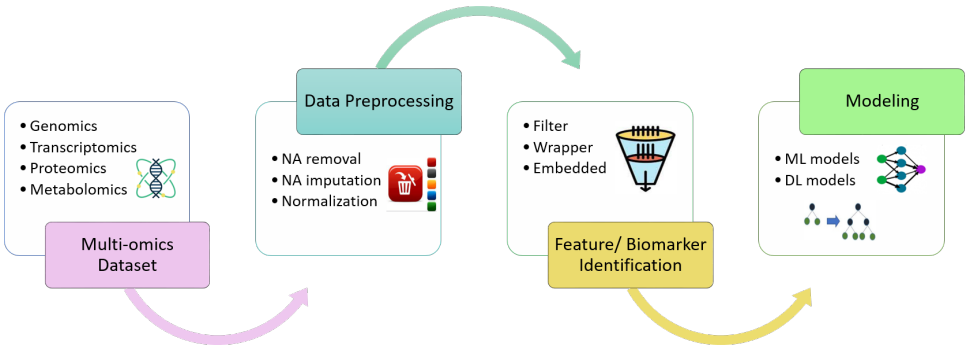


Figure 1.6: Biomarker identification and predictive analysis in multi-omics data

normalization is done in multi-omics data to maintain consistency.

- **Feature/ Biomarker Identification:** The multi-omics dataset is a high dimensional dataset. Therefore, feature selection approaches comprising filter, wrapper, and embedded techniques are applied to select only the relevant features or biomarkers. Various feature selection and extraction algorithms are there comprising principal component analysis (PCA), maximum relevance minimum redundancy (mRMR), Mutual Information (MI), and many more. Similarly, swarm intelligence is also used to find the optimal features/ biomarkers. Some of the common swarm intelligence techniques are particle swarm optimization (PSO), ant colony optimization (ACO), and cat swarm optimization (CSO). The identified biomarkers are validated using biological interpretation tools including DAVID functional analysis and survival analysis tests.

- **Modeling:** A model is built from training data with supervised or unsupervised learning, and then various criteria are used to evaluate the performance of the model. Supervised learning encompasses the process of training models on multi-omics datasets that are accompanied by well-defined labels. This involves utilizing data where a subset already includes the accurate corresponding outcomes. The supervised learning algorithm then analyzes test data and generates accurate results. Supervised learning can be used for both classification and regression problems, depending on whether the outcome variable is categorical (classification) or a real value (regression). Various supervised learning algorithms include Support Vector Machines (SVM), Linear Regression, Random Forests (RF), Adaboost, K-Nearest Neighbor (K-NN), Naïve Bayes, and Decision Trees that can be used for disease diagnosis and prognosis. In unsupervised Learning, the machine works with data that lacks categorization or labeling. It organizes unlabeled data into clus-

ters or groups based on similarities, variations, and differences without prior knowledge of the data. Common unsupervised learning algorithms include hierarchical clustering and K-means clustering [27].

On the other hand, DL is a subset of ML inspired by the structural organization of the human brain. DL analyzes data by employing a hierarchical system of algorithms known as neural networks. These networks are designed to mimic the structure of the human brain and can be trained to recognize patterns and interpret diverse types of data, similar to human cognition. Deep neural networks (DNN) [28] are a type of DL architecture that operates on similar principles to the human brain. They can be applied to various tasks such as classification, clustering, and regression. Neural networks can also be employed to organize or filter unlabeled data based on the similarity between samples. DL encompasses several algorithms, including Convolutional Neural Networks (CNN), and Recurrent Neural Networks (RNN), among others, which are highly valuable in disease classification and prognosis [29].

ML and DL can be used to analyze multi-omics data for the identification of biomarkers required for disease diagnosis and prognosis. By combining ML with multi-omics data, researchers can develop new treatments for diseases, improve patient outcomes, and personalize medicine by focusing on the identified markers.

1.4 Research Problem and Motivation

In the domain of healthcare research, substantial endeavors have been directed toward addressing the complexities of disease treatment and prevention. Particularly, the advent of precision medicine has propelled the widespread utilization of biomarkers to achieve precise disease diagnosis and prognosis. Within this landscape, the abundance of biomedical data presents both significant opportunities and challenges. Biomedical data is high-dimensional and takes various forms, encompassing records, images, and omics data. This multifaceted data landscape garners substantial interest from both medical researchers and data scientists. Moreover, prioritizing diagnostic and prognostic biomarkers allows researchers to strategically address critical clinical needs. Diagnostic biomarkers are crucial for detecting diseases early, while prognostic biomarkers provide insights into disease progression and patient outcomes. These biomarkers not only aid in understanding the disease landscape but also directly impact clinical practice. Their immediate

translation into real-world applications can significantly benefit patient care. On the other hand, predictive, response, risk, monitoring, and safety biomarkers may require extensive studies, including clinical trials, to validate their utility effectively. Focusing on diagnostic and prognostic biomarkers streamlines research efforts and facilitates simpler validation pathways. This strategic approach ensures the efficient allocation of research resources while maximizing the potential impact on clinical practice and patient outcomes. In the context of intricate diseases like cancer, and Alzheimer's, various omics datasets, including genomic, proteomic, transcriptomic, and metabolomic data, can be amassed for a single individual. The integration of these omics datasets holds the potential to enhance predictive capabilities, unveiling insights into otherwise hidden data patterns and disease-related concerns [30].

However, despite the dimensionality of multi-omics data, identifying biomarkers that can accurately diagnose or predict diseases remains a formidable challenge. In recent years, various models and techniques have emerged to facilitate biomarker identification and enhance the understanding of biological systems. ML, DL, and statistical methods, in particular, play pivotal roles in the computational analysis of multi-omics data for biomarker identification. Utilizing ML in conjunction with multi-omics datasets can significantly augment computational mechanisms for biomarker discovery and enhance prediction accuracy. This synergy enables the development of more precise techniques tailored for healthcare applications. Advanced ML and DL techniques prove to be particularly advantageous when dealing with the intricate, extensive, and diverse nature of multi-omics datasets for biomarker identification required for accurate disease diagnosis and prognosis [29].

The raw multi-omics data, in its unprocessed state, lacks utility for healthcare researchers. To render this data suitable for accurate analysis, preprocessing, normalization, and feature/ biomarker identification methods for disease diagnosis, prediction, and survival prediction are imperative. In the realm of healthcare data analysis, there is a strong endorsement for the utilization of ML and DL techniques due to their efficacy and versatility [31]. This work aims to provide a framework that incorporates pre-processing, feature/ biomarker identification, and the development of learning models for diagnosis and prognosis using the identified markers. The present work is an attempt to enhance biomarker identification for survival prediction, disease prediction, and disease subtype classification using ML and DL approaches.

1.5 Objectives

The objectives of this research work are:

- To study and understand various existing tools and techniques used for the identification of biomarkers using multi-omics data.
- To propose, design, and develop an efficient framework for biomarker identification using multi-omics data with the help of ML approaches.
- To test and validate the proposed framework for predictive analysis like survivability, disease prediction, etc.

1.6 Thesis Contribution

The contributions of this research study in the field of Biomarker Identification are listed below:

- A critical review of ML, and DL for biomarker identification using single and multi-omics data analysis for disease prediction, disease survival prediction, disease subtype classification, and treatment/response has been performed. The tools required for biomarker identification in single and multi-omics data are discussed. This would guide researchers to understand the use of computationally intelligent approaches for efficient biomarker identification.
- A framework is proposed for biomarker identification in multi-omics for disease diagnosis and prognosis. Based on this framework, three approaches comprising BioSurv, iMVAN, and HBS-STACK are developed for biomarker identification in multi-omics required for survival prediction, subtype classification, and disease prediction.
- A BioSurv approach is developed for biomarker identification in multi-omics data for survival prediction. Two feature extraction techniques comprising statistical methods and A Random Spatial Local Best Cat Swarm Optimization (RSLBCSO) for biomarker identification are proposed to extract the features. The extracted features from each type are integrated using a concatenation-based approach. Bayesian Optimized Deep Neural Network model is used for survival analysis of Breast Carcinoma (BRCA) and Lung Adenocarcinoma (LUAD). The statistical analysis of BioSurv compared to existing base learning models is performed to show the effectiveness of the proposed work.

- An iMVAN approach for biomarker identification for different subtypes of disease is developed. A multi-modal variational autoencoder (MVAE) is developed to identify the markers which are validated using KEGG analysis and survival analysis. The extracted features/ biomarkers are integrated using similarity network fusion (SNF). The simplified graph convolutional network (SGC) is used as a learning model for disease subtype classification. The performance of iMVAN is evaluated on BRCA, KIPAN, and CESC using multi-omics datasets.

- A HBS-STACK approach is developed for hierarchical biomarker selection and disease prediction in multi-omics datasets. Three-stage feature/ biomarker selection is developed including CpG site aggregation, statistical tests, and light gradient boosting machine recursive feature elimination (LGBMRFE) in multi-omics data of BRCA patients. The top-ranked features have been selected as identified biomarkers which are validated using DAVID analysis. The features along with identified markers are integrated using a concatenation-based approach which is modeled using a developed Stacked model of RF, Naive Bayes (NB), Gradient Boosting Machine (GBM), and DNN. The performance of HBS-STACK is validated on multiple diseases including Kidney Renal Carcinoma (KIRC) and Alzheimer's disease.

1.7 Thesis Organization

After the Introduction to this research work in Chapter 1, the rest of the chapters are organized as follows:

Chapter 2: Literature Survey
In this chapter, the work done by various researchers related to biomarker identification using omics and multi-omics data is explored. It includes a survey of ML and DL techniques for the identification of diagnostic, prognostic, predictive, and other biomarkers for disease prediction, survival prediction, and treatment/ response predictions. The survey is performed for both single and multi-omics datasets. In addition, the survey of existing tools for biomarker identification using multi-omics data is discussed in detail. Finally, based on the studied literature, the challenges related to biomarker identification in multi-omics datasets are discussed in the chapter. Chapter 2 has been derived from:

- Arwinder Dhillon, Ashima Singh, Vinod Kumar Bhalla, " A Systematic

Review on biomarker identification for cancer diagnosis and prognosis in multi-omics: from computational needs to machine learning and deep learning." Archives of Computational Methods in Engineering, vol. 30, no. 2, pp. 917-949, 2023, Springer. [Impact Factor: 9.7]

Chapter 3: Proposed Framework

In this chapter, the framework for biomarker identification in multi-omics data for disease diagnosis and prognosis is proposed. The proposed framework involves different phases, i.e., data acquisition, data preprocessing, feature/ biomarker identification, biological interpretation of identified markers, development of models, and performance evaluation. The data acquisition describes the repositories for multi-omics data collection. Further, the hardware and software requirements required for developing a framework are discussed. By adopting the proposed framework, three approaches comprising BioSurv, iMVAN, and HBS-STACK have been developed for biomarker identification in multi-omics data for survival prediction, subtype classification, and disease prediction, respectively.

Chapter 4: BioSurv: Proposed Biomarker Identification for Survival Analysis

In this chapter, the proposed BioSurv approach for biomarker identification in multi-omics for survival analysis is discussed. This chapter describes the statistical test comprising Fold Change and False Discovery Rate and swarm intelligence technique called Random Spatial Local Best Cat Swarm Optimization for biomarker identification and Bayesian Optimized Deep Neural Network for survival prediction. The experiments are performed using TCGA BRCA and TCGA LUAD multi-omics datasets. The BioSurv is validated using the BRCA dataset from the METABRICS portal. The performance of the BioSurv framework is evaluated for accuracy and area under curve (AUC) parameters. The content of the work presented in Chapter 4 has been taken from:

- Arwinder Dhillon, Ashima Singh, Vinod Kumar Bhalla, "Biomarker identification and cancer survival prediction using random spatial local best cat swarm and Bayesian optimized DNN." Applied Soft Computing, vol. 146, pp. 110649, 2023, Elsevier. [Impact Factor: 8.263]

Chapter 5: iMVAN: Proposed Integrative Multimodal Variational Autoencoder based biomarker identification for disease subtype classification

In this chapter, the proposed iMVAN for biomarker identification for disease subtypes is discussed. This chapter describes the multimodal variational autoencoder

for biomarker identification and simplified graph convolutional networks for subtype classification. The multi-omics dataset is used which is fused using similarity networks fusion. The experiments are performed on Breast Cancer subtypes, and validated on KIPAN and CESC subtypes. The content of the work presented in Chapter 5 has been taken from:

- Arwinder Dhillon, Ashima Singh, Vinod Kumar Bhalla, "iMVAN: integrative multimodal variational autoencoder and network fusion for biomarker identification and cancer subtype classification." Applied Intelligence, vol. 53, no. 22, pp. 1-18, 2023, Springer. [Impact Factor: 5.019]

Chapter 6: HBS-STACK: Proposed Hierarchical Biomarker Selection and Stacked ensemble approach for disease prediction
In this chapter, the proposed HBS-STACK approach for biomarker identification in multi-omics data for disease prediction is discussed. This chapter describes the hierarchical biomarker selection in detail. The extracted features are validated using DAVID analysis. The stacked ensemble model is developed on integrated features for disease prediction. The experiments have been conducted on TCGA BRCA datasets and validated on TCGA KIRC and Alzheimer's disease. The content of the work presented in Chapter 6 has been taken from:

- Arwinder Dhillon, Ashima Singh, Vinod Kumar Bhalla, "HBS-STACK: Hierarchical Biomarker Selection and Stacked Ensemble model for Biomarker Identification and Cancer Prediction in Multi-Omics " Neural Computing and Applications, Springer. [Accepted]

Chapter 7: Conclusions and Future Scope In this chapter, the thesis is concluded by providing conclusions of the research work done and by suggesting possible future directions.

Chapter 2
Literature Survey

In this chapter, the in-depth survey of existing techniques for biomarker identification using omics and multi-omics data with the help of Machine Learning (ML) and Deep Learning (DL) techniques are provided. The chapter also offered a thorough survey of tools developed using programming languages for biomarker identification using multi-omics datasets. The survey of existing works aids in identifying the challenges in biomarker identification. The main focus of this chapter is to discover the opportunities of computational intelligent techniques for biomarker identification using omics and multi-omics datasets.

This chapter begins with Section 2.1 having a detailed description of biomarker identification using ML and DL approaches. The work done on biomarkers for disease prediction, survival prediction, treatment/ response, and other biomarkers comprising monitoring, risk, response, and safe biomarkers is given. In Section 2.2, the existing tools developed for biomarker identification using multi-omics datasets are reviewed. In Section 2.3, the challenges identified from existing literature for biomarker identification are discussed. The chapter is concluded in Section 5.4 summarizing the key findings and insights obtained from the literature review.

2.1 Biomarker Identification using Machine Learning and Deep Learning Techniques

Biomarkers have become increasingly popular for accurate diagnosis and prognosis of diseases in healthcare. While the availability of multi-omics data makes it a lot easier to assess biomarkers for diagnosis and prognosis of diseases, the identification of biomarkers that can accurately recognize or detect diseases in the presence of tens of millions of genes and billions of variants is still a challenging task. The extensive use of Artificial Intelligence (AI), including Machine Learning (ML) and Deep Learning (DL), have gained popularity due to their ability to extract key features from complex datasets. Many researchers have worked on the ML and DL approaches for the identification of biomarkers using single and multi-omics

datasets required for disease prediction, survival prediction, and treatment/ response prediction aiming to enhance medical decision-making and patient care which are discussed further.

2.1.1 Diagnostic Biomarkers for Disease Prediction

Diagnostic markers are the markers that are used to confirm the presence of disease and to identify the markers in different sub-types of disease. The biomarkers can be identified using single-omics and multi-omics datasets with the help of ML, DL, and statistical approaches. The work done on diagnostic biomarker identification using single omics is given in the following section.

2.1.1.1 Biomarker Identification in single omics

Omics data include genomics (DNA Methylation (DM), Copy Number Variation (CNV)), transcriptomics (micro ribonucleic acid (miRNA), messenger RNA (mRNA)), proteomics (Reverse Phase Protein Array (rppa)), metabolomics and Interatomic dataset. Researchers have worked on biomarker identification using a single type of omics data mentioned above, for example, Hanieh et al. [32] utilized ML techniques comprising Support Vector Machine (SVM), Random Forest (RF), and k nearest neighbor (KNN) for the identification of diagnostic biomarkers using miRNA data from gastric cancer. The experiment was performed and the four markers comprising MIR21, MIR133a, MIR129c, and MIR29c as diagnostic markers with 87% Area Under Curve (AUC) value. These markers were validated using Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis and it has been found that the markers play a significant role in Wnt signaling pathways.

Bhaoshan et al. [33] proposed XGBoost for the biomarker identification and cancer prediction using the DM dataset from The Cancer Genome Atlas (TCGA) portal of nine cancer types, including Kidney Renal Papillary Cell (KIRP) and Head and Neck Squamous Cell Carcinoma (HNSCC). The dataset was extracted from TCGA and Gene Expression Omnibus (GEO), and the findings showed that XGBoost accurately identified 151 and 153 biomarkers of HNSC and KIRP respectively.

Ting Jin et al. [34] developed a model using a semi-restricted Boltzmann machine named ECMarker to predict gene expression (gene expr.) biomarkers for the different stages of diseases like the early prediction of cancer. Gene-expression of non-small cell lung cancer (NSCLC) patients has been taken, and the ECMarker model was applied, which achieved an accuracy of 85%. The 9 genes were identified

including KRAS, ALK, BRAF, PIK3CA, NRAS, AKTI, RET, EGFR, and ROS1 as the diagnostic biomarkers. It was also used to prioritize biomarkers genes that were responsible for the early prediction of lung cancer.

Ying Xie et al. [35] used ML methods for the identification of diagnostic biomarkers using the metabolomics dataset of Lung Adenocarcinoma (LUAD). A sample of 110 patients was collected from the Hubei Taihe Hospital (HTH) and passed to principal component analysis (PCA) to select the metabolites. Then the Statistical analysis was performed which considered only those metabolites having a value less than 0.05. Further, the selected metabolites were passed to ML algorithms comprising KNN, SVM, RF, Naive Bayes (NB), Neural Network (NN), and Adaboost and it was found that NN performed best with an accuracy, specificity, sensitivity, and AUC value of 99%. It was also found that 10 metabolomics biomarkers including L-Kynurenine, Proline, Spermidine, Palmitoyl-l-carnitine, Amino-hippuric acid, Phenylalanine, Taurine, L-Valine, o-Tyr, Carnitine plays a critical function in tumor diagnosis.

Fariha Muazzam [36] used Deep Neural Network (DNN) for the identification of diagnostic biomarkers using the RNA-Seq dataset of Breast Carcinoma (BRCA) patients. The size of the dataset was reduced using Kernel Principal component analysis (KPCA) and PCA techniques. Then it was passed to the Stacked DAE for biomarker identification. Further, the identified biomarkers were passed to DNN for the classification of cancer patients. Pathway analysis was performed which identified three genes (PIK3C2G, PCDHB8, WNT10A) to be involved in multiple cancers.

Indu Khattri et al. [37] proposed an ML algorithm for the identification of diagnostic markers using mRNA data of Pancreatic Adenocarcinoma (PDAC) patients. The dataset was collected from Array Express and GEO and passed to pre-processing and statistical analysis tests for the identification of differentially expressed genes (DEGs). The identified genes were then passed to SVM for classification of cancer patients. The experiment was performed and the tested result proved that the proposed framework successfully identified 9 genes comprising IFI27, CTSD, ITGB5, EFNA4, PLBD1, GGH, HTATIP2, CTSA, and IL1R2 with 97% accuracy.

Xin Zhao et al. [38] used RF for the identification of novel diagnostic biomarkers in hepatocellular carcinoma (HCC). miRNA genes of 373 patients were downloaded from the TCGA portal and passed to the RF model for biomarker identification. The experiment was validated on the GSE63046 dataset extracted from GEO. From the results, it was found that the proposed method identified five diag-

nostic biomarkers comprising hsa-miR-224-5p, hsa-miR-10b-5p, hsa-miR-10b-3p, hsa-miR-182-5p and hsa-miR-183-5p.

Oneeb Rehman et al. [39] proposed ML algorithms to validate the importance of miRNA as BRCA biomarkers. The feature selection techniques comprising Least Absolute Shrinkage and Selection Operator (LASSO), Chi-Squared (CHI2), and, Information Gain (IG) were used to rank the features according to their importance. The training of these samples was performed using ML algorithms comprising RF and SVM, and it was investigated that top-ranked miRNAs as biomarkers can be beneficial in predicting BRCA. 11 diagnostic markers comprising hsa-let-7c, hsa-mir-10b, hsa-mir-145, hsa-let-7a-3, hsa-mir-125b-1, hsa-let-7d, hsa-mir-125b-2, hsa-mir-10b, hsa-mir-33, hsa-mir-101-1, and hsa-mir-335 respectively have been identified using the proposed work.

Iman et al. [40] introduced ML algorithms to identify the transcripts for prediction and for guiding the treatment related to prostate cancer progression. Transcripts dataset have been taken from the National Center for Biotechnology Information (NCBI) which was then passed to minimum redundancy maximum relevance (mRMR) to obtain the DEGs. Five ML algorithms including SVM with linear kernel, SVM with Radial Basis Function (RBF) kernel, RF, Decision Tree (DT), and NB were used for modeling, and it has been found that SVM with linear kernel outperforms with 90% accuracy and identified 10 diagnostic biomarkers.

Biao Liu et al. [41] presented two multi-layer feed-forward NN based on DL for the identification of markers by using the DM dataset. To identify genes, a t-statistics test was used and further passed to the LASSO and RF algorithm. 12 CpG markers and 13 promotor markers were identified which were further passed to the DL model that achieved a sensitivity of 92% for CpG markers and 89% for promotor markers.

Reka Toth et al. [42] presented an RF-based classification model for the detection of biomarkers for prostate cancer. DM dataset was downloaded from TCGA and passed to the preprocessing and feature extraction stage to extract the relevant features. It was then given to the RF to identify the biomarker for prostate cancer. The results were evaluated, and it was proved that the RF-based modeling identified the top 30 methylation genes with an AUC value of 77%.

A detailed conceptual survey of the work done by the researchers on diagnostic biomarker identification for disease prediction in single omics is given in Table 2.1.

Table 2.1: Diagnostic Biomarker identification for disease prediction in single omics data

Year	**Algorithm**	**Dataset Used**	**Result**	**Future Studies**
2023 [32]	SVM, RF, KNN	DM data from TCGA (Gastric cancer).	4 miRNAs were identified by SVM as diagnostic biomarkers.	The analysis of miRNAs can be conducted in multiple cohorts using laboratory-based methodologies.
2022 [33]	XGBoost	DM data from TCGA (HNSC, KIRP)	151 and 153 markers for HNSC and KIRP were identified accurately.	Other multi-omics datasets will be considered for better results.
2021 [34]	Semi Restricted Boltzmann Machine	Gene expr. data from TCGA (NSCLC)	ECMarker identified 9 markers with 85% accuracy.	Multi-omics analysis can be carried out to predict biomarkers for disease prediction.
2020 [35]	Statistical Methods, PCA, SVM, NN, RF, Adaboost, KNN and NB	Metabolomics data from HTH (LUAD)	NN outperformed and identified 10 metabolites as diagnostic biomarkers with 99% accuracy.	Clinical information like age, smoking, and past medical history, can be included for better performance.
2020 [36]	DNN, KPCA, PCA, Stacked De-noising autoencoders.	RNA-seq dataset from TCGA (BRCA)	DNN with stacked auto-encoder identified 3 marker with 95% accuracy.	This study can be applied to datasets of larger size and in multi-omics datasets for better performance.
2020 [37]	SVM, Statistical analysis test	mRNA data from ARRAY, GEO (PDAC).	ML methods identified 9 markers with 97% accuracy.	The proposed framework performed well in blood biomarkers and will be ideal for clinical trials.
2020 [38]	RF	miRNA data from TCGA, GEO (HCC).	RF identified 5 diagnostic markers with 89% AUC.	Large sample size can be used in the future for validation.
2019 [39]	RF, SVM, CHI2, IG, LASSO	miRNA data from TCGA (BRCA)	RF successfully identified 11 diagnostic markers with 99%. accuracy	The proposed ML algorithms can be used for the biomarker identification for other diseases.
2019 [40]	mRMR, SVM with linear and RBF kernel, RF, DT, and NB	mRNA dataset from NCBI	11 diagnostic biomarkers were identified with 80% accuracy.	Wet-lab experiments and clinical assays are required to confirm the existence and progression of the identified biomarkers.
2019 [41]	NN, LASSO, RF, t-statistics test	DM datasets from TCGA and GEO	12 CpG markers and 13 promotor markers were identified accurately.	Future studies involve more statistical, ML, and DL algorithms for biomarker identification.
2019 [42]	RF	DM data from TCGA (Prostate Cancer)	RF outperformed and identified top 30 markers with 77% AUC.	Additional biomarkers will be discovered using whole-genome bisulfite sequencing (WGBS) methods in the human genome.

2.1.1.2 Biomarker Identification in multi-omics

Multi-omics dataset is the integration of different types of omics datasets which play an important role in biomarker identification required for disease prediction, and subtype classification. Various authors have worked on biomarker identification using multi-omics datasets, for example, Nivedhitha et al. [43] presented an ensemble of filter approaches comprising ReliefF, CHI2, IG, and mutual information for biomarker identification from mRNA and DM dataset of Alzheimer patients extracted from GEO database. The extracted features from the ensemble approach were integrated using the Jackard index and passed to the deep belief network (DBN) for classification. The experiment was performed and the presented work accurately identified 35 markers as diagnostic markers with 82% accuracy.

Ping et al. [44] proposed a multi-omics attention DNN (MOADLN) for the identification of diagnostic biomarkers and disease prediction. Multi-omics data, including mRNA, miRNA, and DM, were considered from BRCA, Kindey Renal Papillary Cell (KIRP), and Alzheimer patients. Samples were collected and integrated through MODALN. The experiment was performed, and the proposed MOADLN identified the top 20 genes as diagnostic markers from each disease with an accuracy of 83%.

Yanyu et al. [45] presented a deep random forest feature selection (RDFS) method for biomarker identification using CNV and mRNA datasets of gastric cancer patients. The extracted features from RF were passed to DNN for training and testing. The experiment was performed and it was found that the RDFS method accurately identified the top 20 markers from CNV and mRNA, respectively with 98% accuracy.

Min-Koo et al. [46] introduced a novel artificial intelligence method that utilizes a graph convolutional network (GCN) for the identification of diagnostic biomarkers using mRNA, and DM data of NSCLC patients. First, those genes were extracted whose p-value < 0.05 and false discovery rate (FDR) > 0.5. The extracted features were integrated and passed to GCN for model training. The experiment was conducted and the findings revealed that GCN effectively returns the top 15 features from both mRNA and DM as diagnostic biomarkers with 93.7% F1-score value.

Jie Feng et al. [47] proposed joint kernel learning on a multi-omics dataset for the identification of diagnostic genes from LUAD and Liver hepatocellular carcinoma (LIHC). A sample of isoform expression profile, gene expression, and DM data along with their survival information was collected from the Genome Data

Analysis Center (GDAC) and passed to the KPCA method for feature extraction. Then the extracted features were converted to kernel metrics using the Gaussian Kernel function which was then passed to the clustering algorithm. The clustering algorithms divide the features into clusters for different cancers. Further, the performance of clustering was validated using Rand Index (RI) and adjusted RI (ARI) methods. The proposed framework identified GMPS, EPHA10, C10orf54, and MAGEA6 for LUAD and FAU, DEPDC6, VPS24, LOC100133469, RCBTB2, and SLC35B4 for LIHC.

Yong Liu et al. [48] identify the diagnostic markers from skin cancer patients using the SVM with recursive feature elimination (SVM-RFE) approach of ML from gene expr. and DM datasets. The dataset was downloaded from the GEO database and then passed to SVM-RFE to rank the features and identify the DEGs and differentially methylated genes (DMGs). Further, the training was done using RF and LASSO, which identifies the ten diagnostic genes comprising PPARG, LEP, PPARGC1A, IRS1, EBF1, PLIN1, FBXO32, SDC2, PLIN1, ZNF423, and MYOCD with an AUC of 95%.

Ze Zhang et al. [49] proposed DNN to identify the markers from multi-omics data (Gene Expr. and DM) of gastric cancer patients. A sample of patients was taken on which the first p-value and FDR test were applied. After that, mRMR was applied to rank the features. The extracted features were then integrated and passed to DNN for training. The experiment identified eight genes, including RORC, PGC, GPRC5C, KCNE2, PDGFD, KCNE2, PSCA, PPAP2B, and IFITM2, with 98% accuracy.

Ming Zhang et al. [50] used Cox survival analysis and BayesNet model for the identification of diagnostic biomarkers from BRCA patients using DM and gene expression datasets. The dataset was collected and passed to statistical tests for the identification of potential biomarkers. These markers were then passed to the BayesNet model to classify the patients as healthy candidates. Further, the candidate markers were passed to the Cox regression model to calculate the survival value which identified seven differentially methylated sites (DMSs) comprising TUFT1, TRERF1, CCND1, SRGAP1, PER1, ENPP2, and PER1 as diagnostic and prognostic makers.

Meijie Zhang et al. [51] used the RF feature selection method to identify the diagnostic biomarkers using lncRNAs, mRNAs, and miRNAs dataset of osteoporosis patients. A network was created of 105 nodes including 8 miRNAs, 24 mRNAs, 73 lncRNAs, and 515 edges. This network was passed to functional analysis which showed the involvement of DysCeNet in osteoporosis. Further, RF was

used which identified 25 features as diagnostic biomarkers. The identified genes were also validated using the leave-one-out cross-validation (LOOCV) method to show the effectiveness of the proposed work.

Pengfei Liu et al. [52] identifies diagnostic and prognostic markers using ML algorithms including RF, and LASSO-Cox from epigenetic, transcriptomic, and metabolomics datasets. 9398 CPGs and 2478 genes were collected and passed to RF which selected 134 CpGs and 54 genes from the integrated dataset. These were then passed to LASSO for the identification of diagnostic markers. Moreover, prognosis analysis was also performed using univariate Cox and LASSO Cox methods. The proposed framework identified 5 diagnostic and 8 prognostic markers respectively.

Prasoon Joshi et al. [53] proposed a DNN named Sparse Crossmodel Super layered Neural Network (SCR-SNN) for the integration of RNA sequencing and DM data and for the biomarker identification in LUAD patients. The dataset was passed to PCA for data filtering. Further Biomarker selection was performed using SCR-SNN which includes LR with L1 penalty, L1-regularized NN, and L1-regularized cross-modular NN. The proposed method identified 15 markers including WFDC5, TATDN1, LPP, CPLX2, CXCL13, COLI17A1, CEL, CDSN, TMPRSS2, FOXD1, DSC1, LPIN2, MMS4A8, B3GALT2, and AQP10 as the diagnostic markers for LUAD patients. The proposed method was also compared with existing ML algorithms employed on a single omics dataset.

Xiao Ouyang et al. [54] proposed integration methods including Spearman correlation coefficient (SCC), classified information index, fisher ratio, and ensemble of DTs for the identification of biomarkers and classify the LIHC patient into its subtypes. A sample of mRNA expression data, DM, and somatic mutation (SM) data was used and passed to the preprocessing state. After, 34 DEG genes were identified as diagnostic biomarkers using the integration approaches. The identified biomarkers were further analyzed to divide the LIHC patients into 3 subtypes of LUAD.

Nicola Mulder et al. [55] proposed ML algorithms that identify the set of proteins, mRNAs, miRNAs, and DM biomarkers to classify the PDAC into its subtypes accurately. A sample of PDAC patients was obtained from TCGA and cBioPortal which were then passed to a feature extraction technique called neighborhood component analysis (NCA) which identified marker sets involving 49 methylated genes, 50 mRNAs, 20 miRNAs, and 14 proteins. After that, KNN and SVM models were applied, which effectively classified the cancer subtypes with accuracies of 99% and 97% respectively.

Yong-Xia et al. [56] presented a DL framework using a denoising autoencoder DAE to identify subtypes of ovarian cancer OV and to identify genes related to OV. The multi-omics dataset comprising mRNA, miRNA, and CNV was collected using TCGA Assembler and integrated using a DAE. Further, the dataset was passed to the k-mean clustering technique to select the features. These features were then given to the L1-penalized logistic regression (LR) to recognize the subtypes. Also, these features were passed to differential expression analysis and Weighted correlation network analysis (WGCNA) analysis which identified 34 biomarkers related to OV.

Osama Hamzeh et al. [57] used ML algorithms to analyze the Gleason score for prostate cancer and to identify the potential biomarkers for each Gleason group accurately. RNA-Seq data (mRNA and miRNA) were downloaded from the GEO repository and passed to hybrid feature selection techniques for training the model. The experiment was performed, and it has been found that the proposed framework works well with 93% accuracy. Along with that, PIAS3 and UBE2V2 were identified, which will strongly correlate with the progression of prostate cancer.

Xu et al. [58] identified biomarkers related to Cervical Squamous Cell Carcinoma (CESC) by integrating DM and gene expr. data using a hybrid feature selection method. DM and gene expr. profiles of 12 types of cancer have been taken, and adopted ML techniques were applied. The results were evaluated, and it has been found that four cancer-specific markers comprising cg12205729 (GABRA2), cg07211381 (RAB3C), cg26490054 (SLC5A8), and cg20708961 (ZNF257) could identify the tumor cells with sensitivity, specificity and AUC value of 96%, 95% and 92% respectively.

Nguyen et al. [59] used statistical learning and ML algorithms for the identification of diagnostic and prognostic biomarkers in PDAC patients. Transcriptomic (mRNA), Genomic, and protein (rppa) datasets were taken and passed to statistical tests for the identification of diagnostic biomarkers. Further, the survival analysis of the identified was performed using the Cox survival model. The identified biomarkers were also passed to the RF model which will classify the cancer into normal and tumor patients. The proposed framework also showed that the protein expression of identified genes is highly correlated in PDAC patients. The proposed framework identified 4 genes including LAMC2, ANXA2, ADAM9, and APLP2 as diagnostic and prognostic markers.

The work done by researchers on diagnostic biomarker identification for disease prediction in multi-omics is given in Table 2.2.

Table 2.2: Diagnostic Biomarker Identification for disease prediction in multi-omics data

Year	**Algorithm**	**Dataset Used**	**Result**	**Future Studies**
2023 [43]	DBN, Ensemble feature selection	mRNA, DM data from GEO (Alzheimer)	35 markers were identified with 82% accuracy.	The proposed work can be validated in cancer studies in the future.
2023 [44]	MODALN	mRNA, miRNA, DM from TCGA (BRCA, KIRP, Alzheimer)	Top 20 markers from each omic were identified with 83% accuracy.	Clinical and image data will also be considered in the future.
2022 [45]	RDFS	CNV, mRNA from TCGA	Top 20 CNV and mRNA markers were identified.	Advanced neural networks can be applied for better results.
2022 [46]	GCN	mRNA, DM from TCGA (NSCLC)	15 biomarkers from each type was identified with 93.7% F1-score.	More relevant DL algorithms can be applied for the identification of biomarkers accurately.
2021 [47]	KPCA, Spectral Clustering, RI and ARI	Gene expr., DM and isoform expression data from GDAC (LIHC and LUAD)	Proposed method successfully identify 4 genes as diagnostic markers.	Large sample size and ML and DL algorithms can be applied in future studies.
2021 [48]	SVM-RFE, LASSO, RF	Gene Expr., DM from TCGA, GEO (skin)	F accurately identified 10 diagnostic biomarkers with 95% AUC.	Hyperparameter tuning of DL models can improve the performance of the proposed work.
2021 [49]	P-value, FDR, mRMR, DNN	DM and Gene expr. from GEO	8 genes were identified as diagnostic biomarkers with 98% accuracy	CNV data can be integrated along with gene expr., and DM for biomarker identification.
2020 [50]	BayesNet, Cox Regression	DM, Gene expr., and Clinical data from TCGA (BRCA)	7 DMSs were identified as diagnostic markers with 78% AUC.	Treatment therapies can be guided for the identified biomarkers.
2020 [51]	RF, functional analyses	LncRNAs, miRNAs, mRNAs from TCGA	RF successfully identified 25 diagnostic markers with 80% accuracy.	Laboratory researchers can be used to understand the biological functions of identified markers.
2020 [52]	RF, LASSO, LASSO-Cox, Univariate Cox	Multi-omics data from TCGA and GEO (LUAD)	The proposed framework identified 5 diagnostic for LUAD.	DL methods can be applied in future studies.
2020 [53]	DNN, LR, L1-regularized NN and cross modular NN	RNA and DM data from TCGA	SCR-SNN accurately identified 15 diagnostic markers with 89% AUC.	Future research is needed to develop methods for analyzing different diseases.

2020 [54]	SCC classified information index, fisher ratio'	mRNA, DM and SM data from TCGA and GEO	Proposed integrative approach identified 34 diagnostic markers with 99% AUC.	In the future, DL methods can be applied for better results.
2020 [55]	NCA, SVM, KNN	mRNAs, miRNAs, DM dataset from TCGA and cBioportal (PDAC)	KNN accurately identified 50 mRNAs, 49 DMs, 14 rppa and 20 miRNAs with 99% accuracy.	The identified biomarkers can be used for predicting clinical outcomes and guiding treatment strategies.
2020 [56]	DAE, k-mean clustering, L1-penalized LR	mRNA, miRNA, CNV data from TCGA, GSE26712, and GSE32062 (OV).	DL framework accurately identified 19 biomarkers as diagnostic markers.	More clinical features and transfer learning can be used to identify genes related to subtypes of OV.
2019 [57]	Hybrid feature extraction, SVM, RF and NB	mRNA and miRNA data from GEO (prostate cancer).	NB identifies two genes with 95% accuracy and Gleason scores of 7 and 6 respectively.	In the future, a thorough examination of the disease's development, diagnosis, and treatment can be done.
2019 [58]	PCC, Hybrid feature selection, IG, LR	DM and gene expr. dataset from TCGA and GEO (CESC).	The proposed model identified four diagnostic markers with a sensitivity of 96.2%.	The proposed approach can be applied to the development of new epigenetic therapies.
2019 [59]	RF, statistical analysis, Cox regression	Multi-omics dataset from GSE16515 and GSE28735	RF identified 4 diagnostic markers with 90% accuracy.	Future Studies involves the integration of multi-omics data in epidemiological and clinical contexts.

2.1.2 Prognostic Biomarkers for Disease Survival Prediction

Prognostic markers are used to predict the occurrence of a potential clinical condition, disease recurrence, or relapse in an identified sample. The identified markers can also used for disease subtype classification. The work done by the various authors in prognostic markers using single omics is described in the following section.

2.1.2.1 Biomarker Identification in single omics

Jianfeng et al. [60] presented a clustering method to identify the prognostic biomarkers and LUAD subtype prediction. mRNA dataset of LUAD was used from TCGA, GSE203360, and GSE31210. First, DEGs were extracted using the limma package and then stepwise multivariate, univariate Cox, and LASSO were used for the identification of prognostic biomarkers. Clustering was used to subtype the LUAD in 3 clusters. The experiment was performed and it was found that the presented work performed well and identified seven markers as the prognostic biomarkers.

Kountay et al. [61] presented an AI-based DL model to identify the biomarkers for different subtypes of NSCLC patients. First, the input features were passed to autoencoders for feature extraction. The extracted features were then passed to a NN for the classification of NSCLC into its subtypes. The experiment was performed and it was found that the presented work identifies 52 relevant markers with an accuracy of 95.74%. Moreover, out of 52 biomarkers, 28 biomarkers were found to be linked to survival of NSCLC patients.

Eskezeia et al. [62] presented an univariate Cox model to identify the prognostic biomarkers for LUAD patients. The gene expr. data was used and passed to univariate Cox which selected the top DEGs. The extracted DEGs were further passed to a PRPML method which is formed by using four ML algorithms comprising LR, KNN, SVM with RBF, and average neural network (Avnet). The experiment was performed and it was found that the proposed PRPML performed well and identified nine prognostic markers with 81.2% AUC.

Jnanendra et al. [63] presented an ensemble of ML algorithms, including SVM, ANN, KNN, DT, RF, and NB to identify miRNA biomarkers related to BRCA survival. Seven filter feature extraction techniques, including mutual information (MIM), conditional mutual information (CMIM), mRMR, joint mutual information (JMIM), double input symmetrical relevance (DISR), interaction capping

(ICAP), and conditional infomax feature extraction (CIFE) have been used to identify the top features using TCGA portal. The extracted features were passed to the Cox survival model, which identified 27 miRNAs as biomarkers with an HR of more than 1. These markers were identified as poor prognostic markers.

Shuai et al. [64] developed a hybrid feature selection method named Crystall to identify markers and predict the survival time of breast cancer patients. A two-phase model was developed in which the linear support vector regression (lSVR) was first used to extract the methylation features. A rank was generated by lSVR using absolute value, which determines the importance of each feature. The top features were selected, which were passed to an LR to remove the features with small model coefficients. The experiment was performed, and it was found that the Crystall performed well and identified 40 markers with 72% accuracy.

Bhaoshan Ma et al. [65] proposed ML algorithms for the identification of 16 gene prognosis markers for the prediction of LUAD. Clinical and RNA-seq dataset from the TCGA portal was used for the experiment. At first, survival-related genes were identified using Cox, and random survival forest (RSF) method, and then prognostic-related genes were identified from integrated clinical and RNA-seq data. Furthermore, to validate the results, GEO was used. The experiment was performed and compared with existing prediction models. The result was calculated using three metrics comprising hazard ratio (HR), concordance index (CI), and p-value, and it is evident from the results the proposed method outperformed with the CI value of 67%. It was also found that 13 new biomarkers including PITX3, LINC00908, GJB3, MELTF, CRCT1, LOC105370802, BAIAP2L2, GABRA2, RHOV, ARF3, KRT18, TRIM7, ZNF710.AS1 and LOC100996732 were identified as compared to existing studies.

Suman Ghosal et al. [66] make use of ML algorithms to identify the prognostic markers using a noncoding RNA dataset. First data was passed to statistical test to identify the DEGs which were then passed to a multivariate Cox regression model. Further, four ML algorithms comprising LASSO, elastic net, cart, and ridge were used to classify the sample into 5 subtypes of cancer. Then KM analysis was performed which identified 5 lincRNAs (LINC00472, RP4-806M20.3, RP1-40E16.9, RP11-254F7.2, RP11-455B3.1) prognostic markers.

Mohammad Ali et al. [67] used ML algorithms including univariate, multivariate, and Cox proportional hazardous models to identify biomarkers related to OV cancer. Clinical and gene expr. information of OV patients was integrated for selected genes obtained from the Online Mendelian Inheritance in Man (OMIM) database. These were then passed to univariate, multivariate, and combined Cox

Proportional Hazard (CoxPH) regression models which identify the significant biomarkers affecting the survival of the OV patients. Further, Protein-Protein Interaction (PPI) network analysis, Gene Ontology (GO) analysis, and KEGG pathway analysis were performed to identify the essential protein-protein interactions. From the results, it was found that the patients having TL4, BSCL2, CDH1, ERBB2, and SCGB2A1 were less likely to survive as compared to other genes.

Feng Liu et al. [68] aimed to identify prognostic genes of Osteosarcoma using ML. RNA-Seq samples of 94 Osteosarcoma were collected and passed to univariate Cox analysis, LASSO Cox analysis, and multivariate Cox analysis for the identification of prognostic markers. The experiment was performed and the results evidenced that the proposed framework identified four markers (RPL7AP28, RPL11-551L14.1, RP11-326A19.5, and RP4-706A16.3) by dividing the patients into high-risk and low-risk patients.

Jun Yu et al. [69] identify miRNA prognostic markers from esophageal squamous cell carcinoma (ESCC) patients using ML algorithms. A sample of 119 ESCC patients was collected from GEO and TCGA databases where data from TCGA was used as validation set. At first differentially expressed miRNAs were calculated using p-value. The optimal feature subset was selected using Recursive Feature Elimination (RFE). The selected optimal features were passed to the SVM model which classified the patients in early-stage and last-stage samples. Then the risk was calculated using a univariate Cox regression model which identified 5 prognostic markers comprising miR-195-5p, miR-181c-5p, miR-212-3p miR-203, and miR-28-5p for ESCC patients.

Adrian et al. [70] presented a novel hybrid technique called genetic bee colony for the identification of top genes using microarray datasets. At first, the mRMR method was applied, which extracts genes with high correlation and low mutual relevance. The extracted genes from mRMR were then combined to search for the best subset in a narrower search space by using a hybrid of genetic and artificial bee colony algorithms (GBC). The extracted features were trained with SVM with RBF kernel. It was evident from the results that GBC accurately identified seven breast cancer markers with 94% AUC value.

The work done on prognostic biomarker identification for disease survival prediction in single omics is discussed in Table 2.4.

Table 2.3: Prognostic Biomarker identification for disease survival prediction in single omics data

Year	**Algorithm**	**Dataset Used**	**Result**	**Future Studies**
2023 [60]	Clustering, univariate, multivariate cox, LASSO	mRNA from TCGA and GEO (LUAD)	7 biomarkers with an HR close to 1 were identified.	Personalized treatments can be guided based on the identified markers.
2023 [61]	Autoencoders, NN	mRNA data from TCGA (NSCLC)	28 prognostic markers were identified with 95.7% accuracy.	Multi-omics datasets along with pathological images can be used for better results.
2022 [62]	univariate cox, SVM, KNN, LR, Avnet	Gene expr. data from TCGA (LUAD)	9 biomarkers were identified accurately with 81.2% AUC.	Other feature selection algorithms can be considered in the future for better marker identification.
2021 [63]	MIM, JMIM, DISR, CMIM, mRMR, ICAP, CIFE, SVM, DT, RF, ANN, KNN, NB, Cox	miRNA data from TCGA (BRCA)	27 poor prognostic markers were identified with HR > 1.	Large sample size and multiple cancer types can be used to improve the performance.
2021 [64]	Crystall (lSVR+LR)	DM from TCGA (BRCA)	Identified 40 prognostic markers with 72% accuracy.	Multi-omics dataset for BRCA survival prognosis can be used in future studies.
2020 [66]	Multi-variate cox, Ridge, Elastic net, cart and LASSO	Non-coding RNAs dataset from TCGA portal	The proposed algorithms successfully identified 5 lincRNAs as prognostic markers.	In the future, these studies can be extended for further identification of lincRNAs.
2020 [65]	Cox, RSF	miRNA data from TGCA and GEO	13 prognostic markers were identified with 67% CI value.	DL can be used for the prognosis of cancer and provide a more powerful tool for targeted therapy in the future.
2019 [67]	Univariate, Multivariate CoxPH model	Gene expr. dataset from OMIM (OV)	Identified 5 genes with an HR close to 1 for each gene.	The search can be applied in the biomarker identification for different types of cancers.
2019 [68]	Univariate, Multivariate, Lasso cox	miRNA data from TCGA (osteosarcoma)	Four markers were identified with an HR close to 1.	Future Studies involves the validation of experiment using different datasets.
2019 [69]	RFE, SVM, univariate cox	miRNA data from GEO (GSE43732) and TCGA (ESCA)	Five microRNA markers were identified as prognostic markers with an HR close to 1.	DL can be employed for the identification of prognostic biomarkers using large size dataset.
2018 [70]	mRMR, GBC, SVM with RBF kernel	13 publically available microarray datasets	7 BRCA markers were identified with 94% AUC.	Proposed GBC can be applied in multi-omics datasets for accurate biomarker identification.

2.1.2.2 Biomarker Identification in multi-omics

Yongquing et al. [71] presented a framework named LAGProg based on local augmented GCN for the identification of prognostic biomarkers and cancer prognosis prediction. Multi-omics dataset including CNV, mRNA, and DM was used and passed to conditional variational autoencoder (CVAE) to reduce the dimensionality of features. Then the features extracted from CVAE and the actual features were passed to GCN and Cox for cancer prognosis. The experiment was performed and the results evidenced that the GCN with Cox performed effectively and identified 13 prognostic markers for BRCA with 70.4% CI value.

Simak et al. [72] introduced huper-parameter optimized autoencoders (HPOAE), penalized PCA (PPCA), normal autoencoders, Aand coxPH models for the discovery of prognostic biomarkers and survival prediction. Three types of data comprising mRNA, miRNA, and DM of colon patients were used and integrated using HPOAE, PPCA, and normal autoencoder. The integrated features were then passed to coxPH for survival prediction. The conducted experiment performed well and 10 miRNAs, 11 DMs, and 28 mRNAs were identified as prognostic biomarkers accurately.

Xu et al. [73] presented a hybrid of SVR and RFE for the identification of prognostic markers and predicting PDAC clinical prognosis. First, differentially expressed mRNAs were identified using statistical tests (FDR and $Log_2(FC)$), which were then passed to SVMRFE for survival prediction. The proposed work accurately identified 70 markers related to poor prognosis with a mean square error (MSE) of 0.001.

Zhang et al. [74] developed a fusion method named Deep Latent Space Fusion (DLSF) for biomarker identification and cancer subtyping that works by integrating the different types of omics data. Multi-omics data, including miRNA, DM, and gene expr of KIRC patients, were fused with cycle autoencoder (CAE) and passed to k-mean clustering for subtype classification. The presented DLSF performed well and identified 15 prognostic markers with a p-value of less than 0.01.

Yeye et al. [75] presented a sure independence screening procedure based on the distance correlation (DC-SIS) method to extract the features from mRNA and CNV datasets from TCGA and METABRIC (Molecular Taxonomy of Breast Cancer International Consortium) portal related to breast cancer prognosis. The extracted features were combined to form a network and passed to the LR and RF classifiers to perform the experiment. The results evidenced that the proposed work performed well by identifying five breast cancer prognostic markers

accurately with 88% AUC.

Wang et al. [76] developed a new multi-omics integrative approach known as MOGONET for the identification of biomarkers and the classification of cancer subtypes using gene expr, DM, and mRNA data. Combining omics-specific and cross-omics correlation learning, MOGONET aims to classify multi-omics data efficiently and effectively. MOGONET surpassed other state-of-the-art techniques with 82% accuracy and identified the top 30 markers as prognostic markers.

Hua et al. [77] presented a denoising autoencoder (DCAP) with XGBoost for biomarker identification and cancer survival prediction using miRNA, mRNA, DM, and CNV data. First, multi-omics data was passed to a DAE to reduce the dimensionality of the dataset, followed by a Cox proportional hazard model to accurately estimate the risk of cancer patients. Further, XGBoost was used to identify the markers by computing the feature importance of each feature. The presented work accurately identified nine markers with a CI value of 0.66.

Kailun Zhou et al. [78] used ML algorithms for the identification of prognostic markers of PDAC patients. mRNA, SNP, and CNV dataset was used and passed to GISTIC 2.0 and Mutsig 2.0 to preprocess the omics data. 54 candidate genes were identified which were then integrated and passed to the LASSO risk prediction model. A total of 9 markers comprising TSPYL4, UNC13B, KLHDC7B, MICAL1, AIM1, KLHL32, DCBLD1, ARHGAP18, and CACNA2D4 were identified by the proposed work as the prognostic markers.

Ning Zhao et al. [79] suggested a technique to rank genes by calculating a score for the identification of biomarkers. Multi-omics data comprising somatic copy number alteration (sCNA), gene expr., DM and miRNA data of 13 cancer types were taken, and the multivariate Cox proportional method was used to rank the gene. The experiment was performed, and higher ranks genes were identified as the prognostic biomarkers. Further CI was used to validate the results, and it was found that in comparison to single omics, multi-omics works well with a CI value of 0.95. In fact, when contrasting the genes related to 13 types of cancer, 7 genes were shown to be linked with a number of cancer prognoses.

Yu-Heng Lai et al. [80] proposed a DNN framework using novel biomarkers to predict the survival of NSCLC. A sample of 614 patients having gene expr. and clinical data was taken and integrated with 15 biomarkers to develop an integrative DNN model. The biomarkers were discovered using the StepMiner algorithm. The experiment was performed, and it was found that the proposed framework works well by accurately predicting the survival of NSCLC patients with 70% accuracy.

Lei Cui et al. [81] proposed the DL framework U-net for the identification of

biomarkers used in the survival prediction of LUAD patients. A sample of 191 patients was from the TCGA portal, and U-net was applied to segment the images. The CoxPH model was used to predict survival. Four biomarkers were discovered guiding the survival of LUAD patients.

Wnju Mo et al. [82] used RF for the identification of prognostic biomarkers using the multi-omics dataset of BRCA patients. Sample of mRNA, SNP, CNV, and clinical information was taken and integrated together which were then passed to RF for feature selection. This technique identified 120 candidate genes. These genes were then passed to the Cox regression model for the identification of prognostic genes. The experiment was performed and it was evident from the experiment that the proposed algorithm successfully identified 6 genes including CD24, PRRG1, IQSEC3, MRGPRX, RCC2, and CASP8 as the prognostic markers.

Qianxing Mo et al. [83] presented a clustering approach for the identification of the prognostic value of muscle-invasive bladder cancer (MIBC) patients using a multi-omics dataset. A sample of 388 patients including DM, SM, mRNA, and CNA was used and passed to the iClusterBayes method. This will divide the data into two clusters, i.e., basal and luminal subtypes clusters. These clusters were validated using the Markov Chain Monte Carlo (MCMC) method. A total of 42 genes were identified which were further passed to statistical analysis tests including Fisher's exact test, two-sampled t-test, and Analysis of Variance (ANOVA) methods which identified 7 genes as prognostic markers.

Zhiqiang Chang et al. [84] proposed a pipeline to identify the dosage-sensitive markers in CC. The somatic copy number, mRNA, and gene expr. data were used and passed to the Wilcoxon rank-sum test for the identification of DEGs. The genes with a p-value greater than 0.3 were selected and passed to Cox regression analysis. Finally, PCC was calculated which identified 6 driver genes including WDR5B, NDUFB4, IQCB1, GTF2E1, SEC22A, and KPNA1 as poor prognostic markers.

Yang Yuan et al. [85] developed clustering algorithms in multi-omics data to identify the prognostic biomarkers related to brain tumors. A sample of 117 glioblastoma patients including mRNA, CNV, SNP, DM, and clinical information was used to perform the experiment. MutSigCV was used in the analysis of SNP data which decreased the number of false positives. For CNV data, GISTIC was used to extract the important copy number genes. Then the genes were integrated and passed to the cluster of cluster analysis (CoCA) algorithm which divided the data into two clusters HX-1 and HX-2. The survival analysis of these clusters was

performed which identified 3 methylation and 15 gene mutations as the prognostic markers.

Yanhui Jia et al. [86] used the PCA method for the dimensionality reduction of huge feature space of multi-omics data comprising miRNA, mRNA, DM, and SNP data of PDAC and identified 11 prognostic biomarkers. First, the dimensionality is reduced using PCA and then the most relevant features are selected using Filter methods including chi-square and t-test. The proposed method identified 12 markers comprising hsa-mir-1224, hsa-mir-1179, hsa-mir-129-1, hsa-mir-1251, hsa-mir-129-2, MAPK8IP2, DPP6, CPE, IL20RB, MSI1, FMN2, and S100A2 with a HR close to 1.

Tzhong et al. [87] proposed a DL framework called autoencoder CoxPH model (AE-cox) for the identification of biomarkers and survival prediction of LUAD patients. miRNA, mRNA, DM, and CNV datasets of LUAD patients was extracted from TCGA and GEO portal and passed to AE to reduce the dimensionality of features. Further, cox analysis was performed to identify the poor prognostic markers and classify cancer as poor and good survival patients. The proposed experiment accurately identified ten genes with a 0.65 CI value.

Nitish Kumar et al. [88] presented survival models for the identification of biomarkers in PDAC patients. A sample of 153 PDAC patients was taken from the TCGA database for multi-omics analysis which consists of gene expr., DM, miRNA and lncRNA data. The preprocessing and feature extraction were performed to identify the genes that are positively correlated with survival. For survival analysis, Cox and KM estimations were done. The experiment was performed, and results proved that the presented work performed well by accurately identifying 5 genes with an AUC value of 95%.

Ruang Zhang et al. [89] developed prognostic models for the early prediction of LUAD using trans-omics biomarkers. Authors integrate the clinical, DM, and gene expr. dataset of 825 patients and used the Ranger algorithm for screening the biomarkers associated with prognosis. Biomarker information and clinical data were fused using icluster plus algorithm to divide the patients into high-mortality and low-mortality rates. The experiment was performed, and it was evident from the results that the developed method improved the performance of the trans-omics model by 18.3% with an AUC value of 87.2%.

Chen Peng et al. [90] proposed a DL framework called Capsule Network-based Modelling of Multi-omics data (CapsNetMMD) for the detection of BRCA-related genes. A sample of 770 BRCA patients, including DM, miRNA, and CNA have been taken and converted into a matrix form. It was then passed

to CapsNetMMD for the detection of BRCA-related genes. The experiment was performed, and the results were evaluated. The results were also compared with different ML algorithms comprising XGBoost, NN, SVM, Adaboost, and KNN and it was marked that CapsNetMMD outperformed with 90% Accuracy.

Jayeon Lim et al. [91] used a DL framework called Artificial Neural Network (ANN) for the analysis of genetic data and for the discovery of disease-related genes. TCGA dataset of BRCA patients was taken, and the experiment was performed. For parameter optimization, the lasso penalty activation function was used. The model was compared using the Youden J index with other ML algorithms, including meta LR and meta-SVM. It was estimated from the results that the suggested DL framework was more robust in the identification of genes related to BRCA.

The work done by researchers on prognostic biomarker identification for disease survival prediction in multi-omics data is given in Table 2.4.

2.1.3 Predictive Biomarkers for Response/ Treatment

A predictive biomarker is a test that can be used to classify people who are more likely to react to a certain medicinal substance or chemical product. Asymptomatic gain, increased longevity, or an adverse effect may be the result. In this research, a predictive biomarker is considered as a gene prioritization problem where the gene can signify the occurrence of some particle disease with some known disease genes. The work done on predictive biomarker identification using single omics for treatment/ response is given in the forthcoming section.

2.1.3.1 Biomarkar Identification in single omics

Siquan et al. [92] used ML algorithms comprising WGCNA and RF to identify the prognostic biomarkers using gene expr. dataset of Alzheimer patients. First, the WGCNA was conducted which screened a total of 3718 genes. Then the RF was used to rank the genes which identified 5 prioritized genes comprising FAM71E1, AP4M1, DDB2, DOC2A, and GPR4 as predictive biomarkers. These identified biomarkers were highly enriched and associated with immune cell designations. The nomogram was built which showed the high predictive power of identified biomarkers.

Tianyi Zhao et al. [93] presented a GCN for the prioritization of protein-coding genes using the lncRNAs dataset. A sample of the lncRNA dataset was used and

Table 2.4: Prognostic Biomarker identification for disease survival prediction in multi-omics data

Year	**Algorithm**	**Dataset Used**	**Result**	**Future Studies**
2023 [71]	GCN, CVAE, Cox	mRNA, DM, CNV from TCGA	13 prognostic markers were identified with 70.4% CI value.	The study of relationships between the survival risks and transcriptional regulation can be done.
2023 [72]	CoxPH, HPOAE, PPCA, normal autoencoders	mRNA, miRNA, DM from TCGA (colon cancer)	10 miRNAs, 11 DMs, and 28 mRNAs were identified with 71% accuracy.	Large sample sizes and supervised DL methods can be used in the future for better performance.
2022 [73]	RFE, SVR	mRNA, miRNA from TCGA(PDAC)	70 prognostic markers were identified with 0.001 MSE.	Treatment therapies can be guided using the identified biomarkers for future studies.
2022 [74]	CAE, K-mean Clustering	Gene expr., DM, miRNA data TCGA (KIRC)	15 prognostic markers were identified with a p-value < 0.01.	The presented work helps practitioners to understand the biological significance of identified biomarkers.
2022 [75]	DC-SIS, LR, RF	mRNA, CNV from TCGA and METABRIC (BRCA)	5 prognosis genes were identified with 88% AUC value.	Complete follow-up information can improve the performance for biomarker identification and cancer prognosis.
2021 [76]	NN to rank features, GCN	gene expr., DM, and mRNA from TCGA (BRCA, KIPAN)	30 biomarkers were identified with 82% accuracy and 82.5% F1-score.	More data types can be used to improve the performance for biomarker identification.
2021 [77]	Denoising AE, DNN, XGBoost	miRNA, mRNA, CNV, DM from TCGA, GEO (BRCA)	9 markers were identified with 0.66 CI value.	Parameter tuning of learning models can be done to improve the performance results.
2021 [78]	GISTIC 2.0, Mutsig 2.0, LASSO, univariate cox	mRNA data from TCGA and gene expr. from GSE28735,and GSE62452	The presented method identified 9 prognostic markers with 87% accuracy.	Future Studies include the verification of genes in vivo and in vitro. A thorough investigation will be done in future studies.
2020 [79]	Multi-variate cox model	DM, Gene expr., sCNA and miRNA data from TCGA	7 prognostic markers were successfully identified with 95% CI value.	Some non-parametric algorithms can be applied to study biomarkers in future studies.
2020 [80]	DNN, step miner	Gene expr. and clinical data from GEO	15 prognostic markers were identified with HR in the range of 1.2-7.063.	This framework can be employed to predict the survival of other cancers.

2020 [81]	U-net, coxPH	Pathological images dataset from TCGA	Four prognostic markers have been identified with 68% CI value.	The results can be improved by integrating multi-omics datasets using the proposed framework.
2020 [82]	RF, Cox Regression	mRNA, SNP, CNV and clinical from UCSC (BRCA)	Proposed models identified 6 prognostic markers with 80% AUC value.	Experimental Validation is required because of the limited clinical information present in the current research.
2020 [83]	iCluterBayes, Fishers test, t-test, ANOVA	DM, SM, mRNA and CNV data from firehose [94]	6 genes were identified as prognostic markers.	Future studies involve the identification of markers from large sample-size datasets.
2020 [84]	Wilcoxon Rank-sum test, cox regression, PCC	CNA, mRNA and gene expr. data from TCGA (CC)	The proposed work selects 6 prognostic genes having p-value < 0.05.	Target drugs or treatment methods can be developed in future studies.
2020 [85]	CoCA, GISTIC 2.0, and Mutsig	SNP, CNV, DM, mRNA and clinical from TCGA	CoCA identified 2 DMs and 15 gene mutation genes as prognostic markers.	Treatment therapies can be provided based on the identified markers.
2020 [86]	PCA, Cox	mRNA, miRNA, and DM from TCGA (PDAC)	12 genes were identified with an HR close to 1.	Treatment therapies can be guided based on the identified markers.
2020 [87]	AE, AE-Cox	mRNA, miRNA, CNV, DM from TCGA, GEO (LUAD)	10 markers were identified with 0.65 CI value.	Biomarkers for multiple cancer types can be identified using the proposed work.
2019 [88]	LR, Cox Regression	DM, gene expr., miRNA and lncRNA data from TCGA (PDAC)	5 prognostic biomarkers were identified with HR of each gene lies in the range of 1-2.	Artificial intelligence can be employed in the future for better performance.
2019 [89]	Ranger, iCluster plus, multi-variate cox	Clinical, DM and gene expr. dataset TCGA (LUAD)	The proposed work identified 7 prognostic markers with 81% CI.	Futures studies involve the exploration of biological evidence for the identification of markers.
2019 [90]	CapsNetMMD	DM, miRNA, and CNA dataset from TCGA (BRCA)	CapsNetMMD outperformed and identified top 5% genes with 90% sensitivity.	In future, the predicted genes with prognostic values in BRCA may serve as candidates for ecologists and medical scientists.
2019 [91]	ANN, Youden J index, meta LR and meta SVM	mRNA, DM, CNV dataset from TCGA (BRCA).	ANN outperformed Meta-SVM by successfully identifying the prognostic biomarkers.	More than two types of data sources to construct a multimodal learning model can be used for better prediction.

passed to the feature selection technique in which the gene expr. and the position of the gene was identified and a gene network was created. This network was then passed to GCN which prioritizes the target genes of lncRNAs. The method was also validated and compared with existing methods and it was found that GCN works well with an AUC and Area under Precision Recall (AUPR) value of 90% and 91% respectively.

Yu Zhang et al. [95] proposed the Prioritization of autism genes using a Network-based Deep-learning Approach (PANDA) for the identification of genes by prioritizing them. The protein dataset was extracted from Simons Foundation Autism Research (SFARI) and OMIM data. First, a human molecular interaction network (HMIN) was constructed in which nodes represent the proteins corresponding to their gene and edges represent the interaction between the genes. This network was then passed to GCN which trains the dataset and prioritizes the genes on the basis of their influence on the patients. The experiment was performed and it was evident from the results that the proposed framework worked well by selecting 10 genes including RUNX1T1, MAG12, GRIA3, MVCRP2, AKAP6, PTPRD, AUTS2, MYO9A, AB12, and PLXNA2 respectively.

Xue Jiang et al. [96] presented a generative adversarial network (GAN) with DAE as the generator and MLP as the discriminator (GAN-DAEMLP) to prioritize genes by taking the miRNA dataset. The sample of the dataset was taken and passed to GAN-DAEMLP for disease and non-disease genes which calculates the disease and non-disease prediction score. Finally, a risk score was calculated and a genes risk list was generated. The experiment was performed and it was proved from the results that the GAN-DAEMLP performed best by selecting 10 disease-related genes (TRAIP, Bsgnt2, Ugt8a, PPP3CA, PMEPA1, RGS4, PPP3R1, CHN1, and ST8SIA3).

Jonghyun Nam et al. [97] designed a method Gene Ranker for the identification of genes using gene expr. dataset. First, a PPI network was created which was used as a base network. Then, an add-on network was generated using WGCNA. An integrated network was developed and passed to a gene ranker algorithm to generate a score. The higher rank genes including OTC, B3GNT9, and Clorf167 were identified as the predictive markers. Along with that, the 10 known genes including CSNK2A3, IFNL2, UCN3, POU3F4, TIW1, IL22, UCN2, PSG1, HTRA1 and CD68 were also identified. These genes have a strong relation with the above-mentioned identified genes.

The work done on predictive biomarker identification for treatment/ response in single omics is discussed in Table 2.5.

Table 2.5: Predictive Biomarker Identification for treatment/ response in single omics data

Year	Algorithm	Dataset Used	Result	Future Studies
2022 [92]	RF, WGCNA	Gene Expr. from GSE5281, GSE48350 (Alzheimer)	5 genes were identified accurately with high predictive power.	Experimental Validations are required for identified biomarkers.
2020 [93]	GCN	lncRNAs dataset from TCGA	The presented method prioritize the candidate genes with 92% AUPR.	DMs and CNVs can be considered for better results.
2020 [95]	PANDA, GCN	Protein data from SFARI and OMIM database (Autism)	PANDA successfully identified 10 genes with an accuracy of 89%.	This framework help researchers to learn and perform better by understanding the complex genetic architecture of diseases and disorders.
2020 [96]	GAN-DAEMLP	miRNA expression dataset from CHDI Foundation [98]	GAN-DAEMLP identified top 9 genes as predictive markers with 90% AUC value.	The proposed work can be applied to the identification of cancer biomarkers.
2019 [97]	WGCNA network	Gene Expr. data from the GEO Database	Proposed method successfully identify 3 genes as predictive markers with 76% AUC.	Future work is extended to identify disease categories. Along with that, different methods or algorithms can be created to integrate the networks.

2.1.3.2 Biomarkar Identification in multi-omics

Runzhi et al. [99] developed asmbPLS-DA based on an adaptive sparse multi-block partial least square discriminant analysis method for the identification of predictive biomarkers and for the classification of different types of diseases. Multi-omics dataset was considered and passed asmbPLS-DA which finds the covariance among the latent variables. The experiment was performed and presented asmbPLS-DA accurately identified the 14 mRNAs and 6 miRNAs as predictive biomarkers of response with 97% accuracy.

Jae et al. [100] presented NTriPath for the identification of gene signatures for Gastric Cancer patients. The SM and gene expr. dataset was used in which SM was passed to NTriPath and gene expr. was passed to the clustering algorithm. Then the output from NTriPath and unsupervised clustering were integrated and passed to SVM with linear kernel to generate a risk score for the identification of prognostic and predictive biomarkers. The predictive biomarkers were further used to identify the response to a specific treatment in Gastric cancer patients.

Poria et al. [101] presented Reboust Rank Aggregation (RRA) and WGCNA to identify the predictive biomarkers in CC patients. mRNA, miRNA, and lncRNA dataset was used and passed to RRR and WGCNA to identify the DEGs. A total of 37 DEGs were identified which were further passed to the regulatory network, survival analysis, and ML algorithms comprising RF, XGBoost, SVM, and LASSO. The experiment was performed and one gene comprising LINC00974 was identified as the diagnostic, prognostic, and predictive biomarker.

Yang Wu et al. [102] proposed an ML framework called a semi-supervised non-negative matrix factorization model called MapGene to prioritize the candidate genes using high functional modules and gene interactions dataset. First, a PPI network was made of both disease interactions and network interactions, and then module correlation (MC) was calculated using the MapGene algorithm which identified the top rank genes as predictive markers. The proposed framework was also compared with several base models and it was found that the MapGene outperformed with a precision and recall value 0f 87% and 90% respectively.

Haixia et al. [103] developed an integrative rank method comprising iRank, and Constrained Page Rank (CPR) to prioritize the cancer genes using multi-omics data of HCC patients. A multiplex network was generated using multi-omics data by calculating the differentially mutual information (DMI). This DMI was then passed to the PageRank algorithm and the final rank was obtained by aggregating the rank of multiple networks. The proposed method iRank was compared with other existing algorithms and it outperformed with an accuracy of 81%.

Jenfeng Zia et al. [104] proposed a method for Driver gene discovery with an improved random walk method (Driver_IRW) using the integration of transcriptomic and interaction network data. Authors constructed different networks for different types of cancer using edges from PPI and DCG networks. Then the edge, betweenness, and Katz centralities were found using the constructed network. These scores were integrated and passed to a random walk with an improvement method to calculate their rank. Finally, top-ranked genes were selected as the predictive markers. The proposed method was compared with several traditional methods and it was found that Driver-IRW performed best by accurately prioritizing the driver genes.

Anais Baudot et al. [105] proposed a random walk with a restart method to prioritize the genes on multiplex (RWR-M) and multiplex heterogeneous networks (RWR-MH). First, a graph of the PPI network, pathway interaction, and co-expressed genes was created. The integrated network consists of 17559 nodes and 1659084 edges which were then passed to RWR-M and RWR-MH to explore the

different functionalities and associations of the graph. This was then applied to Wiedemann Rautenstrauch syndrome patients which identified three genes (FIG4, RNF113Aand LMNA) that were strongly related to the disease.

Zhen Zeng et al. [106] proposed a tree-based ensemble model called random interaction forest (RIF) to prioritize candidates and generate the predictive scores. First, a DT was created and then the rank was calculated. The authors identified the top 10 genes and compared the results with other existing methods.

Naoya et al. [107] used the joint non-negative matrix factorization (JNMF) method for the discovery of predictive biomarkers using a multi-omics dataset. The authors took the dataset in three matrix form and applied the JNMF method which generates four clusters by reducing the dimensionality of the matrices. This method also reduces noisy values and selects the relevant features. This method successfully identifies the two candidate genes comprising PLX4720 and HER2 as predictive biomarkers and also finds the association between the genes and the drugs given for treatment

Christos Dimitrakopolos et al. [108] developed a Network-based Integration of Multi-omics data (NetICS) method for prioritizing the cancer genes by integrating genetic aberrations, mRNA and miRNA, and DM datasets. A bidirectional network diffusion was created which generates a rank list for each sample. This rank list was then passed to rank aggregation techniques which generated a global ranking. NetICS identified the top 5% genes from both BRCA (TP53, PTEN, ERBB2, and CDH1) and LUAD (EGFR, AKT1, KRAS, PIK3CA, and NRAS) respectively.

Yuanfang Guan et al. [109] designed a feature selection method and SVM to prioritize the predictive genes using gene expression and DM data. The PCC of genes was calculated and their correlation scores were combined to generate a rank. By using this, 10 most predictive features including ASAP2, BCL9L, PTPRF, PTPN12, ANXA1, AJUBA, CYTIP, SH3D19, CMTM4, EIF2C2 were selected.

Di Zhang et al. [110] developed a network-based approach for the identification and prioritization of predictive genes by integrating gene expr., mutation, and PPI datasets. This approach works by identifying the neighbor genes. A relationship among the various differentially co-expressed genes (DCGs), and functional genes was made and then weight was calculated to check the impact of DSCs on the functional genes. This procedure was applied to three datasets including KIRC, thyroid carcinoma (THCA), and HNSC to identify the genes. The experiment was performed and it was found that the proposed method identified the top 5 genes

including EGFR, EP300, NRAS, LYN, PTPN11, TP53, PIK3CA, EGFR, EP300, FADD, PBRM1, SETD2, BAP1, SRC and EP300 for THCA, HNSC, and KIRC respectively.

Huihui Fan et al. [111] integrate multi-omics data including genome, epigenome, and transcriptome data which identify and prioritize the functional differentially methylated regions (fDMRs). Authors first filter the DMRs and based on the expression alteration scores, ranks were generated and further aggregated to identify and prioritize the genes. This method identified 10 genes including VIM, PCDH10, SFRP1, ADAMTS1, SLIT2, CDH4, SFEP2, HS3ST2, and CHFR using ranks. Further, classification and survival analysis of identified genes was performed.

Qianlan Yao et al. [112] proposed a method MetPriCNet to prioritize and predict the metabolites using a multi-omics dataset. The authors constructed a composite network of genomic, phenome, metabolome, and interactome datasets. This network consists of 25269 nodes and 11,926,113 edges. This network was then passed to MetPriCNet which calculated their global distance similarity. This method was applied to BRCA patients and it was found that the higher rank metabolite in 3 genes including TP53, AKT1, and BARD1, and the third-ranked magnesium ion metabolite interact with 4 seed genes consisting of CDH1, KRAS, CHEK2, and CDS1.

Pia Kinaret et al. [113] proposed fuzzy logic as feature selection and RF for prioritizing the genes using multi-class. Four gene expr. dataset was taken and passed to the fuzzy pattern discovery method to select the most relevant and class-specific features (FP). Then the selected feature set (FP) was passed to RF which works by removing the redundant features and ranking the genes by using a Mean decrease accuracy score. The proposed method works well with an accuracy of 96%.

The work done on predictive biomarker identification for treatment/response in multi-omics data is discussed in Table 2.6.

2.1.4 Identification of other Biomarkers

There are four other types of biomarkers comprising safety, monitoring, risk, and response biomarkers. Limited work is done on the above-mentioned markers using both single and multi-omics datasets, for example, Andrew et al. [114] presented the Connor-Davidson Resilience Scale (CD-RISC) method for the identification of risk biomarkers using DM datasets of 78 adults. The risk score was computed to select the features which were then passed to LR and SVM for the classification

Table 2.6: Predictive Biomarker Identification in multi-omics for treatment/ response

Year	**Algorithm**	**Dataset Used**	**Result**	**Future Studies**
2023 [99]	asmbPLS-DA	mRNA, miRNA data from TCGA	14 mRNAs, and 6 miRNAs were identified with 97% accuracy.	The raw data was used without pre-processing which can be done in future works.
2022 [100]	NTriPath, SVM, unsupervised clustering	SM and gene expr. from TCGA (gastric)	32 gene signatures were identified as prognostic and predictive markers with HR close to 0.5	The investigation of the molecular mechanisms underlying the prognostic and predictive capabilities of the 32-gene signature can be done.
2022 [101]	RRA, RF, WGCNA, SVM, LASSO	miRNA, mRNA, lncRNA from TCGA (CC)	LINC00974 was identified as a prognostic, diagnostic, and predictive marker.	The role of LINC00974 can be further assessed in the CC patients.
2021 [102]	Semi-supervised method MapGene	Multi-omics dataset from DisGeNet and String database	MapGene outperformed with 87% precision value.	Sparse connections are there because of large modules which can be broken down into smaller modules for better performance.
2020 [103]	iRank, CPR	DM, mRNA, SM, miRNA, CNV from TCGA (HCC).	iRank outperformed by accurately prioritizing the cancer genes with 81% accuracy.	Presented iRank can be applied to other cancer types.
2020 [104]	Driver-IRH	Transcriptomic and network data of BRCA, HNSC, KIRC, and THCA from TCGA	Driver_IRW successfully identified top 10 genes for each cancer type with 90% recall value.	This method can be applied to classify patients in different subtypes of cancer by using the identified genes.
2019 [105]	RWR-M, RWR-MH	Multi-omics dataset from TCGA, and OMIM database.	RWR-M and RWR-MH outperformed the basic RWR method with an accuracy of 89% and 82% respectively.	Other biological networks can be constructed for better gene identification.
2019 [106]	RIF	Clinical data from TCGA	Proposed method identified top 10 genes as predictive markers.	To adjust the nuisance covariates, regression models can be used for better performance.
2018 [107]	JNMF	Genomic, transcriptomic from cBioPortal	2 predictive biomarkers were identified accurately by JNMF.	The pathological images along with multi-omics can be considered for better performance.

2018 [108]	NetICS	Genetic, mRNA, miRNA, DM from TCGA	Top 5% genes from both LUAD and BRCA patients were identified with 89% AUC value.	More complex mutational patterns along with the genomic and transcriptomic data can be integrated for better performance.
2018 [109]	PCC, SVM	Gene Expr. and DM data from synapse [115]	The proposed method identified 10 predictive genes with 80% accuracy.	This research can be extended to additional datasets and algorithms for better results.
2017 [110]	Network based approach	PPI, gene expr., mutation data from TCGA, OMIM, GEO (KIRC, HNSC)	Proposed method outperforms various existing methods by accurately identifying top 5 genes for each cancer type with 80% accuracy.	In the future, gene expr., CNV and DM data will be integrated to construct a network. Further, optimal treatment techniques can be guided to patients by integrating the dataset.
2015 [111]	Cox	Multi-omics dataset from GEO	The proposed method identified 10 predictive markers with 86% AUC value.	DL can be employed in the future for better performance.
2015 [112]	MetPriCNet	Multi-omics data from STRING, OMIM, TCGA portal.	MetPriCNet prioritizes and predicts the candidate genes with an AUC value of 91%.	This proposed framework can be used in different fields of biomedicine like disease prediction, drug discovery, and target discovery.
2014 [113]	RF, Fuzzy Logic	Four multi-class gene expr. datasets from GEO, and St. Jude Research [116]	The proposed framework performed well by and successfully prioritized the genes with 96% accuracy.	DL algorithms can be used in the future for better results.

of low and high-risk patients. The experiment was performed and the proposed work identified three markers comprising AARS (cg18565204), FBXW7 (cg17682313), and LINC01107 (cg07167608) as high-risk markers with 72.3% and 87.1% accuracies for LR and SVM respectively.

Kong et al. [117] proposed an ML framework called NetBio, which utilizes network-based studies to effectively discover the response biomarkers for immune checkpoint inhibitor (ICI) treatment. A collection of over 700 patient samples treated with ICI was done which were accompanied by clinical outcomes and transcriptome data. Then NetBio-based predictions were than and the findings demonstrated that the proposed work effectively forecasted the responses to ICI treatment. Furthermore, the utilization of NetBio-based prediction demonstrates a higher level of effectiveness compared to existing state-of-the-art works.

Danquing et al. [118] presented the limma and RebustRankAggreg method to identify the monitoring biomarkers related to CRC. The DEGs were extracted which were further passed WGCNA for biomarker identification. The experiment was performed and the findings showed that the proposed work identified four biomarkers comprising AMPD1, ABCC13, TMIGD1, and SCNN1B as monitoring biomarkers with an AUC value of 70%.

Jungho et al. [119] presented machine learning and network-based analysis for the identification of response biomarkers using genomic and transcriptomic data of bladder cancer and CRC patients. First PPI network was created which was further passed to ridge regression for training. The findings showed that the proposed work performed effectively and identified top markers as response biomarkers.

Yun et al. [120] employed WGCNA to identify gene modules that have a strong association with the risk of BRCA metastasis. First, a total of 21 network hub genes were selected with the highest level of significance. Subsequently, PPI networks were employed to further investigate the biomarkers exhibiting the highest number of connections among gene modules. The PPI networks discovered five genes as monitoring biomarkers. Additionally, the validation of the prognostic value and DEGs was conducted using data obtained from TCGA and KM Plotter. The examination of the ROC curve demonstrated that the mRNA expression levels of the five hub genes exhibited exceptional diagnostic efficacy in distinguishing BRCA from surrounding tissues.

Anna et al. [121] Monte Carlo feature extraction and NB model for identification of response biomarkers using two gene expr. datasets of BRCA patients undergoing radiation therapy. The datasets used were with different dose treat-

ments. The validation of genes was done using the Jonckheere–Terpstra test. The experiment was performed and NB successfully identified three response biomarkers comprising GADD45A, ZMAT3, and NAMPT with an accuracy of 93.5%.

The work done on other biomarker identification in single and multi-omics data is given in Table 2.7.

Table 2.7: Identification of other biomarkers in single and multi-omics dataset

Year	**Algorithm**	**Dataset Used**	**Type of Biomarker**	Future Scope
2023 [114]	CD-RISC, LR, SVM	DM dataset of 78 adults	Risk	Other factors including age, history can be included for better results.
2022 [117]	NetBio	Transcriptomic and Clinical data from TCGA	Response	The presented approach can be applied to different diseases along with cancer.
2020 [118]	Limma, RobustRankAggreg, WGCNA	DM dataset from TCGA, GEO (CRC).	Monitoring	Multi-omics datasets can be considered along with DL algorithms for biomarker identification.
2020 [119]	PPI, ridge regression	Genomic and Transcriptomic data from GEO.	Response	Paired datasets providing molecular changes before and after drug treatment can be used with ML for better results.
2019 [120]	WGCNA, PPI, KM	Gene expr. from GEO, TCGA (BRCA)	Monitoring	Biological significance of the identified markers can be done.
2019 [121]	Monte Carlo, NB	Two gene expr. data from GEO (BRCA)	Response	This method will help clinicians to study the impact of particular doses in the future.

2.2 Tools used for Biomarker Identification

It is a challenge for researchers without bioinformatics skills to identify the biomarkers by analyzing a high volume of multi-omics data. Therefore, many researchers developed tools for biomarker identification created from omics technologies, for example, Ganxun Li et al. [122] developed an IMOPAC web server with the potential to simplify the interpretation of pharmacogenomic profiles derived from cell lines by using transcriptomic, metabolomic, epigenetic, genetic and proteomic datasets. The user-friendly query interface along with tailored data storage enables users to interactively examine and display multi-omics variations across genes and pathways and to link these alterations with treatment responses across cell lines from varied cancer types. The developed tool can potentially discover under-

lying biological mechanisms and facilitate pharmacogenomics exploration in the identification of clinically relevant biomarkers.

Anqi et al. [123] designed a proprietary web-based tool, named Comprehensive Analysis in Multi-Omics of Immunotherapy in Pan-cancer (CAMOIP) for the identification of prognostic markers. The tool uncovers the underlying mechanisms governing biomarker expression, functionality, and immunotherapy in pan-cancer. This may be accomplished conveniently through the utilization of the CAMOIP platform, hence facilitating and promoting the advancement of immunotherapeutic research. The CAMOIP platform offers significant evidence that bridges the gap in information between cancer genome data and immunotherapy.

Furkan M. Torun et al. [124] developed an open-source ML tool called omics-learn for biomarker discovery. Genomic and proteomics dataset was used for the experiment. Python libraries were used to develop the tool and it can be downloaded using a local server. This tool used the XGBoost model for training the dataset. The visualization and web interface of omics-learn was built using StreamLit.

Salim Ghannoum et al. [125] presented an open-source pipeline named DIscBIO to identify the genes by using the transcriptomic data. The authors used two scRNA-seq datasets to demonstrate the pipeline capabilities. The first dataset contains circulating tumor cells from BRCA patients. The second was a cell cycle regulation dataset from myxoid liposarcomas. All of the analyses were accessible as notebooks with R coding, explanatory language, output data, and images. The pipeline was implemented in four steps including data preprocessing, cellular clustering, retrieving DEGs, and biomarker identification. In myxoid liposarcoma, DIscBIO worked well by defining a small subset of cells with potentially aggressive and stem-like properties.

Dongqiang Zeng et al. [126] developed a tool Immuno-Oncology Biological Research (IOBR) for the identification of gene signatures based on a multi-omics dataset. This tool provides batch analysis of the gene signatures and their association with lncRNA profiling, clinical phenotypes, genetic characteristics, and the signatures produced from single-cell RNA sequencing data. Moreover, this tool integrates deconvolution methodologies with various signature construction tools for the identification of gene signatures. This tool is freely available to use and it is an effective and flexible tool.

Huan Dong et al. [127] developed an Online Survival analysis web server for Diffuse Large Cell Lymphoma (OSdlbcl) for the identification of prognostic value for some specific gene. Clinical follow-up information and gene expr. profiles of

1100 samples were used from TCGA and GEO databases. Moreover, DM data was also used for prediction purposes. This tool will develop a Kaplan-Meier (KM) plot which will give p-value, HR, and log-rank for some specific gene symbol. OSdlbcl is a modern web server that combines public gene mutation data, gene expr., and clinical follow-up data to deliver prognosis assessments for the discovery of DLBCL biomarkers.

Harpreet Kaur et al. [128] developed a web server called HCCpred for the identification of both diagnostic and prognostic biomarkers from gene expr. dataset in HCC patients. Raw data was extracted from 30 studies and passed to feature extraction techniques. The extracted genes were then passed to model training which successfully identified three genes (FCN3, CLEC1B, and PRC1). Further survival analysis was done using univariate Cox models.

Amrit Singh et al. [129] presented a framework Data Integration Analysis for Biomarker discovery using Latent Components (DIABLO) using a multi-omics dataset. This tool was capable of identifying the biomarkers from both simulated and real multi-omics data. mixOmics was used to implement the tool. The presented framework successfully identified novel and existing biomarkers including mRNAs, miRNAs, Cpgs, proteins, and metabolites.

Dvir Netanelly et al. [130] developed a framework Profiler of Multi-omics data (PROMO) for analyzing, pre-processing, clustering, and visualizing the single omics and multi-omics data simultaneously. Further, this tool was also used for biomarker discovery and survival analysis. This tool consists of a package of multiple ML algorithms and statistical methods like t-test, and chi-square test for the analysis of the dataset. For biomarker identification, statistical tests were used for identification of DEGs which were further passed to Cox models for survival analysis. This tool fills the gaps of all the available tools and was easily available on the web.

Harpreet Kaur et al. [131] developed a tool called CancerLSP for the identification of biomarkers in LIHC. Genomic and epigenomic data, i.e., transcripts and Cpg methylation data was downloaded from the TCGA portal and passed to ML models (SVM, RF, NB, SMO, and J48). These algorithms were implemented in Weka which successfully identified 21 Cpg sites and 20 transcript profiles related to LIHC.

Sangaralingam et al. [132] presented O-miner which was a powerful online platform for combining and analyzing multi-omics data. The method aids in the discovery of important pathways and the prioritization of biomarkers in databases that include gene, transcriptome, and DM data, as well as clinical and biological

data. The pipelines created for the tool make use of Bioconductor packages, and statistical methods and run in R and Python environments.

Mickael et al. [133] developed a biomarker discovery tool called BioDiscML using multi-omics data comprising genomic, proteomic, and pathological datasets. BioDiscML followed a variety of ML algorithms to identify the optimal set of biomarkers. This tool has the advantage of using a vast range of ML classifiers within a completely integrated framework that often includes data pre-processing, making it easier for non-ML experts to complete their tasks.

Xiaoyu et al. [134] proposed an integrative analysis tool called iProFun for biomarker identification using Proteomic, CNA, and DM datasets. This tool was used on OV patients. Proteomic data was extracted using the iTRAQ (isobaric Tags for Relative and Absolute Quantification) method, CNA data using the Clinical Proteomic Tumor Analysis Consortium (CPTAC) data portal, and DM using the TCGA firehose portal. The collected data was preprocessed and integrated for further evaluation. The identified genes by the iProFun tool serve as a drug target for OV patients.

Zefang Tang et al. [135] developed a web server Gene Expression Profiling Interactive Analysis (GEPIA2) for biomarker identification by using the gene expr. dataset. GEPIA2 works efficiently for 84 cancer subtypes. In this, gene expr. quantification was extended from the gene level to the transcriptomic level. This tool also helps to classify the cancer based on different subtypes. This website is freely accessible and implemented using HTML, JavaScript, and PHP language.

Qiang Wang et al. [136] developed an online survival web server OScc for the validation of the prognostic biomarkers from gene expr. dataset. This tool was tested on 4 gene expr. datasets retrieved from GEO and TCGA platforms. This tool will generate a survival curve for the p-value, HR, and log-rank test. Based on the values achieved, treatment will be provided to the high-risk patients.

Yeongjun Jang et al. [137] developed a web application called Cancer Patient Stratification and Survival Analysis (CAPSAA) for the evaluation of predictive values of candidate biomarkers by dynamically visualizing the survival stratification for different subgroups of patients. The subgroups were made from gene expr., CNA and mutation data downloaded from TCGA coherent. Hierarchical clustering was done to divide the patients into subgroups. This tool is freely available and the source code is uploaded on the GitHub platform.

Magali Champion et al. [138] software algorithm AMARETTO for the identification of cancer genes by integrating gene expr., DM and CNV datasets. Then co-expressed target genes were connected to the driver genes which were known

as regulatory modules. Then these driver genes were converted into a network for the identification of cancer genes. AMARETTO was applied on patients from 11 different sites and it was considered as the best tool for gene identification.

Xie et al. [139] developed a repository MOBCdb for the integration of genetic, clinical, transcriptomic, and epigenomic results. The database was created to enable users to collect data from BRCA patients' SNV, gene expr., miRNA, and DM. An interface is available in MOBCdb for concurrently visualizing multi-omics data from different samples. This data was also subjected to a survival study using MOBCdb's survival module. MOBCdb aids precision medicine by detecting new markers in different subtypes of BRCA through its comprehensive web interface.

Mohammed et al. [140] developed a pipeline named CancerDiscover to predict classes of cancer and to identify the cancer biomarkers. The tool assists with normalization and offers various function filtering approaches to select the best-performing functions. High-throughput raw datasets can be analyzed automatically and reliably with CancerDiscover. Various models for identifying cancer types and subtypes can be created using the proposed integrative pipeline. CancerDiscover is an open-source platform that is free to download.

Jasmine Chong et al. [141] presented an update to MetaboAnalyst (version 4.0) for the analysis of metabolomic data. This tool has added four new features to previous version of MetaboAnalyst including real-time R command monitoring and show, as well as the introduction of the MetaboAnalystR kit, an MS Peaks to Pathways module for predicting pathway behavior from untargeted mass spectral data using the mummichog algorithm, a Biomarker Metaanalysis module for comprehensive biomarker recognition using multiple metabolomic datasets, and a Network explorer module which integrates transcriptomic, metagenomics, and metabolomics dataset. Based on the most recent evidence from the HMDB, the underlying knowledge bases (compound databases, metabolite collections, and metabolic pathways) have also been revised.

Chun-Jie Liu et al. [142] developed a web server GSCALite for the analysis of gene sets related to cancer. This tool includes the identification of DEGs from mRNA expression, CNV, DM, and SNP data and the survival analysis using these genes, detection of genomic variation along with survival analysis, cancer pathway activity analysis, and identification of drug sensitivity related to genes. GSCALite is a user-friendly web server for dynamic study.

The work done on the tools for biomarker identification in multi-omics data is discussed in Table 2.8.

Table 2.8: Tools for Biomarker Identification in multi-omics dataset

Year	Dataset	Tool	Technology	Link
2023 [122]	mRNA, metabolomic, proteomic, genetic, DM	IMOPAC	PHP, R, HTML, CSS3, Apache	http://www.hbpding.com/IMOPAC
2022 [123]	Gene Expr., mutation	CAMOIP	R/ Jupiter	https://www.camoip.net/CAMOIP/
2022 [124]	Genomic, Proteomic	Omics-Learn	Python	https://omiclearn.com/
2021 [125]	scRNA-seq	DIscBIO	R/ Jupiter	https://github.com/ocbe-uio/DIscBIO
2021 [126]	lncRNA,RNA, genomic	IOBR	R environment	https://github.com/IOBR/IOBR
2020 [127]	Gene Expr., Gene Mutation, DM	OSdlbcl	J2EE/ MySQL	https://bioinfo.henu.edu.cn/DLBCL/DLBCLList.jsp
2020 [128]	Gene Expr.	Web Server	Cloud	https://webs.iiitd.edu.in/raghava/hccpred/
2019 [129]	Genomic, Metabolome	mixOmics	R/Bioconductor	http://mixomics.org/
2019 [130]	Genomic, mRNA	PROMO	Matlab	http://acgt.cs.tau.ac.il/promo/
2019 [131]	Genomic + Epigenomic	CancerLSP	Weka	http://webs.iiitd.edu.in/raghava/cancerlsp/
2019 [132]	mRNA, genome, DM	O-miner	R/ Python	http://www.o-miner.org
2019 [133]	Genomic, Proteomic, pathological	BioDiscML	Java/ Weka	https://github.com/mickaelleclercq/BioDiscML
2019 [134]	CNA, DM, Proteome	iProFun	R environment	https://github.com/songxiaoyu/iProFun
2019 [135]	Gene expr., RNA Sequencing	GEPIA2	Javascript/ PhP/ Pearl	https://gepia2.cancer-pku.cn/#index
2019 [136]	Gene expr.	OSCC	R/ javascript/ cloud	http://bioinfo.henu.edu.cn/CESC/CESCList.jsp
2019 [137]	CNV, Gene Expr.,SM	CAPSAA	Clojure/ Fig Wheel	http://capssa.ewha.ac.kr/
2018 [138]	Gene expr., CNV, DM	AMARETTO	R programming	https://bitbucket.org/gevaertlab/pancanceramaretto
2018 [139]	Gene expr., SNP, DM, microRNA	MOBCdb	Perl, R, MySQL	http://bigd.big.ac.cn/MOBCdb/
2018 [140]	Gene Expr., Sequencing	CancerDiscover	WEKA, Affy R	https://github.com/HelikarLab/CancerDiscover
2018 [141]	Metabolome, transcriptome, metagenome	MetaboAnalyst 4.0	Prime faces/ R/ Google Cloud Server	https://github.com/xia-lab/MetaboAnalystR
2018 [142]	mRNA, CNV, SNP, DM	GSCALite	R scripts/ maftools	http://bioinfo.life.hust.edu.cn/web/GSCALite/

2.3 Challenges in Biomarker Identification

Some problems have been faced while performing the review of existing techniques for biomarker identification using omics and multi-omics data which are shown in Figure 2.1 and are described below.

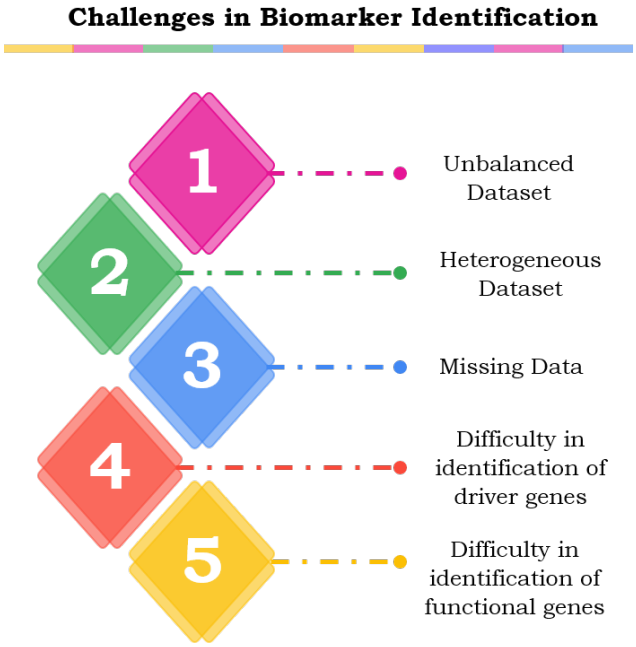


Figure 2.1: Challenges in Biomarker Identification [1]

- **Unbalanced dataset:** For biomarker identification, omics data including genome, transcriptome, protein, metabolites, and peptides are used. The available dataset is present in an unbalanced form. It means that the variables and attributes are too big than the sample size. This leads to an overfitting problem. Therefore, it is very difficult to identify biomarkers using an unbalanced dataset. This problem can be eliminated by integrating the different types of datasets and using that integrated dataset for biomarker identification. The feature extraction techniques called mRMR can also be employed to solve this problem [38, 63].
- **Heterogeneous datasets:** In biomarker identification, some of the molecular profiles are highly heterogeneous. They can be divided into categorical and continuous and sometimes may be scattered into multiple inputs. It makes the biomarker identification difficult. Therefore, different ML algorithms including graph network, clustering approaches, and DL techniques can be applied to remove heterogeneity [41, 47].

- **Missing Data:** In multi-omics biomarker identification, data missingness is a major challenge. Image noise, batch impacts, and hybridization failures all cause data missingness in microarray data [143]. Due to this complication, appropriate imputation of missed values based on practice, a mixture of methods, and trial and error is required. One of the most common ML algorithms i.e. KNN is used to impute the missing values. Instead, a median of the rows with missing data can be computed and imputed in place of the missing value.

- **Difficulty in the identification of important biomarkers:** There are different types of omics data. Sometimes it is not possible to identify the markers on the basis of a single type of data. For example, the genes can be identified using genomic data, but these may not be enough for disease prediction, survival prediction, and treatment/ response prediction. Therefore, integrated omics are required to identify the cancer biomarkers. Hence, multi-omics is required to identify the driver genes required for disease diagnosis and prognosis [33, 36].

- **Difficulty to identify functional genes:** Genomic data focus on DM data to identify mutations related to cancer. The DNA involves different changes starting from small somatic mutations, several insertions, deletions, and large copy number data for the identification of cancer mutations. The mutation further varies in different subtypes of cancer. Therefore, it is difficult to identify which function gene is growing the cancer. To solve this challenge, different DL techniques and gene prioritization algorithms are required [111].

2.4 Conclusion

This chapter discussed the review of techniques and tools used for biomarker identification using single and multi-omics with the help of ML and DL approaches. It sets the stage for the subsequent exploration of computational intelligent techniques for biomarker identification using multi-omics datasets. Additionally, the chapter acknowledges the need for appropriate approaches to be employed for biomarker identification due to the increasing complexity of multi-omics datasets resulting from the continuous collection of real data in omics data repositories. It unveils the potential for these techniques to address the identified challenges and open up new avenues for enhanced analysis and interpretation of biomarkers

using multi-omics data. In the next chapter, the framework developed using ML techniques for efficient biomarker identification in multi-omics data for disease diagnosis and prognosis is discussed.

Chapter 3
Proposed Framework

The literature survey carried out in the previous chapter clearly indicates the need to develop feature selection, machine learning (ML), and deep learning (DL) based framework for biomarker identification in multi-omics for disease diagnosis and prognosis.

The current chapter delineates the proposed framework aimed at fulfilling the research objectives outlined in the thesis. First, the requirement specifications required to develop the framework are discussed which is followed by the development of the framework proposed for biomarker identification in multi-omics for disease diagnosis and prognosis. The requirement specifications include the exhaustive study concerning hardware, software configuration, and their calibration for multi-omics data analysis and further use of that data for biomarker identification. The proposed framework includes six stages comprising data acquisition, data preprocessing, feature/ biomarker identification, biological interpretation, modeling, and performance evaluation.

Section 3.1 gives detailed information about the requirement specification. The minimum hardware and software requirements are described which is followed by the discussion of the proposed framework for biomarker identification in Section 3.2. Finally, Section 3.3 concludes the chapter.

3.1 Requirement Specifications

To address the objectives stated for the Thesis, the foremost requirement is to understand the required specifications for hardware and software. For the same, the specifications are described.

3.1.1 Software Requirements

The following tools are used to address the research objectives.
Anaconda (Spyder): Spyder is a freely available and open-source scientific com-

puting environment that is implemented in the Python programming language. It is specifically tailored to cater to the needs of scientists, engineers, and data analysts [144]. This software integrates the advanced functionality of a comprehensive development tool, including extensive editing, analysis, debugging, and profiling capabilities, with the data exploration, in-depth inspection, and visually appealing visualization capabilities of a scientific package, resulting in an unparalleled combination. This study utilizes Python and R language to implement the proposed frameworks. Several libraries for Python, such as *numpy, matplotlib, tensorflow, torch, pandas, keras*, and *os*, have been installed and utilized for experimentation purposes. The implementation of multi-omics data preparation is carried out using the R programming language, necessitating the use of RStudio as explained in the next section.

RStudio: R is widely used programming language for statistical computation and data visualization. It provides a wide range of statistical and graphical techniques and can be easily extended [145]. One notable advantage of the programming language R is its ability to effortlessly construct charts of high quality suitable for publishing, including the incorporation of mathematical symbols and formulas. The source code of the programming language R is distributed as free software, adhering to the conditions specified in the General Public License (GNU) provided by the Free Software Foundation. The RStudio Integrated Development Environment (IDE) is a comprehensive suite of integrated tools that have been specifically developed to optimize productivity in the programming languages R and Python. The software package comprises a console, an editor with syntax highlighting capabilities that facilitate immediate code execution, and a collection of comprehensive tools for tasks such as graphing, reviewing past actions, troubleshooting, and organizing one's workspace. In the current study, multi-omics data preparation is utilized within the RStudio environment. The installation of all necessary packages is accomplished by executing the install.packages() command. The installed packages utilized for model training and testing encompass limma, h2o tcgabiolinks, summerizedExperiment, dplyr, Hmeasure, tidyverse and caret. The process of normalizing the dataset is also implemented in the statistical programming language R through the utilization of the *scale* and *normalizebetweenarrays* functions. The hardware requirements required to implement the thesis objectives are discussed in the following section.

3.1.2 Hardware Requirements

System's Configuration: The minimum system configurations required for the Thesis are given as follows:

- Processor: Intel® Core™ i5 processor.
- Memory: 8 GB of RAM.
- Hard Disk: 20 GB of hard disk space required.
- Display: 1024 X 768 or higher-resolution display with 24-bit colors.

The following section details the methodology adopted for biomarker identification on the multi-omics dataset.

3.2 Framework for Biomarker Identification

Computational intelligent techniques comprising Machine Learning (ML) and Deep learning (DL) [146] have been adopted for the analysis of multi-omics data required for biomarker identification for disease diagnosis and prognosis. The workflow of the framework developed for the identification of biomarkers in multi-omics data is shown in Figure 3.1. It consists of data acquisition, data preprocessing, feature/ Biomarker identification, biological interpretation, modeling, and evaluation. The description of the above-mentioned stages is discussed in the following sections.

3.2.1 Data Acquisition

The multi-omics data required for biomarker identification is collected from various public repositories such as The Cancer Genome Atlas Portal (TCGA) [22], Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) [23] and Religious Orders Study and Rush Memory and Aging Project (ROSMAP) [24]. It involves the collection of a multi-omics dataset comprising genomics, transcriptomics, and proteomics which is generated by using high-throughput technologies. The complete description of the multi-omics repository used in this study is given in Table 3.1 and is described below.

3.2.1.1 TCGA

It is a prominent project with the largest collection of omics data encompassing 53 different cancer types and comprising 20,000 samples. This project includes gene

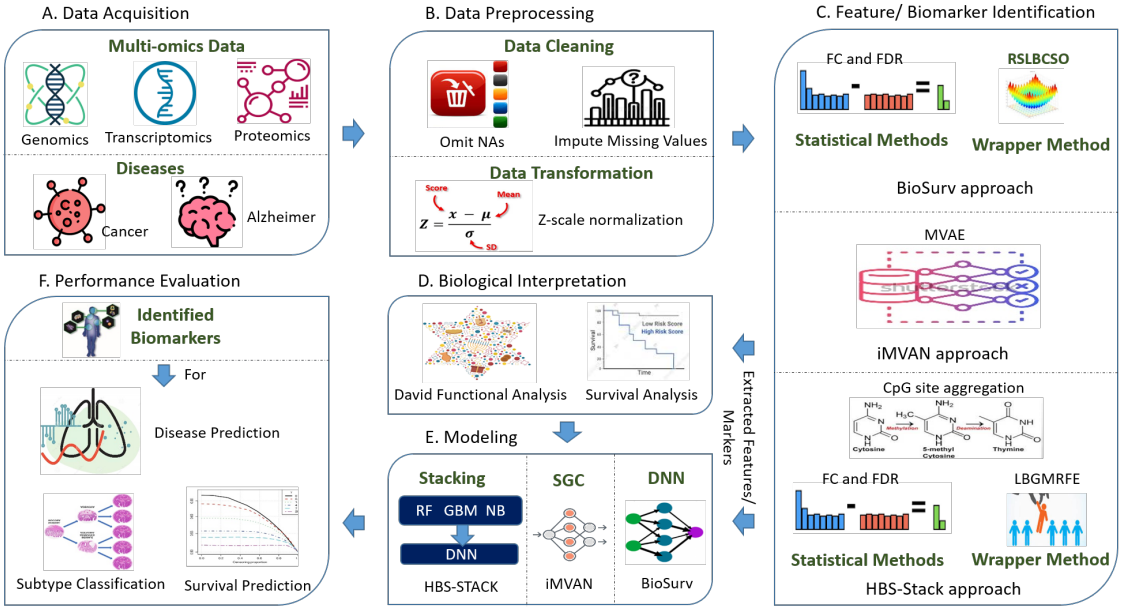


Figure 3.1: Workflow of proposed framework for biomarker identification in multi-omics

expression, DM, CNV, SNP, clinical and pathological image data. This initiative aims to generate, combine, analyze, and interpret profiles of cancer patients [22]. The publicly available TCGA data is extensively used to enhance cancer diagnosis, treatment, and prevention strategies. The multi-omics data is collected from the TCGA portal using the steps as follows:

- Access the TCGA portal using the link: `https://portal.gdc.cancer.gov/`.
- Select the omics data types (e.g., genomic, transcriptomic, proteomic, clinical, etc.) and specific types of cancer.
- Select the files to download which include processed data, raw sequencing data

Additionally, one web portal called LinkedOmics (`https://linkedomics.org/login.php`) is present which contains the pre-processed multi-omics datasets comprising genomic, transcriptomic, proteomic, and clinical data downloaded from the TCGA portal. The LinkedOmics platform offers a distinct opportunity for biologists and clinicians to conveniently access, evaluate, and compare multi-omics data related to cancer, both inside and across various types of tumors. Five types

Table 3.1: Description of multi-omics repositories

Repository	**Disease**	**Omics Types**
TCGA [22]	53 Types of Cancer	Genomic (DNA Methylation (DM), Copy Number Variation (CNV)), Transcriptomic (messenger Ribonucleic Acid (mRNA), micro RNA (miRNA), Gene Expression), Proteomic (Reverse Phase Protein Array (rppa)), Single Nucleotide Polymorphism (SNP), Clinical
METABRIC [23]	Breast Cancer (BRCA)	CNV, mRNA, DM, SNP, clinical
ROSMAP [24]	Alzheimer	mRNA, miRNA, DM

of cancer including breast carcinoma (BRCA), lung adenocarcinoma (LUAD), cervical and endocervical cancer (CESC), pan kidney coherent (KIPAN), and kidney renal carcinoma (KIRC) from TCGA have been utilized for disease survival prediction, disease subtype classification, and disease prediction and are discussed in Chapters 4, 5, and 6, respectively.

3.2.1.2 METABRIC

It is a Canada-UK project that includes gene expression, SNP, CNV, and clinical data of BRCA patients [23]. Using the underlying multi-omics biomarkers, this project intends to subclassify breast cancers into further groups. This database discovered 10 previously unidentified subtypes of breast cancer and new therapeutic targets, which will help in formulating the best course of treatment for breast cancer. The multi-omics data is collected from METABRIC using the steps as follows:

- Access the cBioPortal using link: `https://www.cbioportal.org/datasets/`.
- Select the specific breast cancer data type and molecular profile.
- Select particular samples and genes.
- Download the selected profile data along with clinical data.

METABRIC dataset has been utilized for validation of biomarker identification in multi-omics for disease survival prediction and is discussed in Chapter 4.

3.2.1.3 ROSMAP

The ROSMAP dataset is obtained via the Accelerating Medicines Partnership: Alzheimer's Disease (AMP-AD) Knowledge Portal [24]. The ROSMAP initiative comprises two distinct longitudinal clinical-pathologic cohort studies focused on Alzheimer's disease conducted at Rush University. These studies are known as ROS (Religious Orders Study) and MAP (Memory and Aging Project). It contains mRNA, miRNA, and DM multi-omics datasets of Alzheimer patients. ROSMAP is used as a validation dataset for biomarker identification in multi-omics required for disease prediction and is discussed in Chapter 6.

After data acquisition, data preprocessing is done to make multi-omics data suitable for feature selection/ biomarker identification and learning models. The data preprocessing is described in the following section.

3.2.2 Data Preprocessing

Data preprocessing is a technique employed to transform raw data into a format that is suitable and useful for analysis. Data preprocessing includes comprehensive data cleaning strategies for addressing null values and data normalization. The removal of null values, sometimes known as "NA" values, involves two main steps: feature filtering and sample filtering. Feature filtering involves the removal of features that exhibit a significant proportion of null values throughout the samples. On the other side, sample filtering involves the removal of samples that include a substantial proportion of null values across their features i.e. omitting the rows having more than 20% missing value. To handle the remaining missing values, null values imputations are required. Null value imputation entails the estimation of missing values using the information that is currently available. One of the main methods to impute the missing value is the K nearest neighbor method. KNN's non-parametric approach and its capability to handle both numerical and categorical data are notable strengths. Without presuming a specific data distribution, KNN can adapt well to diverse datasets, particularly those with complex or unknown distributions. Its ability to calculate similarity using distance metrics allows it to accommodate various feature types effectively, broadening its utility across different datasets and problem domains. These qualities render KNN a versatile and valuable tool for tasks like classification, regression, and missing value imputation across a spectrum of data scenarios. [147] The KNNimputer [148] function is used which searches for n neighbors surrounding the missing value (NA) and computes the Euclidean distance. The shortest distance value is imputed in

place of the missing value. This approach enables the retention of a higher volume of data while simultaneously enhancing the accuracy and effectiveness of the analysis. Further, data normalization of multi-omics data is undertaken with the objective of eliminating or rectifying inconsistencies in order to ensure the uniformity of data across different samples. Two main techniques, i.e., z-scale and min-max normalization (normalize between arrays) have been used for normalization. Z-score normalization, also known as standardization, transforms the data to have a mean of 0 and a standard deviation of 1. In min-max normalization, each feature's values are scaled to a fixed range, typically between 0 and 1. The min-max normalization is preferred in HBS-STACK because raw data is used directly. Preserving the original distribution of the data can be important in certain analyses, and min-max normalization allows you to do that while ensuring that the values are bounded within a specific range. This can be especially beneficial in scenarios where you need to maintain the interpretability of the data or when the original scale of the features is meaningful. Figure 3.2 shows the preprocessing steps for the analysis of multi-omics data required for biomarker identification and prediction purposes. As multi-omics data is a high-dimensional dataset, it is

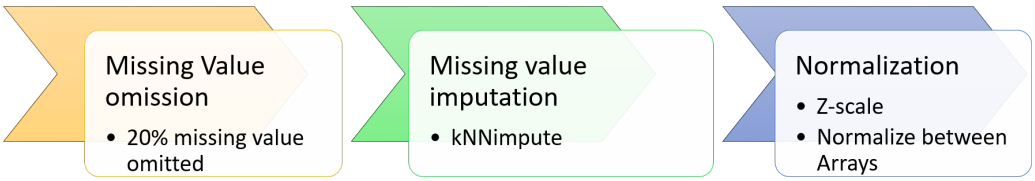


Figure 3.2: Steps of Data Preprocessing

challenging to handle the high dimensionality for accurate diagnosis and prognosis. Therefore, to select the highly important features/ biomarkers, feature/ biomarker identification techniques are required which is discussed in the following section.

3.2.3 Feature/ Biomarker Identification

Feature or biomarker identification is a method of choosing the most useful and informative features required for disease diagnosis and prognosis [149]. In bioinformatics, features are considered as biomarkers which are identified using feature selection, feature extraction, and statistical methods. In the current research, three approaches comprising BioSurv, iMVAN, and HBS-STACK have been proposed in which the feature selection, feature extraction, and statistical test have been used for feature/ biomarker identification. The summary of the proposed

techniques is shown in Figure 3.3 and is discussed in the forthcoming section.

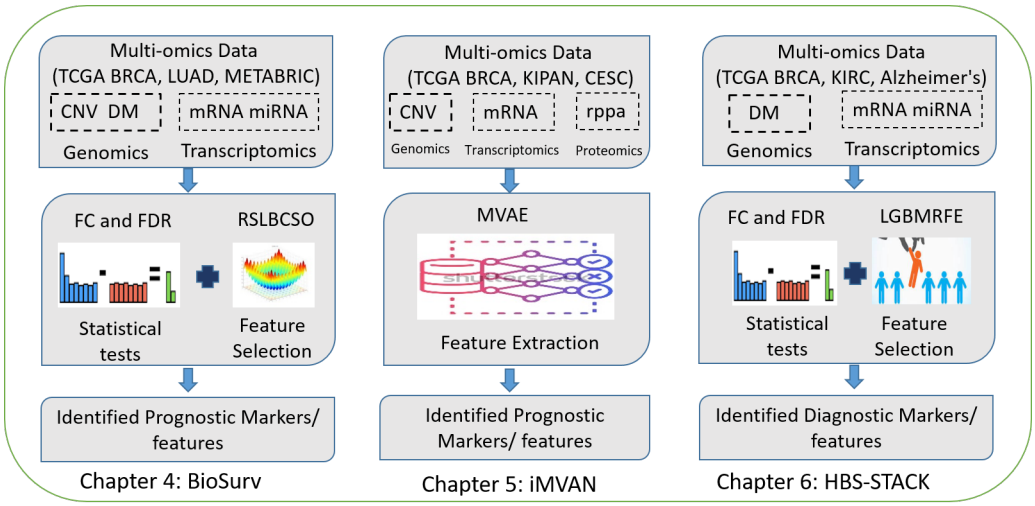


Figure 3.3: Proposed biomarker/feature identification approaches

3.2.3.1 Feature Selection

Feature selection is the process of selecting a subset of the most suitable and relevant features from the high-dimensional feature set in a dataset. Feature selection techniques are of 3 types including filter method, wrapper method, and embedded methods, and are shown in Figure 3.4.

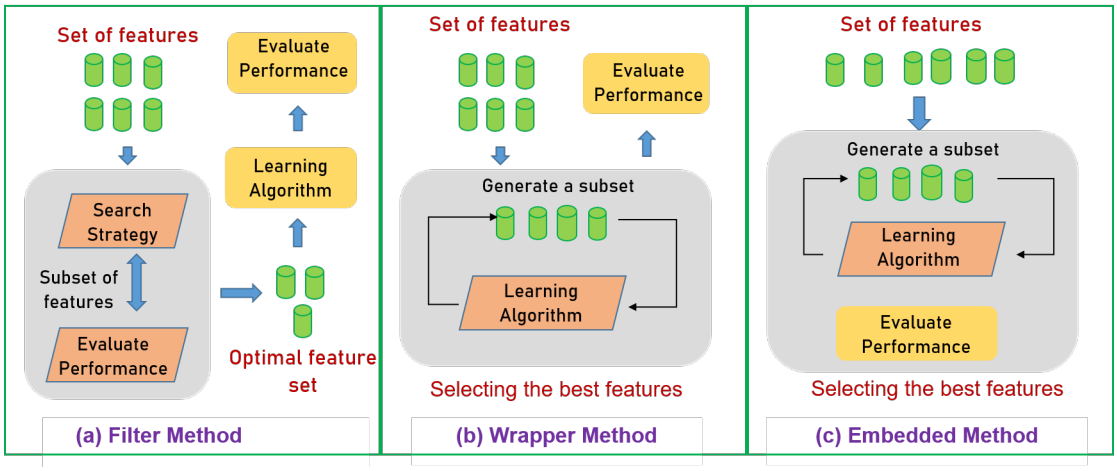


Figure 3.4: Feature Selection Methods [1]

Filter method: Filter method works by assigning a rank to the features and selecting only higher-rank features. There are different filter method techniques including Pearson Correlation Coefficient (PCC), chi-square (CHI2), t-test, and

Analysis of Variance (ANOVA) [150] which works by finding the correlation between the features and target variable. In biomarker identification, chi-square and t-test are used by various researchers to rank the differentially expressed genes and select the top-ranked genes.

Wrapper Methods: Wrapper methods work by selecting the features iteratively and evaluating their performance using a classifier. Initially, there is no feature set. Each time a feature is added and performance is checked. This is done until the most relevant features are not selected [151]. The wrapper method selects the features in two ways including forward feature selection and backward feature selection. Two wrapper methods comprising Cat swarm optimization (CSO) and Light Gradient Boosting Machine with Recursive Feature Elimination (LGBM-RFE) have been used for disease diagnosis prognosis and are described as follows.

- **Cat Swarm Optimization:** CSO [152] is an optimization algorithm inspired by nature and based on the collective behavior of cats. It was proposed in 2006 by Chuang-Bin Chiu, who drew inspiration from the predatory behavior of feral cats. Cats are renowned for their superior hunting abilities, which include individual and social behaviors like exploration, tracking, and coordination. The cats move and search in the solution space following their positions, velocities, and fitness values, which indicate the quality of their solutions. The algorithm updates the positions and velocities of the cats iteratively based on their fitness values and predefined principles, including random walks and social interactions. The velocity of the cats is updated using the following Eq. (3.1):

$$\begin{aligned} vel_i(j+1) = w * vel_i(j) + c_1 * r_1 * (per_i(j) - cur_i(j)) + \\ c_2 * r_2 * (pop_{gl}(j) - cur_i(j)) \end{aligned} \tag{3.1}$$

  Where $vel_i(j+1)$ denotes the updated velocity of cat i at time step $j+1$, $vel_i(j)$ is the current velocity of cat i at time step j, w represents the inertia weight required to control the influence of the previous velocity on the new velocity, c_1 and c_2 are the cognitive and social parameters, respectively, which control the influence of the personal best ($per_i(j)$) and global best ($pop_{gl}(j)$) solutions on the new velocity, r_1 and r_2 are random numbers uniformly drawn from the [0,1] range, $per_i(j)$ is the personal best solution of cat i at the time step j, which represents the best solution found by cat j so far, $pop_{gl}(j)$ is the global best solution among all the cats in the population at time step

j, which represents the best solution found by any cat in the population, $cur_i(j)$ is the current position of cat i at time step j, which represents the current candidate solution. Similarly, the position of the cat is updated using Eq. (3.2) as given below:

$$upos_i(j+1) = pos_i(j) + vel_i(j+1) \tag{3.2}$$

where $upos_i(t+1)$ is the updated position of cat i at time step $j+1$, $pos_i(j)$ is the current position of cat i at time step j, $vel_i(j+1)$ is the updated velocity of cat i at time step $j+1$, which is calculated using the Eq. (3.1). However, CSO may be biased towards exploitation, meaning it may focus more on intensively exploiting the areas of the search space around the global best position and local best position [153]. It may limit its ability to explore other regions of the search space. Therefore, a better strategy with a better balance between exploitation and exploration is required, allowing it to explore a more prominent search space and potentially find better solutions. Hence, a random spatial local best cat swarm optimization (RSLBCSO) is proposed to solve the limitation of basic CSO.

RSLBCSO adds more randomness and local search to the basic CSO algorithm to improve its ability to explore and exploit solutions to optimization problems. In RSLBCSO, each cat keeps track of its position and velocity in a multi-dimensional search space. The cats move around in the search space by changing their velocities based on their previous positions, the best positions the swarm found, and a set of random factors. The velocities are then used to change the positions of the cats, and their fitness is calculated. The random spatial local best update allows RSLBCSO to explore a more extensive search space and escape from local optima more effectively than CSO, which only uses the global best position as the reference for updating the local best positions. RSLBCSO have been developed for biomarker identification in multi-omics for disease survival prediction and is discussed in detail in Chapter 4.

- **LGBMRFE:** LightGBM [154] is a gradient-boosting framework that employs decision trees (DT) as the fundamental models for constructing a robust prediction model. The ensemble learning methodology referred to is a method that amalgamates the predictions generated by numerous DTs. The LightGBM framework employs a gradient boosting technique to iteratively adjust the weights of decision trees, aiming to minimize the loss function.

A gradient score is computed to generate a rank of the features to determine the contribution of the feature in the reduction of the loss function. The highly important features are extracted and passed to RFE for further feature identification. RFE is a widely employed technique in the field of feature selection, which aims to identify and retain the most significant features within a given dataset. The process involves iteratively training a machine learning model (LightGBM) and afterward removing the characteristics with the lowest importance. LGBMRFE feature selection is used for disease prediction and is discussed in detail in Chapter 6.

Embedded Methods: Embedded method combines the function of both filter and wrapper method. It works by integrating the feature selection algorithm with the training algorithm and selecting the feature subset [155]. Least Absolute Shrinkage and Square Estimator (LASSO) is one of the most common techniques of feature selection which is implemented by several researchers in biomarker identification.

3.2.3.2 Feature Extraction

Feature Extraction is the technique of reducing the feature space of high-dimensional multi-omics data to low-dimensional feature space [156]. This low-dimensional feature space consists of important information only required for biomarker identification, disease detection, and prognosis. There are different techniques available for the extraction of features for integrated omics, including Principal Component Analysis (PCA), Canonical Correlation Analysis (CCA), Linear Discriminant Analysis, Autoencoders, and Non-Negative Matrix Factorization (NNMF). One of the feature extraction techniques using autoencoders called multimodal variational autoencoder (MVAE) is developed for biomarker identification in multi-omics for disease subtype classification.

Multimodal Variational Autoencoders (MVAE): Variational Autoencoders (VAEs) [157] are deep generative models that can develop a meaningful data manifold from high-dimensional input data. There are three components to VAE: an encoder, a sampling module, and a decoder. Multimodal means combining data with different modalities. As the multi-omics data is a combination of different types of omics data including genomics, transcriptomics, and proteomics, therefore, multimodal variational autoencoders have been developed which after combining the multi-omics data are passed to encoders, sampling modules, and decoders to generate a latent data or a feature matrix. The detailed working of MVAE is discussed in Chapter 5.

3.2.3.3 Statistical Methods

Statistical approaches are employed to identify markers whose expression levels exhibit significant differences between two or more experimental conditions or groups. Two statistical techniques comprising Fold Change ($Log_2(FC)$) and False Discovery Rate (FDR) are used to identify the differentially expressed genes and markers and are described as follows.

- **False Discovery Rate (FDR)**: FDR [158] is a statistical term used in situations where multiple statistical tests are going on simultaneously, like in genomics, transcriptomics, bioinformatics, and other fields. The FDR is the number of false discoveries or wrongly rejected null hypotheses compared to the total number of findings or rejected null hypotheses. In other words, it shows how many false positives (FP) are expected from all the positives. A standard FDR threshold is 0.05, meaning no more than 5% of the discoveries will likely be FP. By adjusting the FDR, researchers can find a balance between finding significant results and making as few false discoveries as possible. The FDR can be calculated in several ways, such as with the Benjamini-Hochberg procedure, Storey's q-value, and Bonferroni correction. In the present research, the Student t-test [159] is performed to calculate the p-values. Then the Benjamini–Hochberg method is applied to calculate the FDR value on the significant p-value.

- **Fold Change (FC):** Fold change is a statistical measure that quantifies the relative change in a value between two conditions or groups [158]. Mathematically, fold change is calculated as the ratio of two values, typically the mean or median of a particular measurement in one condition or group (e.g., experimental group) divided by the mean or median of the exact size in another condition or group (e.g., control group) and is given by the Eq. (3.3) below:

$$FC = \frac{\frac{\sum_{i=1}^{n} x_g^n}{T}}{\frac{\sum_{i=1}^{n} y_g^n}{N}} \tag{3.3}$$

  where T represents the patients with short-term survivor, x_g^n is the g^{th} gene at n^{th} low survival sample, N is the long-term survivor sample, and y_g^n is the g^{th} gene at n^{th} high survival sample. In the current research, $log_2(FC)$ value is calculated and only those genes are selected using $|log_2(FC)|$ is greater than 0.5.

FDR serves a crucial role by controlling the proportion of false positives among

significant results. This is particularly significant in genomics studies, where numerous hypotheses are simultaneously tested. By curbing the risk of falsely identifying non-existent differences as significant DEGs, FDR ensures the reliability of findings. In contrast, Fold Change offers biologically meaningful insights into the magnitude of expression differences between conditions. While statistical significance is pivotal, Fold Change values convey the practical significance of gene expression alterations. They illustrate the extent of expression level differences between experimental conditions, aiding in the interpretation of biological implications. Together, FDR and Fold Change strike a balance between statistical rigor, biological relevance, and interpretability. They furnish researchers with robust tools to discern DEGs accurately while minimizing false positives and enhancing the understanding of gene expression dynamics. The FDR and FC methods have been used in biomarker identification for survival prediction and disease prediction in Chapters 4 and 6, respectively. Once the feature or biomarkers are identified, the validation is done to prove the significance which is discussed in the following section.

3.2.4 Biological Interpretation of Identified Biomarkers

Biological interpretation, within the realm of genomics and molecular biology, pertains to the act of attributing biological significance and meaning to experimental findings or data. This task entails comprehending the ramifications of experimental observations with regard to biological processes, functions, and mechanisms. The process of biological interpretation aids researchers in deriving significant findings from data obtained from many biological investigations, including genomics, transcriptomics, and proteomics. Different tools are used for biological interpretation comprising DAVID functional analysis and survival analysis and are discussed below.

3.2.4.1 DAVID Function Analysis

DAVID (The Database for Annotation, Visualization and Integrated Discovery) is a widely employed tool in the scientific community for investigating the biological context of gene lists and discerning the prevalence of biological terminology or pathways that may shed light on the activities and involvement of genes in distinct biological processes [160]. The primary purpose of utilizing DAVID is for the functional annotation and enrichment analysis of genes or proteins. The provision of functional annotations, gene ontology concepts, pathways, and other biologi-

cal features assists researchers in the interpretation of the biological significance of a given set of genes. The DAVID tool facilitates the analysis and interpretation of experimental findings by establishing connections between genes and their established or anticipated activities and relationships. The steps to biologically interpret the markers using the DAVID analysis tool are as follows:

- Go to David Functional analysis tool using `https://david.ncifcrf.gov/`.
- Go to start analysis and paste the list of identified markers.
- Choose the list type and submit the list as a new list.
- Select the new list and start the analysis of the extracted features.

The DAVID analysis has been utilized for validation of identified markers in Chapters 4, 5, and 6, respectively. The prognostic analysis can also done of the identified markers which is given in the following subsection.

3.2.4.2 Survival Analysis for prognostic biomarkers

Survival analysis is the process of responding to the occurrence of any event as required by our interests. The use of survival analysis has increased recently in several industries, including advertising, e-commerce, finance, and telecommunications, to determine when a client should make a purchase or when it is beneficial to exercise a stock option. Additionally, "survival" is used in biological systems to determine how long a patient survives after therapy and in mechanical systems to examine failure [161]. In medical situations, the event can be death, being alive, or the return of the illness. In medicine, the survival analysis method can identify even a patient's risk variables for survival. According to their length of survival, the patients in the study are classified as long-term survivors and short-term survivors. To identify the prognostic markers, three statistical tests comprising Kaplan Meier, Cox Proportional Hazard (CoxPH), and Concordance Index (CI) have been used which are described below:

Kaplen Meier (KM) Method: The KM estimator is widely utilized in survival analysis and is regarded as a non-parametric statistical method. This approach provides an estimation of the likelihood for a patient to survive beyond a designated time period. When the time is equal to zero, the method yields a probability of 1. As the time tends towards infinity, the method yields a probability of 0 [162].

Additionally, the KM Plotter tool [163] is utilized to construct KM survival curves and perform statistical analyses such as hazard ratio estimation and p-value

calculation. The steps to analyze the prognostic markers using KM Plotter are as follows:

- Go to KM Plotter tool using the link: `https://kmplot.com/analysis/`.
- Enter the identified genes and select the overall survival.
- Click on Draw Kaplen-Meier Plots.
- Select the poor and good prognostic markers based on Hazard Ratio and p-value.

CoxPH: CoxPH is employed to determine the chance of an event occurring, namely the survival of an individual beyond a certain time period. The utilization of the hazard function is essential for the purpose of comparing the survival rates among patients in this particular methodology. The CoxPH model is implemented in R using the "survival" package [164].
Concordance Index (CI): CI is a classification variable whose values range from 0 to 1, with 0 representing the worst value and 1 representing the best. Higher values of the concordance index indicate more excellent model performance. CI is computed using concordance.index function of the survcomp package [165].
The validation of identified prognostic markers has been done in Chapters 4 and 5, respectively. The validated markers along with the selected features are integrated and passed to modeling for performance evaluation which is discussed in the following section.

3.2.5 Modeling

The extracted features/ biomarkers are integrated using a concatenation-based integration and transformation-based integration. The detailed working of integration methods is given in Section 1.2.1.1. One of the transformation-based integrations is similarity network fusion (SNF). The SNF [166] method integrates many omics datasets, building a network for each, to provide a comprehensive view of the condition under investigation. By computing and combining Patient Similarity Networks (PSNs) for each data type, SNF is better than conventional single-data analysis methodologies. This facilitates the usage of complementary information from multi-omics data. SNF has been used for integration in Chapter 5. The concatenation-based integration is used in Chapters 4 and 6, respectively. The integrated features are then passed to modeling stage. Both ML and DL can be used in the modeling of datasets. The biomarkers can be identified for disease

prediction, survival prediction, and disease subtype classification. Various ML and DL comprising Naive Bayes (NB), Random Forest (RF), Gradient Boosting Machine (GBM), and Deep Neural Network (DNN) have been used for biomarker identification [167]. Using these models, three approaches have been proposed comprising BioSurv, iMVAN, and HBS-STACK. In BioSurv, DNN has been used for survival prediction whose parameters are optimized using Bayesian optimization. In iMVAN, DL method Simplified Graph Convolutional Network (SGC) has been proposed for disease subtype classification. In HBS-STACK, stacking of four ML models comprising NB, RF, GBM, and DNN has been proposed for disease prediction. The proposed approaches have been shown in Figure 3.5 and are described below.

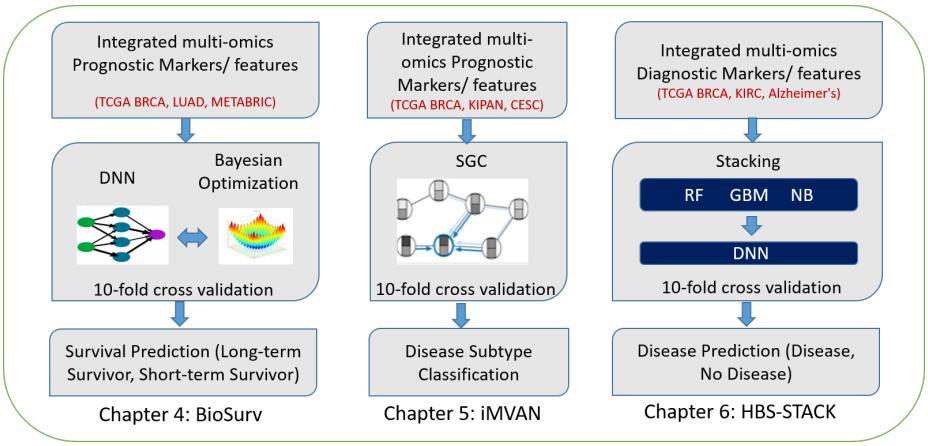


Figure 3.5: Proposed models using modeling phase of proposed framework

3.2.5.1 Naive Bayes (NB)

NB is a basic yet effective predictive modeling technique. Two kinds of probabilities can be derived from the training dataset directly. These probabilities are the probability of an individual class and the conditional probability given an x value for each class [168]. This Bayes theorem will be used to make new predictions on the test data when training is complete. A bell-shaped curve (Gaussian Distribution) is formed when the dataset used is in a real form which makes it easier to estimate the probabilities. We aim to select the best prediction/ hypothesis (h) from the given dataset (d). The hypothesis can be assigned the new class for the test dataset (d). To select the best hypothesis, we must have prior knowledge. Therefore, we use the Bayes theorem to identify the best hypothesis or best class by using prior knowledge. Bayes theorem is given by the following Eq. (3.4):

$$P(h|d) = \frac{(P(d|h) \times P(h))}{P(d)} \tag{3.4}$$

Where $P(hd)$ is a hypothesis (h) probability with the given dataset (d). This is also known as posterior probability. $P(dh)$ is the probability of dataset (d) when the given hypothesis (h) is true. $P(h)$ is the prior probability which is the probability of h being true. $P(d)$ is the dataset probability. Once the posterior probability for different hypotheses is calculated, the highest probability hypothesis is selected as the final result. This highest probability hypothesis is known as maximum posteriori (MAP), and it is inscribed as Eq. (3.5):

$$\text{MAP (h)} = \max(P(h|d)) \tag{3.5}$$

3.2.5.2 Random Forest (RF)

RF is a predictor that combines several decision trees on different subsets of data and averages the results to increase the dataset's prognostic accuracy. Instead of depending on a single decision tree, the RF collects the results from every tree and anticipates the output value relying on the majority voting rule. It can solve both regression and classification problems [169]. In classification, voting is performed, and the highest voted result is selected. In regression, the average of all the predictions from each tree is calculated and used as the final result. The structure of RF is shown in Figure 3.6.

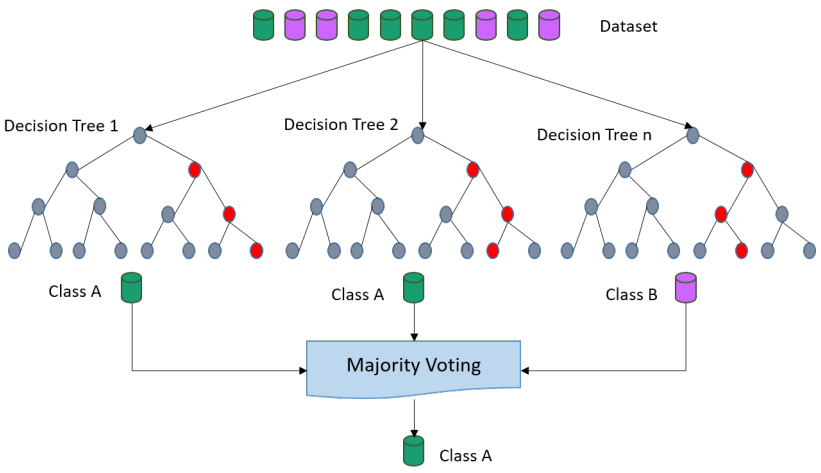


Figure 3.6: Structure of RF

3.2.5.3 Gradient Boosting Machine (GBM)

GBM works by combining different weak learners to build a strong learner. The weak learners correspond to different decision trees connected sequentially, where each tree tries to reduce the errors from the previous tree. The objective of GBM is to minimize the loss function. Loss is also known as the mean square error

(MSE) and is defined by Eq. (3.6):

$$\text{MSE} = \text{loss} = \sum (y_i - y_i^p)^2 \tag{3.6}$$

Where y_i is the target variable at the ith position, y_i^p is the prediction variable at the i^{th} position. $L(y_i, y_i^p)$ denotes the loss function [170]. The predictions should be in such a way that the loss function should be minimum. The gradient descent function can be used in which by changing the learning rate, we can the loss where it is minimum. It is given by the following Eq. (3.7):

$$y_i^p = y_i^p - \alpha \times 2 \times \sum (y_i - y_i^p) \tag{3.7}$$

Where α denotes the learning rate and $\sum(y_i - y_i^p)$ represents residuals sum. When the sum of the residuals is minimum or 0 or close to 0, the predicted value becomes close to actual values, which automatically reduces the lost function.

3.2.5.4 Deep Neural Network (DNN)

DNNs are feed-forward artificial neural networks (ANN) with multiple hidden layers of neurons used to perform various classification tasks. DNN performed well on text, voice, sounds, and other functions, which required innovative thinking. When a system employs multiple layers of nodes to extract high-level functions from incoming data, it is called a DNN [171]. It entails translating the facts into a more abstract and creative component. The structure of DNN is shown in Figure 3.7.

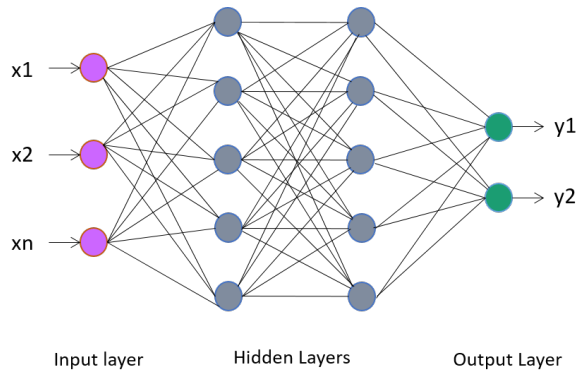


Figure 3.7: Structure of DNN

There is more than one hidden layer (h) present in DNN. Each input, hidden, and output layer contains some nodes called a neuron. The output of the i^{th} layer acts as an input for the j^{th} layer. The final output y is obtained by applying functional transformations and activation functions using some weights (wi, j) and

bias (b) values. Mathematically, it is given by the following Eq. (3.8):

$$y_i^h = f(\sum_{k=1}^{K}(w_(i,j)x_j + b_i)) \tag{3.8}$$

Where h is the hidden layers, x_j denotes the input at the j^{th} layer, $w_(i,j)$ represents the weights, and b denotes the bias value. The activation function used in the equation is non-linear. Different activation functions are there, like tanh, Rectified Linear Unit (ReLU), and sigmoid, which can be used to perform different computations. DNN has been utilized in biomarker identification in multi-omics for survival prediction and is given in Chapter 4.

3.2.5.5 Graph Convolutional Neural Network (GCN)

GCNs [172] accept as input a graph that already has some of its nodes labeled, and then they make label predictions for the rest of the graph's nodes. The structure of GCNs is quite complex. Therefore, the GCNs are simplified by developing a simplified graph convolutional networks (SGC). SGC sequentially reduces the nonlinearities and collapses weight matrices between succeeding layers. It does this by smoothing the node input features by employing powers of the normalized adjacency matrix in conjunction with self-loops. SGC has been developed for biomarker identification in multi-omics required for disease subtype classification. The detailed working of GCN and SGC is given in Chapter 5.

3.2.5.6 Stacking

Stacking works by using multiple heterogeneous weak models at first-level training to train only a portion of the problem but not the whole problem. The stacking has been proposed for biomarker identification in multi-omics required for disease prediction and detailed in Chapter 6. The training of the weak learner is done in parallel. Hence, different base learners have been built, which can be used to make first-level or intermediate predictions. Afterward, a new model called meta-model or meta learner is added, which will make predictions on the same class variable by considering the intermediate predictions as features. The GBM, RF, and NB have been used as base-learners and DNN has been used as a meta-learner.

3.2.5.7 Hyper-parameter tuning

Hyperparameter tuning is used to fine-tune a model's performance by adjusting its hyperparameters. There are three types of hyperparameter tuning comprising

grid search, manual search, and bayesian optimization.

Grid Search is the most fundamental method for tuning hyperparameters. A grid of hyperparameter values is defined. The tuning algorithm sequentially performs an exhaustive search of this space and trains a model for each possible combination of hyperparameter values [173].

Random Search: Random search differs from grid search in that values are sampled from a statistical distribution for each hyperparameter [174]. A sampling distribution for each hyperparameter to conduct a random search is defined. The number of utilized hyperparameter combinations can be controlled or limited using a random search.

Bayesian Optimisation: Bayesian optimization is a prominent method for optimizing expensive-to-evaluate black-box functions [175]. Bayesian optimization integrates statistical models [176], typically Gaussian processes, with acquisition functions to guide the search for the optimal solution.

The bayesian optimization have been employed to tune the parameters of a DNN model and is given in detail in Chapter 4. Additionally, it is common practice to do cross-validation in conjunction with model training in order to evaluate how effectively a model generalizes to data that it has not previously encountered which is discussed in the following section.

3.2.5.8 Cross Validation

Cross-validation refers to the procedure of evaluating the efficacy of algorithms through the partitioning of the dataset into two distinct subsets. One portion of the data is allocated for the purpose of training the model, while the other portion is reserved for validation [177]. The purpose of cross-validation is to ensure that each component of the original dataset has an equal opportunity to be included in both the training and testing sets. The final result obtained from the cross-validation process is utilized to assess the robustness of the model. K-fold cross-validation is a commonly employed technique for the purpose of validation. In the context of k-fold cross-validation, the validation dataset is partitioned into k subsets, with $k-1$ subsets utilized for training the model and the remaining subset employed for assessing the performance of the trained model. This process is iterated multiple times until all components have undergone both the training and testing phases. The final results are produced by averaging performance parameters at the conclusion of iterations. If the obtained outcomes closely resemble or are equivalent to the training outcomes, it can be inferred that the models are functioning accurately. Figure 3.8 shows the working of K-fold cross-validation.

One of the primary benefits of employing this methodology is that each individual data point is included just once during the validation process, resulting in a reduction of bias and variance in the overall performance of the model.

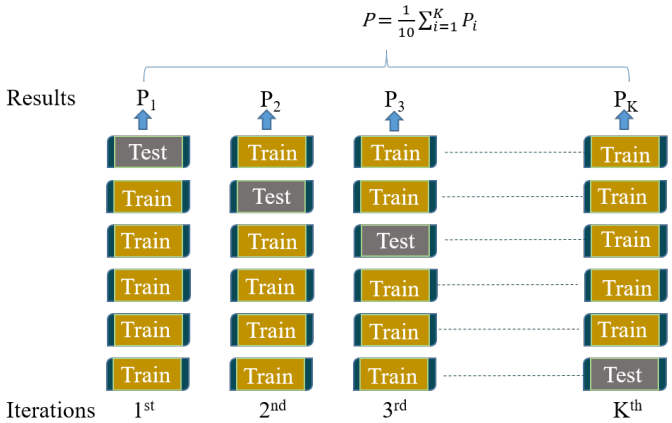


Figure 3.8: K-fold Cross Validation

Cross-validation has been employed for biomarker identification in multi-omics for disease survival prediction, disease sub-type classification, and disease prediction in Chapters 4, 5, and 6, respectively. To evaluate the performance of trained models various performance parameters have been used which are discussed in the following section.

3.2.6 Performance Evaluation

The performance of biomarkers identification for disease prediction, survival prediction, and subtype classification is done using the following parameters.

- **Accuracy:** The accuracy metric is determined by dividing the number of properly predicted observations by the total number of observations, as represented by the following Eq. (3.9).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.9}$$

  Where TP stands for True Positive, TN for True Negative, FP for False Positive, and FN for False Negative

- **Sensitivity:** Sensitivity gives the rate of correctly identified positive instances. The calculation involves determining the proportion of accurately

anticipated positive observations in relation to the overall number of positive observations and is given by Eq. (3.10).

$$Sensitivity = \frac{TP}{TP + FN} \tag{3.10}$$

- **Specificity:** Specificity gives the true negative rate. The metric is determined by dividing the number of falsely anticipated positive observations by the total number of negative observations. It is given by the following Eq. (3.11).

$$Specificity = \frac{TN}{TN + FP} \tag{3.11}$$

- **Precision:** Precision is determined by dividing the number of accurately anticipated positive observations by the total number of positive observations made and is represented mathematically by an Eq. (3.12).

$$Precision = \frac{TP}{TP + FP} \tag{3.12}$$

- **F1-score:** The F1-score can be defined as the mathematical average of precision and recall, specifically calculated using the harmonic mean. The algorithm effectively manages the trade-off between precision and recall, rendering it particularly valuable in scenarios where there exists a class imbalance. The mathematical expression for the F1-score is given as Eq. (3.13).

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

- **Matthews Correlation Coefficient (MCC):** It is the best metric to use one value to show true and false negatives and positives in a confusion matrix. The MCC metric is computed using the following Eq. (3.14).

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN) \times (TP + FP) \times (TN + FN) \times (TN + FP)}} \tag{3.14}$$

- **Area Under ROC Curve (AUROC):** The receiver operating characteristic curve (ROC) is used to show the performance of the model using the true positive rate (TPR) and false positive rate (FPR) at all thresholds. The area under the ROC plot is calculated and is termed AUC.

Based on the proposed framework, three approaches have been developed comprising BioSurv, iMVAN, and HBS-STACK for biomarker identification in multi-omics

for survival prediction, subtype classification, and disease prediction. The detailed working of BioSurv, iMVAN, and HBS-STACK is given in Chapters 4, 5, and 6, respectively.

3.3 Conclusion

This chapter discussed the hardware and software requirements and proposed framework for multi-omics biomarker identification for disease prognosis and diagnosis using ML and DL approaches. The six phases are provided including data acquisition, data preprocessing, feature/ biomarker identification, biological interpretation of identified markers, modeling, and performance evaluation. Based on the proposed framework, three approaches comprising BioSurv based biomarker identification for survival analysis, iMVAN based biomarker identification for disease subtype classification, and HBS-STACK based biomarker identification for disease prediction have been developed on the multi-omics dataset. In BioSurv, the first statistical tests comprising FC and FDR have been employed for feature identification which are then passed to RSLBCSO for selecting the most optimized features. The extracted features from each omic type are integrated and passed to Bayesian optimized DNN for survival prediction. In iMVAN, a multimodal variational autoencoder (MVAE) is developed for biomarker identification. The fusion of multi-omics datasets is done using similarity network fusion (SNF). The output of MVAE and SNF is integrated and passed to SGC for disease subtype classification. In HBS-STACK, a hierarchical biomarker selection (CpG sites aggregation, statistical tests, and LGBMRFE) is proposed to identify the biomarkers, which are then passed to a stacked ensemble for disease prediction. The biomarkers identified from BioSurv, iMVAN, and HBS-STACK are validated using DAVID and KM plotter analysis. The performance is evaluated using performance parameters comprising accuracy, sensitivity, specificity, F1-score, MCC, and AUC. The outlined proposed framework comprises four consistent phases: data acquisition, data preprocessing, biological interpretation, and performance evaluation across all approaches. However, the feature/biomarker identification and modeling phases vary for each approach, i.e., BioSurv, iMVAN, and HBS-STACK, respectively. The detailed working of BioSurv, iMVAN, and HBS-STACK is presented in the forthcoming sections.

Chapter 4

BioSurv: Biomarker Identification for Survival Analysis

The previous chapter demonstrates the proposed framework for the identification of biomarkers for disease diagnosis and prognosis in multi-omics data. The hardware and software requirements and the computational methodology using ML and DL approaches are discussed. By following the proposed framework, a BioSurv approach for biomarker identification in multi-omics for survival prediction is presented.

In this chapter, a detailed presentation and discussion of the BioSurv approach based on the proposed framework developed using a Random Spatial Local Best Cat Swarm Optimization (RSLBCSO) for Biomarker identification and Bayesian Optimized Deep Neural Network (DNN) for survival prediction is provided. The approach called BioSurv, is trained and evaluated in multi-omics data specifically related to breast carcinoma (BRCA) and lung adenocarcinoma (LUAD).

Section 4.1 focuses on the discussion of biomarker identification and survival prediction using the proposed RSLBCSO and Bayesian Optimized DNN. Section 4.2 presents the experimental analysis and results, while the statistical analysis showing the effectiveness of the proposed BioSurv is discussed in Section 4.3. Finally, the chapter concludes with Section 4.4.

4.1 Overview of BioSurv

The BioSurv approach is proposed for biomarker identification for disease survival prediction in multi-omics datasets of Breast Carcinoma (BRCA) and Lung Adenocarcinoma (LUAD) patients using a novel Random Spatial Local Best Cat Swarm Optimization (RSLBCSO) and Bayesian optimized Deep Neural Network. The BioSurv is developed using five phases, as shown in Figure 4.1. It consists of Data collection, Data Preprocessing, Feature selection, Biological Interpretation and Modeling.

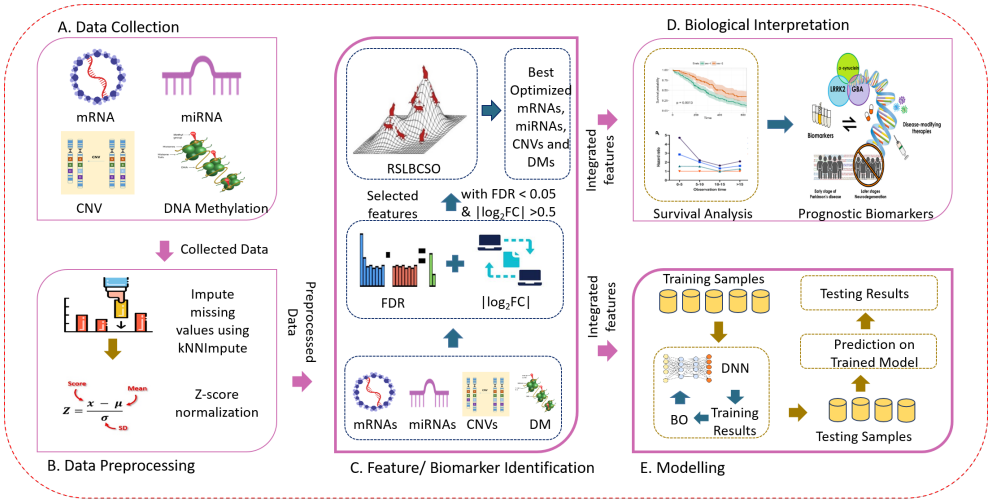


Figure 4.1: Workflow of the BioSurv [2]

Phase 1. Data collection: The data utilized in this study is sourced from the TCGA (The Cancer Genome Atlas) portal [22], which provides a comprehensive collection of real cancer data, encompassing genomic (DNA Methylation (DM), Copy Number Variation (CNV)), transcriptomic (micro Ribonucleic acid (miRNA), messenger Ribonucleic acid (mRNA)), proteomic (Reverse Phase Protein Array (rppa)), metabolomics and clinical datasets. TCGA is a prominent cancer genomics program encompassing 53 different cancer types and comprising a vast repository of 20,000 samples. Researchers extensively rely on the publicly available TCGA data for biomarker identification and to enhance cancer diagnosis, treatment, and prevention strategies. The complete description of TCGA is given in section 3.2.1.1

Phase 2. Pre-processing: After collecting the multi-omics data, a thorough cleaning process is carried out to ensure its quality and reliability. The cleaned data is then passed through a pre-processing pipeline, where it undergoes various transformations and formatting procedures to prepare it for analysis.

Phase 3. Feature Selection: A feature selection algorithm including statistical tests and RSLBCSO is employed to identify the most important features/ biomarkers used for training and testing the data set. This algorithm effectively selects the most relevant and informative features/ biomarkers, enhancing the accuracy and performance of the subsequent analysis.

Phase 4. Biological Significance: The selected features/ biomarkers from each omics type are then integrated and passed to KEGG (Kyoto Encyclopedia of Genes and Genomes) analysis and survival analysis for the identification of prognostic biomarkers.

Phase 5. Modeling: The modeling phase involves training and testing of selected machine learning (ML) models. To ensure optimal performance, the hyperparameters of these models are optimized using the Bayesian optimization technique. The prediction models are then trained using the optimized parameters. During the training phase, a dedicated training set is utilized to train the models. Finally, a testing set is employed to make predictions using the trained model, evaluating its accuracy and effectiveness. The performance of the trained ML models is assessed by evaluating several key parameters, including accuracy, sensitivity, specificity, precision, AUC, and concordance index (CI) value. These metrics serve as important indicators for prediction and survival analysis. By analyzing and comparing these performance measures, the effectiveness and reliability of the models in predicting and analyzing survival outcomes can be accurately determined. The phases of biomarker identification and disease survival prediction are described in the following sections.

4.1.1 Data Collection

Cancer is a multifaceted and diverse illness that deregulates cellular activities on various molecular levels, comprising RNA, DM and CNV which can lead to the development of many types of cancer. It is essential to note that molecules from various levels are related to one another in reprogramming the cell's activities [178]. Every year billions of people are affected due to cancer. The low survival rate of the patients leads to the need for accurate identification of biomarkers for survival prediction to improve the quality of life and opt for personalized treatment of cancer patients [179].

The four types of omics datasets, including mRNA and miRNA at the transcriptomic level and DM, and CNV, at the genomic level for BRCA and LUAD patients have been used to identify biomarkers and predict the survival of cancer patients. As it is a survival analysis task, clinical data is required to determine the patient's status (dead or alive), and time (survival time). The dataset is downloaded from the Linked Omics Portal [180], which contains multi-omics datasets from TCGA. This downloaded dataset consists of a different number of samples for BRCA and LUAD cancers, i.e., BRCA contains 1093 mRNA, 755 miRNA, 1080 CNV, 783 DM, and 1097 clinical samples, and LUAD comprises 515 mRNA, 450 miRNA, 516 CNV, 458 DM, and 522 clinical samples. Therefore, to find the common samples, the Venn diagram is used, and finally, 616 samples for BRCA and 435 samples for LUAD have been obtained. The five-year survival is used, and the study is divided into two classes comprising short-term survivors and long-term

survivors. The short-term survivors are labeled 0, and the long-term survivors are labeled 1. The detailed summary of the BRCA and LUAD datasets with samples and labels is given in Table 4.1.

Table 4.1: Description of Dataset

Dataset	**BRCA**	**LUAD**
Total no. of samples	616	425
Cut-off years	5	5
Long-term survivors	130	64
Short-term survivors	486	361
Median Survival	38.55 months	29.98 months

4.1.2 Data Preprocessing

Data preprocessing is the process of cleaning, transforming, and preparing raw data into a suitable format for further analysis or modeling. Data preprocessing is crucial because the quality and accuracy of the data used in the study directly impact the results and performance of the final output. The detailed working of data preprocessing is given in section 3.2.2. The total number of features in mRNA, miRNA, DM, and CNV are 20155, 823, 335855, and 24776 for BRCA and 19988, 809, 336284, and 24776 for LUAD respectively. To preprocess the dataset, first, the missing values (NAs) from each mRNA, miRNAs, DMs, and CNVs are imputed using the K Nearest Neighbor Impute (KNNImputer) function [148]. Then, to make consistency in the values of the omics dataset, normalization is required. Therefore, z-scale normalization is performed on mRNA, miRNA, and DM datasets. The values in the CNV dataset are in the form of -2, -1, 0, 1, and 2, hence, these values are used directly without any normalization. However, many features exist, so feature selection techniques are applied to reduce the features. The techniques are described in the following subsection.

4.1.3 Feature Selection

It is a challenging task to identify the features from high-dimensional omics data required for prediction purposes. The dataset used in this research is high-dimensional. It can lead to the curse of dimensionality if not treated properly [181]. Additionally, more features compared to a total number of samples can

sometimes lead to overfitting problems and poor performance. To solve these challenges, feature selection is employed to reduce the feature space while selecting the most important/ relevant features. In this study, the feature selection is performed in two stages: first, the statistical analysis test is performed, and then features are optimized using RSLBCSO. The complete detail is discussed below:

4.1.3.1 FDR and $log_2(FC)$

Two statistical analysis tests comprising FDR and $|log_2(FC)|$ are performed to reduce the dimensionality. The complete detail of FDR and $log_2(FC)$ is given in Section 3.2.3.3. The tests are applied on mRNA, miRNA, DM, and CNV datasets of BRCA and LUAD cancer and select only those features with $FDR < 0.05$ and $|log_2(FC)| > 0.5$. These methods return 2332 mRNAs, 39 miRNAs, 1176 CNVs, and 2112 DMs for BRCA and 1304 mRNAs, 26 miRNAs, 983 CNVs, and 1828 DMs for LUAD samples. This is still a huge feature set that needs to be reduced for training and testing. Therefore, a swarm optimization technique RSLBCSO is proposed, which returns the most optimized features for each data type. The RSLBCSO is described in the following subsection.

4.1.3.2 Random Spatial Local Best Cat Swarm Optimization (RSLBCSO)

RSLBCSO adds more randomness and local search to the basic CSO algorithm to improve its ability to explore and exploit solutions to optimization problems. The CSO algorithm is described in subsection 3.2.3.1. In RSLBCSO, each cat keeps track of its position and velocity in a multi-dimensional search space. The cats move around in the search space by changing their velocities based on their previous positions, the best positions the swarm found, and a set of random factors. The velocities are then used to change the positions of the cats, and their fitness is calculated. This random spatial [182] local best update allows RSLBCSO to explore a more extensive search space and escape from local optima more effectively than CSO, which only uses the global best position as the reference for updating the local best positions. This can potentially lead to better convergence and higher-quality solutions in RSLBCSO. Mathematically, the velocity update equation in RSLBCSO is represented by Eq. (4.1) as follows:

$$\begin{aligned} vel_i(j+1) = w * vel_i(j) + c_1 * r_1 * (per_i(j) - cur_i(j)) + \\ c_2 * r_2 * (pop_{gl}(j) - cur_i(j)) + r_3 * (dsp_i(j) - cur_i(j)) \end{aligned} \tag{4.1}$$

Where r_3 is a randomly generated number drawn from [0,1] range, $dsp_i(j)$ is a randomly generated local best spatial position within a specific neighborhood of

cat i at time step j, which promotes exploration in the search space. The additional term $r_3 * (dsp_i(j) - cur_i(j))$ introduces a random spatial local best component to the velocity update equation. This promotes exploration by adding a random spatial displacement to the current position of the cat, allowing it to explore new areas in the search space beyond its personal and global best solutions. Following the Eq. 4.1, the position of cats is updated and given by the following position update Eq. (4.2):

$$upos_i(j+1) = pos_i(j) + vel_i(j+1) + r_4 * (dsp_i(j) - pos_i(j)) \tag{4.2}$$

Where r_4 is a randomly generated number drawn from the [0,1] range. Figure 4.2 and Algorithm 4.1 give a complete description of RSLBCSO.

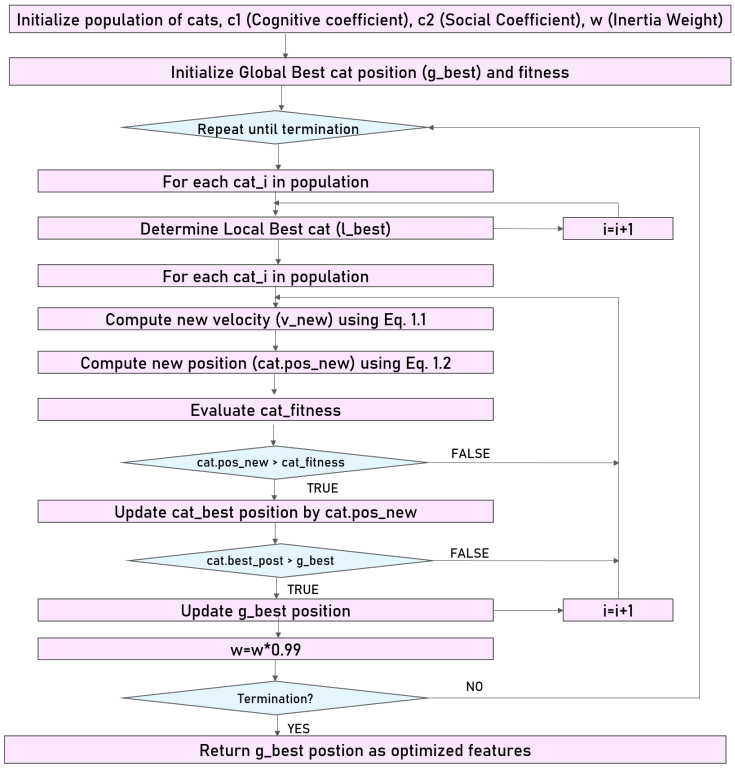


Figure 4.2: RSLBCSO flowchart [2]

The RSLBCSO begins with initializing the population, variables, global best position, and fitness, followed by a for loop for $n_{iterations}$. Next, the loop run for each cat in the population in which the local best cat is determined. This is used in the update velocity equation. Then, again a for loop is called in which the new velocity (v_new) and new position (cat.pos_new) are computed. This is followed by an evaluation of fitness function (cat_fitness). Then an if loop is called, which will check whether the cat.pos_new is greater than cat_fitness or not. If it is

Algorithm 4.1 Algorithm of RSLBCSO

INPUT X_{train} (Training feature matrix), X_{test} (Testing feature matrix), y_{train} (Training labels), y_{test} (Testing labels), n_{cats} (Number of cats in the swarm), $n_{\text{iterations}}$ (Number of iterations), *lbest_cat* (Local best cat) c_1 (Cognitive coefficient), c_2 (Social coefficient), and w (Inertia weight)

OUTPUT *g_best* (Global best feature subset)

BEGIN

X_{train}, X_{test}, y_{train}, y_{test}, n_{cats}, $n_{\text{iterations}}$, c_1, c_2, w
$n_{\text{samples}}, n_{\text{features}} \leftarrow X_{\text{train}}$
swarm $\leftarrow$ initialize_swarm($n_{\text{cats}}, n_{\text{features}}$)
g_best_pos $\leftarrow$ get_g_best(swarm, $X_{\text{train}}, X_{\text{test}}, y_{\text{train}}, y_{\text{test}}$)
for $i \leftarrow 1$ to $n_{\text{iterations}}$ **do**
    for cat $\leftarrow$ swarm **do**
        Determine lbest_cat based on cat.best_fitness
    end for
    for cat $\leftarrow$ swarm **do**
        v_new $\leftarrow$update_velocity(cat, global_best_pos, c_1, c_2) using Eq. 4.1
        cat.pos_new $\leftarrow$ update_position(cat) using Eq. 4.2
        cat_fitness $\leftarrow$ fitness(cat.pos_new, $X_{\text{train}}, X_{\text{test}}, y_{\text{train}}, y_{\text{test}}$)
        if (cat.pos_new $>$ cat_fitness) **then**
            cat.best_pos $\leftarrow$ cat.pos_new
        end if
        if (cat.best_pos $>$ global_best_pos) **then**
            g_best $\leftarrow$ get_g_best(swarm, $X_{\text{train}}, X_{\text{test}}, y_{\text{train}}, y_{\text{test}}$)
        end if
        $w \leftarrow w \times 0.99$ ▷ Update inertia weight
    end for
end for
g_best_pos as the optimized features

End

greater, then the cat_best position is updated by cat.pos_new. Otherwise, it will go looking for the next cat. Then, it will check whether the cat.best_pos is greater than g_pest. If it is greater, then the new g_best position is computed; otherwise, it will go for the next cat. Once, all the cats in the population are traversed, the inertia weight (w) is multiplied by 0.99. Multiplying the inertia weight by 0.99 in each iteration gradually reduces its value over time, decreasing the algorithm's

exploration capability and increasing the exploitation capability. As the algorithm progresses, particles are more likely to exploit the local and global best solutions found so far, leading to a more focused search around these solutions. This can help the algorithm to converge faster and find better solutions in the later stages of the optimization process. The value of 0.99 is chosen empirically and can be tuned based on the specific problem being solved. The fitness function calculates the fitness of a cat's current position by training a DNN classifier on the training data using the subset of features represented by the cat's current position and then computing the accuracy of the classifier on the test data. The frequency is also computed, which tells how often a feature is selected. The RSLBCSO returns the most optimized mRNAs, miRNA, DMs, and CNVs having feature counts of 102, 20, 86, and 94 for BRCA and 86, 18, 69, and 89 for LUAD, respectively. The complete details of feature extraction are given in Table 4.2. Further, the KEGG and survival analyses are performed to identify the prognostic biomarkers, which are discussed in the following section.

Table 4.2: Extracted features after Statistical test and RSLBCSO

Dataset	**BRCA**			**LUAD**		
	A	**B**	**C**	**A**	**B**	**C**
mRNA	20155	2332	102	19988	1304	86
miRNA	823	39	20	809	26	18
CNV	24776	1176	94	24776	983	89
DM	335855	2112	86	336284	1828	69

A: Features before extraction, B: Features after statistical analysis, C: Features after RSLBCSO

4.1.4 Biological Interpretation

Biological interpretation enables the derivation of information pertaining to fundamental biological processes and the etiology of genetic disorders. Different tools are there to understand the biological activity going on in the human body. By utilizing these tools and employing biological data mining techniques, it is possible to transform sequence data into valuable insights [183]. The tools used are DAVID and Survival Analysis which are discussed in detail in Section 3.2.4. First, the extracted features from RSLBCSO are validated using KEGG analysis from DAVID function enrichment analysis tool [160] for biomarker identification. This tool helps to find the biological meaning behind the genes. A total of 28 mark-

ers enriched in various pathways, cell adhesion, cell growth, and cell proliferation have been identified. Further, the identified biomarkers are passed to the survival analysis test, identifying the poor and good prognostic markers. 5-year survival analysis for both BRCA and LUAD is performed. The Kaplan Meier (KM) plots have been used to show the expression level of genes. The Cox Proportional Hazard (coxPH) model is used to compute the Hazard Ratio (HR) and p-value. Moreover, the extracted features from RSLBCSO are integrated and passed to Bayesian optimized DNN for training and testing, which is discussed in the next section.

4.1.5 Modeling

The extracted features along with identified markers are integrated using a concatenation-based approach. The complete description of the integration of features is given in Section 1.2.1.1. The integrated features are then passed to model training and testing. A Bayesian-optimized Deep Neural Network (DNN) is presented for training and testing of the extracted features for survival prediction of BRCA and LUAD patients which are discussed in the forthcoming sections.

4.1.5.1 Deep Neural Network (DNN)

DNNs are a form of artificial neural network having multiple hidden layers between the input and output layers [184]. The complete working of DNN is given in Section 3.2.5. In BioSurv, first, an input layer is created with a total of 303 and 211 input features for BRCA and LUAD respectively. This is followed by hidden layers with a number of neurons and dropout rate as attributes. The Relu activation function is used at this stage. These hidden layers enable the network to learn hierarchical data representations, capturing more complex patterns and characteristics than neural networks with fewer hidden layers. At the end, the output layer is given with a sigmoid activation function [185]. The learning rate, dropout, and hidden layers are tuned with Bayesian optimization, discussed in the following subsection.

4.1.5.2 Bayesian Optimization

Bayesian optimization is a prominent method for optimizing expensive-to-evaluate black-box functions [175]. It is frequently employed in ML, DL, engineering, and other disciplines that require optimizing a function with an unknown analytical form. Bayesian optimization integrates statistical models, typically Gaussian processes, with acquisition functions to guide the search for the optimal solution.

The basic concept is to model the unknown function using a probabilistic surrogate model, such as a Gaussian process, which provides a posterior distribution over the function values given the observed data. This model is then utilized to iteratively select the next point to evaluate based on an acquisition function that strikes a balance between exploration and exploitation. In Bayesian optimization, the acquisition function quantifies the value of sampling a specific point in the search space, considering both the predicted function value at that point (exploitation) and the uncertainty of the prediction (exploration). Expected Improvement (EI), Probability of Improvement (PI), and Upper Confidence Bound (UCB) are typical acquisition functions. These acquisition functions permit Bayesian optimization to explore various regions of the search space and sample points likely to improve the optimization and avoid sampling points unlikely to improve the optimization [186]. In the present research, DNN parameters are optimized using BO. The algorithm describing the optimization of DNN is given in Algorithm 4.2.

Algorithm 4.2 Bayesian Optimization for DNN

INPUT: DNN model $f(*)$, hyperparameter search space $\mathcal{P}$, acquisition function $a(\cdot)$, number of iterations T

OUTPUT: Optimal hyperparameters θ^*

BEGIN:

1: Initialize dataset $\mathcal{D} = \{\}$
2: **for** $t = 1$ to T **do**
3: Fit DNN model $f(\cdot)$ with hyperparameters θ_t using $\mathcal{D}$
4: Evaluate acquisition function $a(\cdot)$ to select next hyperparameters θ_{t+1}
5: Update dataset $\mathcal{D} \leftarrow \mathcal{D} \cup \{(\theta_t, f(\theta_t))\}$
6: **end for**
7: **return** $\theta^* = \arg\max_{\theta \in \mathcal{D}} f(\theta)$

End

The EI acquisition function is used in the current research, which returns the best hyperparameters. Three hyperparameters of DNN comprising learning rate, number of hidden layers, and dropout are tuned, and it returns the best hyperparameters with 0.08, 241, and 0.249 values, respectively, for the parameters as mentioned above. The 10-fold cross-validation is used, and the performance is evaluated by taking an average of each fold. The working of cross-validation is given in Section 3.2.5.8. The detailed flowchart of the BioSurv is given in Figure 4.3 and the algorithm is given in Algorithm 4.3. The experiments use the BioSurv applied on TCGA-BRCA and TCGA-LUAD datasets containing mRNA, miRNA,

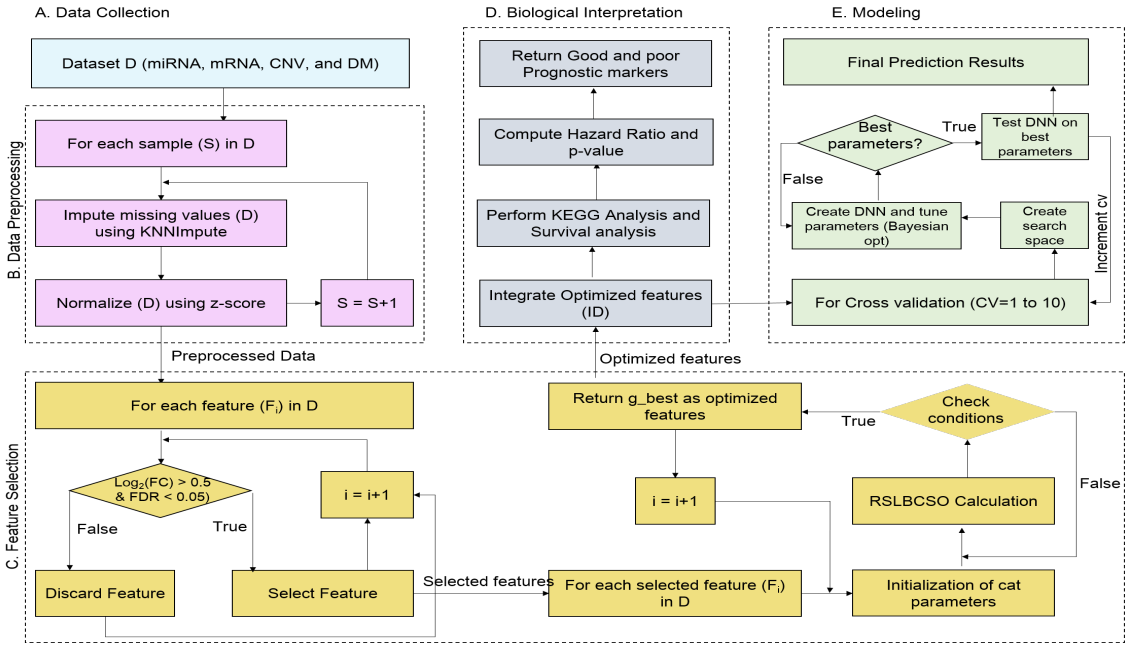


Figure 4.3: Flowchart of BioSurv [2]

CNV, and DM profiles. The BRCA and LUAD have 616 and 425 samples, respectively. The 10-fold cross-validation is performed in which nine folds are used for training and one fold for testing. This is repeated until all the folds serve a test set. The average of each fold result is calculated to achieve the final result.

4.2 Experimental Setup and Results

Section 3.1 gives a complete description of the experimental hardware and software used. The libraries used to perform the implementation are NumPy, pandas, for DNN, sklearn for datasets and performance metrics, a lifeline for KM plots, skopt for Bayesian optimization, and matplotlib. The R packages used are survcomp for CI computation, tidyverse, and dplyr. The models comprising XGBoost, Support Vector Machine (SVM), Random Forest (RF), Gradient Boosting Machine (GBM), KNN, and Decision Tree (DT) are implemented using the scikit-learn (`https://scikit-learn.org/stable/`) library.

4.2.1 Experimental Setup

The minimum hardware requirement to implement the work is 8 GB RAM with an i5 processor. Python 3.9.0 and RStudio 4.2.2 are used to implement the work. DAVID functional Analysis is used to perform the pathway analysis.

Algorithm 4.3 Algorithm of BioSurv

INPUT: Dataset D (miRNA, mRNA, CNV, DM), $D = X_i, Y_i$, $i = 1$ to n

OUTPUT: Prognostic Biomarkers (N), Performance Parameters (P)

BEGIN

Step 1: Preprocess Dataset
for each sample S in D **do**
    $D \leftarrow$ KNNImpute(D)
    $D \leftarrow$ z_score(miRNA, mRNA, and DM)
end for
Step 2- Extract Features F_i
for each F_i in (miRNA, mRNA, DM, and CNV) **do**
    Compute $|log_2(FC)|$ and FDR
    if ($|log_2(FC)| > 0.5$) and ($FDR <0.05$) **then**
        Select F_i
    end if
end for
for selected features F_i in D **do**
    Repeat until F_i is NULL
    Compute fitness for each F_i using RSLBCSO
    Select F_i with best fitness
end for
Step 3- Biological Interpretation
for each F_i (miRNA, mRNA, DM, CNV) **do**
    Perform KEGG analysis and Compute HR
    Return good and poor prognostic biomarkers N
end for
Integrate selected features from RSLBCSO
Step 4: Modeling
Impose 10-fold cross-validation i.e. $D = D_1, D_2, ..., D_{10}$
for CV=1 to 10 **do**
    Train DNN
    Tune hyperparameters with Bayesian optimization
    Return best hyperparameters
    Test DNN using best hyperparameters and Return P
end for
Compute mean of P

End

4.2.2 Experimental Steps

The following steps have been implemented for the BioSurv approach based on biomarker identification for survival prediction of BRCA and LUAD patients in multi-omics datasets:

- Two datasets comprising LUAD and BRCA are downloaded with omics types miRNA, mRNA, CNV, and DM. The downloaded data is passed to the KNNImpute method to impute the missing values. Further, the normalization is performed using z-score normalization.
- Second, the feature/ biomarker identification is presented in which first, the statistical analysis is performed in which those features are selected whose $|log_2(FC)| > 0.5 and FDR < 0.05$. Secondly, the selected features are passed to RSLBCSO for the extraction of optimized features.
- Third, the KEGG analysis and survival analysis are performed, which returns good and poor prognostic markers based on computed HR and p-value.
- At last, DNN is trained whose parameters are optimized with Bayesian optimization. The model is tested on the best parameters achieved, and performance is evaluated using six performance parameters comprising accuracy, sensitivity, specificity, precision, Area Under Curve (AUC), and CI. The description of performance parameters is given in Section 3.2.6.

4.2.3 Results and Discussions

The results section is divided into two phases comprising identified biomarkers and prediction results and are described below:

4.2.3.1 Identified Biomarkers

The features/ biomarkers extracted from RSLBCSO are validated using KEGG and survival analysis for the identification of prognostic biomarkers. KEGG analysis is performed to identify the markers from mRNA, miRNA, CNV, and DM, which are highly enriched in BRCA and LUAD patients. The KEGG analysis identifies the markers that are responsible for cell growth and development and signal transduction [187] and whose p-value is less than 0.05. A total of 6 mRNAs, 5 miRNAs, 4 CNVs, and 6 DM markers for BRCA and 9 mRNAs, 4 miRNAs, 5 CNVs, and 6 DM markers for LUAD patients, respectively, are identified. The six mRNA markers from BRCA are FGFR3, YWHAG, NFKB2, RAB2A, CHEK1,

and ATG5. FGFR3 plays an important role in cell growth and development and is identified as a poor prognostic marker in triple-negative BRCA (TNBC) [188]. YWHAG is involved in signal transduction and cell cycle pathways and has been identified by Mei et al. [189] as a poor prognostic marker. NFKB2 is often mutated in malignancies and is identified as risk signatures in BRCA [190]. RAB2A is responsible for cancer cell proliferation and migration. Wang et al. [191] identified RAB2A as a prognostic marker highly expressed in high-risk patients. CHEK1 plays an important role in signal transduction and is identified as a poor prognostic marker [192]. ATG5 plays a critical role in the process of autophagy, which is the cellular process that involves the degradation and recycling of cellular components. Grandvallet et al. [193] identify it as a poor prognostic marker and is responsible for cell migration in TNBC patients. The miRNA markers identified by KEGG analysis for BRCA are miR-106b, miR-132, miR-222, miR-143, and miR-98. miR-106b and miR-132 are involved in various cellular processes, including cell proliferation, differentiation, and apoptosis. miR-106b is identified as a cancer progression marker by targeting the PTEN marker [194], and miR-132 is identified as a potential biomarker for therapeutic targets in BRCA patients [195]. miR-222 plays a crucial role in cell survival, proliferation, and growth and is identified by Kim et al. [196] as a molecular marker in BRCA. miR-143 and miR-98 regulate apoptosis and metabolism and are identified as diagnostic biomarkers in the BRCA HER2 subtype [197]. The four identified CNV markers are STK11, ROCK1, IL13, and SMC3. STK11 plays a crucial role in autophagy and signal transduction and is identified by Firooz et al. [198] as a driver gene in BRCA. ROCK1 and IL13 are involved in cell migration, proliferation, and contraction. ROCK1 is identified as a target gene for miR-202 [199], and IL13 is identified as a risk biomarker affecting the overall survival of BRCA patients [200]. SMC3 is involved in the cell cycle pathway and is identified as a hub gene down-regulated in BRCA [201]. Next are the DM markers comprising TSC2, ARNT2, AXIN1, DLL1, LAMA5, and PLCG2 identified in KEGG analysis. TSC2 [202] mutations play a role in signal transduction, which regulates cell growth, proliferation, and survival. ARNT2 is involved in various cellular processes, including cell proliferation, differentiation, and apoptosis, and is identified by Liu et al. [48] as a poor prognostic marker in the BRCA Luminal B subtype. AXIN1 expression is significantly lower in BRCA tissues and is identified as the target gene leading to tumor progression [203]. DLL1 is involved in cell proliferation, invasion, and migration and is identified as a poor prognostic marker [204]. PLCG2 plays a key role in several biological processes, including immune responses and cell growth.

It is identified as a prognostic marker up-regulated in BRCA tissue [205]. LAMA5 [206] is involved in cell adhesion, migration, proliferation, and differentiation.

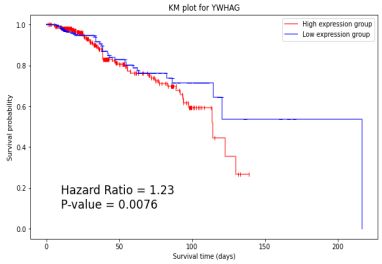
Furthermore, the mRNA biomarkers identified for LUAD using KEGG analysis are FN1, ITGA3, PRKAA2, SGK2, CASP8, FZD3, RHOA, TGFB1, and RRAS. These markers are involved in cell growth, proliferation, migration, and apoptosis. FN1 and ITGA3 have been identified as prognostic markers that show poor overall survival of LUAD [207]. PRKAA2 has been identified by Yao et al. [208] as a driver gene that leads to poor prognosis in LUAD patients. Zeng [209] identified SKG2 as a prognostic risk signature related to the overall survival of LUAD patients. CASP8 and FZD3 are novel markers identified for the first time in LUAD. CASP has been identified as autophagy-pyroptosis-related genes, which is highly expressed in LUAD patients [210]. FZD3 has been identified by Kohansal et al. [211] as a target gene in gastric cancer that inhibits cancer progression. Lin et al. [212] identified RHOA and TGFB1 as immune-related markers with a strong positive correlation in LUAD. RHOA has been identified as a highly expressed gene in LUAD [213]. The four miRNAs of LUAD patients are miR-132, miR-155, miR-221, and miR-222. miR-132 has been identified as tumor suppressor gene [214] and miR-155 as a diagnostic marker in LUAD patients [215]. The markers miR-221 and miR-222 have been identified by Guo et al. [216] as tumor progression markers that are highly over-expressed in lung cancer patients. Moving ahead, the CNV markers identified for LUAD are BPIFB1, MAP2K4, NLRP1, CYCS, and ITCH, in which ITCH has been found to be a tumor suppressor gene that inhibits the cell proliferation in lung cancer [217]. MAP2K4 has been identified by Wang et al. [218] as a target gene that regulates the proliferation and apoptosis of lung cancer cells. Zhang et al. [219] identified CYCS as a risk prognostic signature associated with high-risk in LUAD patients. NLRP1 has been identified as a poor prognostic marker in LUAD patients [220]. BPIFB1 has been identified as a candidate gene highly expressed in LUAD patients [221]. The DM markers of LUAD identified through KEGG analysis are TYK2, AP2A2, NEU1, SYNJ2, GIT1, and TBCD, where He et al. [222] have identified TYK2 as a prognostic marker associated with immune infiltration. AP2A2 has been identified as risk-associated genes [223]. NEU1 has been identified as a poor prognostic marker by Zhao et al. [224] and acts as an independent prognostic factor for LUAD patients. Hou et al. [225] identified SYNJ2 as a prognostic that is up-regulated and leads to poor survival prediction in lung cancer patients. Tao et al. [226] has identified GIT1 as a poor prognostic marker in LUAD and liver cancer patients. TBCD has been identified as a diagnostic and prognostic marker affecting the overall survival

of LUAD patients [227]. The biomarkers identified using KEGG analysis are then passed to the coxPH model to compute each marker's p-value and HR and are given in Table 4.3. HR depicts the risk for each gene, i.e., if the value of HR is close to 1 or greater than 1, then that gene is identified as a poor prognostic marker; otherwise, the marker is treated as a good prognostic marker. The 5-year survivability is considered in the current study.

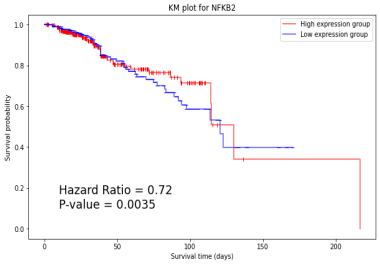
Table 4.3: P-value and HR for identified biomarkers

Type	**BRCA**			**LUAD**		
	Marker	**p-value**	**HR**	**Marker**	**p-value**	**HR**
mRNA	FGFR3	0.003	0.79	RRAS	0.03	1.19
	YWHAG	0.007	1.23	FN1	0.005	1.15
	NFKB2	0.003	0.72	ITGA3	0.02	1.11
	RAB2A	0.0001	1.46	PRKAA2	0.006	0.78
	CHEK1	0.006	1.31	SGK2	0.01	1.12
	ATG5	0.01	1.27	CASP8	0.01	1.14
	-	-	-	FZD3	0.007	0.72
	-	-	-	RHOA	0.01	1.22
	-	-	-	TGFB1	0.006	1.28
miRNA	miR-106b	0.05	1.08	miR-132	0.02	1.21
	miR-132	0.01	1.29	miR-155	0.007	1.25
	miR-222	0.001	1.35	miR-221	0.013	1.12
	miR-143	0.037	1.10	miR-222	0.005	1.16
	miR-98	0.04	1.33	-	-	-
CNV	STK11	0.007	0.73	BPIFB1	0.01	1.21
	ROCK1	0.002	0.56	MAP2K4	0.01	0.73
	IL13	0.02	0.98	NLRP1	0.04	0.77
	SMC3	0.005	0.70	CYCS	0.005	1.07
	-	-	-	ITCH	0.001	1.22
DM	TSC2	0.001	1.61	TYK2	0.04	0.75
	ARNT2	0.009	1.02	AP2A2	0.03	0.74
	AXIN1	0.001	1.44	NEU1	0.002	0.75
	DLL1	0.02	0.73	SYNJ2	0.01	1.15
	LAMA5	0.008	0.78	GIT1	0.005	0.95
	PLCG2	0.001	0.69	TBCD	0.03	0.77

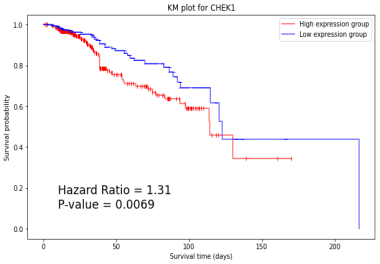
From the results, it is found that 4 mRNAs, 5 miRNAs, 1 CNV, and 3 DMs of BRCA and 7 mRNAs, 4 miRNAs, 3 CNVs, and 1 DM of LUAD patients have HR ratios greater than one and close to one. These markers are identified as poor prognostic markers; the remaining markers are good prognostic markers. The KM plots of the poor prognostic markers are made and are shown in Figure 4.4 below.



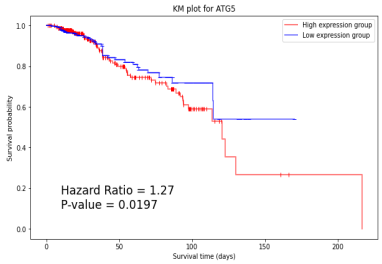
(i) YWHAG (BRCA)



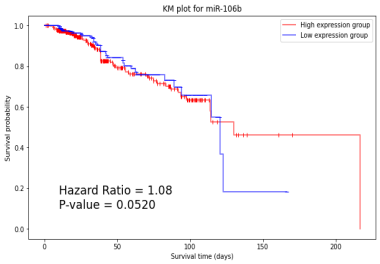
(ii) NFKB2 (BRCA)



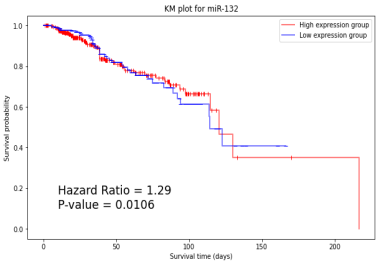
(iii) CHEK1 (BRCA)



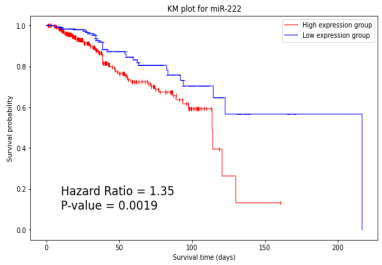
(iv) ATG5 (BRCA)



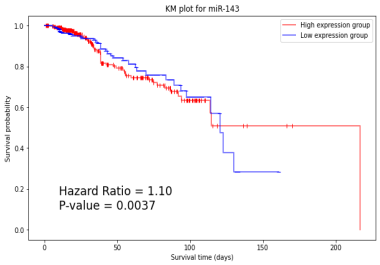
(v) miR-106b (BRCA)



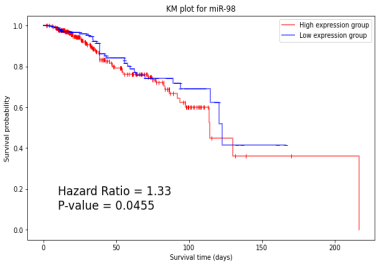
(vi) miR-132 (BRCA)



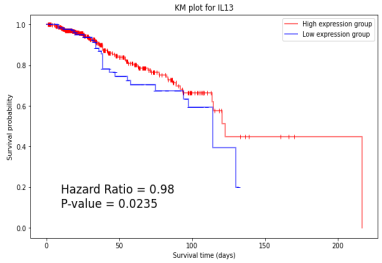
(vii) miR-222 (BRCA)



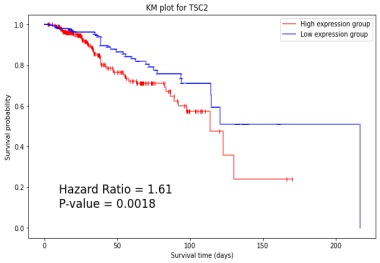
(viii) miR-143b (BRCA)



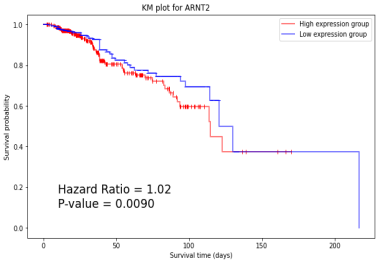
(ix) miR-98 (BRCA)



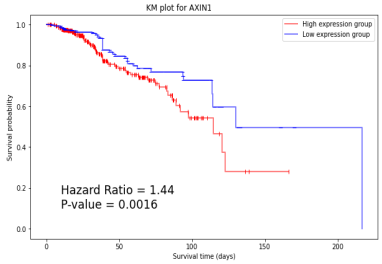
(x) IL13 (BRCA)



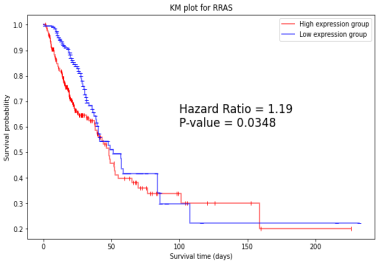
(xi) TSC2 (BRCA)



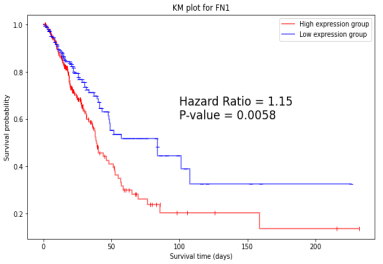
(xii) ARNT2 (BRCA)



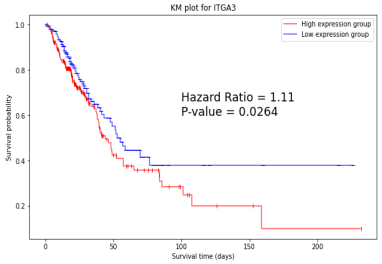
(xiii) AXIN1 (BRCA)



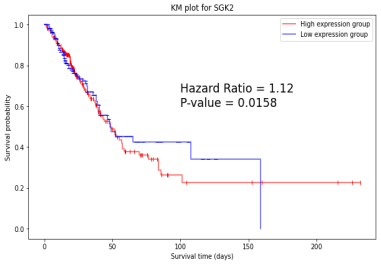
(xiv) RRAS (LUAD)



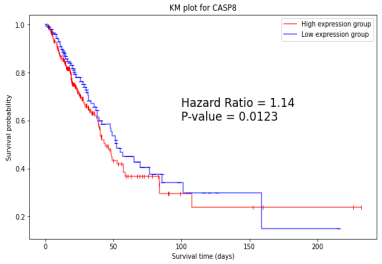
(xv) FN (LUAD)



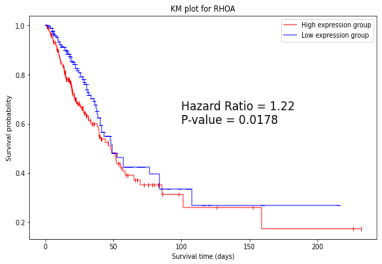
(xvi) ITGA3 (LUAD)



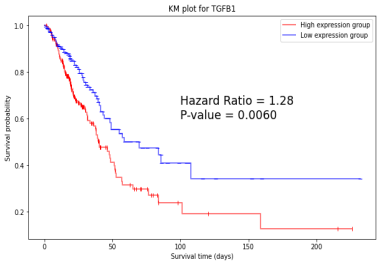
(xvii) SGK2 (LUAD)



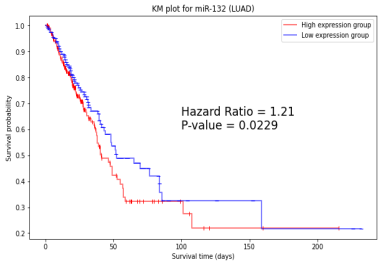
(xviii) CASP8 (LUAD)



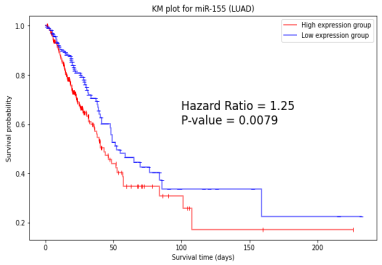
(xix) RHOA (LUAD)



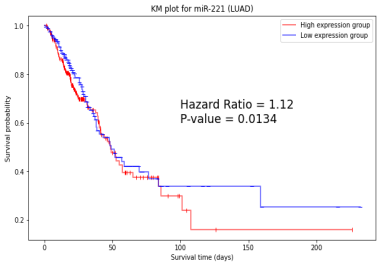
(xx) TCFB1 (LUAD)



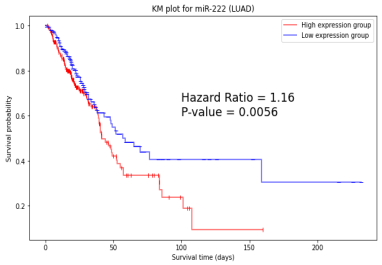
(xxi) miR-132 (LUAD)



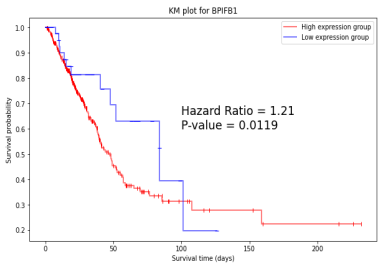
(xxii) miR-155 (LUAD)



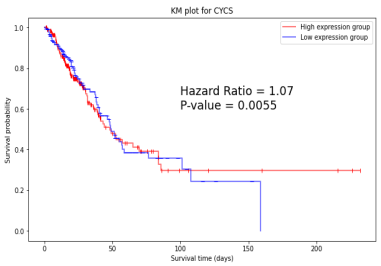
(xxiii) miR-221 (LUAD)



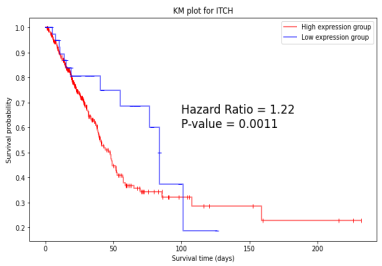
(xxiv) miR-222 (LUAD)



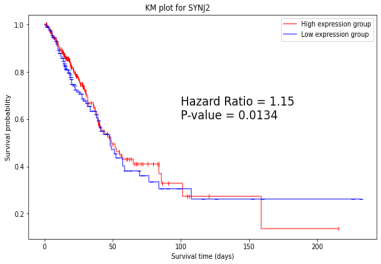
(xxv) BFIFB1 (LUAD)



(xxvi) CYCS (LUAD)



(xxvii) ITCH (LUAD)



(xxviii) SYNJ2 (LUAD)

Figure 4.4: Poor Prognostic Markers for BRCA and LUAD samples [2]

4.2.3.2 Prediction Results

To test the performance of extracted features/ biomarkers from RSLBCSO, a Bayesian optimized DNN is used. The features are integrated using concatenation-based integration and passed to a Bayesian optimized DNN for survival prediction of short-term survival and long-term survival. 10-fold cross-validation is performed in which one fold is used for testing and nine folds for training. This is repeated until all the folds are tested, and the final result is the average of each fold. The performance is evaluated using six performance parameters comprising accuracy, sensitivity, specificity, precision, AUC, and CI. The Bayesian Optimized DNN is applied on single omics, i.e., on miRNA, mRNA, CNV, and DM alone, and the multi-omics, i.e., integrated miRNA+mRNA+CNV+DM dataset. Performance is evaluated, and it is found that the BioSurv performed well on integrated multi-omics data with accuracy, sensitivity, specificity, precision, and AUC value of 91.60%, 88.02%, 89.12%, 90.01%, and 90% respectively for BRCA and 90.1%, 87.5%, 88.3%, 86.4%, and 86%, respectively for LUAD samples. The results are shown in Table 4.4.

Table 4.4: Results of BioSurv on single and integrated omics

Cancer	Dataset	Accuracy	Sensitivity	Specificity	Precision	AUC
BRCA	mRNA	0.78	0.76	0.76	0.74	0.61
	miRNA	0.85	0.82	0.82	0.86	0.83
	CNV	0.83	0.80	0.79	0.76	0.65
	DM	0.82	0.78	0.80	0.79	0.79
	Integrated	0.91	0.88	0.89	0.90	0.90
LUAD	mRNA	0.87	0.84	0.82	0.74	0.78
	miRNA	0.85	0.81	0.80	0.78	0.70
	CNV	0.82	0.78	0.81	0.80	0.77
	DM	0.87	0.84	0.85	0.82	0.81
	Integrated	0.90	0.87	0.88	0.86	0.86

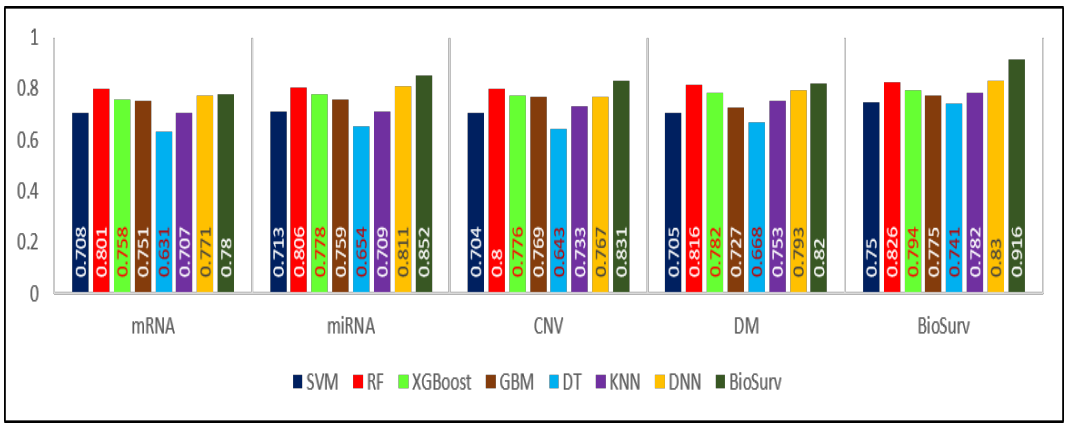
Additionally, the results of BioSurv with several state-of-the-art models comprising DNN, XGBoost, RF, SVM, GBM, KNN, and DT have been compared. It is found that BioSurv outperforms all the models and shows an improvement of approximately 8% and 3% in terms of accuracy for both BRCA and LUAD cancers, respectively. The comparative results of BioSurv with existing methods for BRCA and LUAD are given in Table 4.5 below.

Table 4.5: Results of BioSurv and existing models on BRCA and LUAD samples

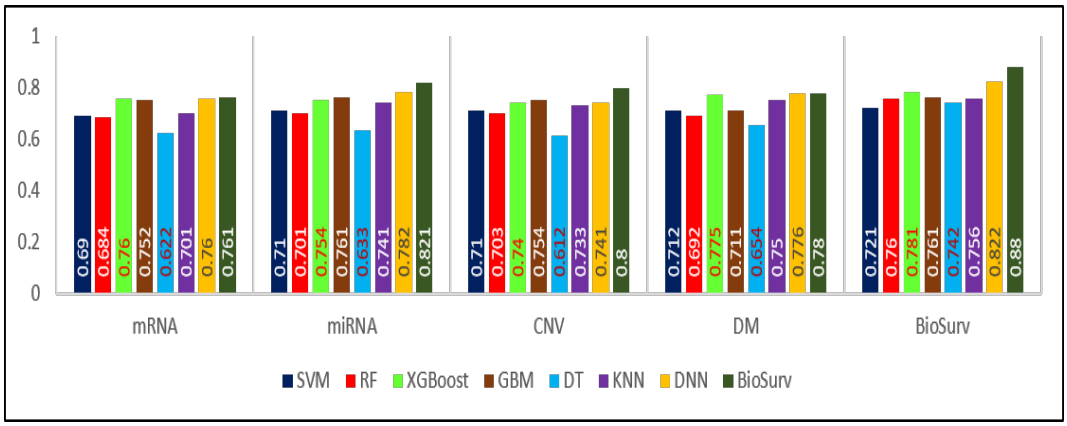
P	Models	BRCA					LUAD				
		A	**B**	**C**	**D**	**E**	**A**	**B**	**C**	**D**	**E**
Accuracy	SVM	0.708	0.713	0.704	0.705	0.75	0.664	0.729	0.711	0.745	0.761
	RF	0.801	0.806	0.80	0.816	0.826	0.789	0.781	0.793	0.796	0.804
	XGBoost	0.758	0.778	0.776	0.782	0.794	0.781	0.773	0.745	0.764	0.798
	GBM	0.751	0.759	0.769	0.727	0.775	0.75	0.765	0.77	0.765	0.781
	DT	0.631	0.654	0.643	0.668	0.741	0.632	0.711	0.719	0.703	0.732
	KNN	0.707	0.709	0.733	0.753	0.782	0.773	0.762	0.781	0.796	0.806
	DNN	0.771	0.811	0.767	0.793	0.83	0.799	0.8	0.811	0.852	0.874
	BioSurv	0.78	0.852	0.831	0.82	**0.916**	0.875	0.858	0.825	0.87	**0.901**
Sensitivity	SVM	0.69	0.71	0.71	0.712	0.721	0.666	0.74	0.712	0.74	0.757
	RF	0.684	0.701	0.7032	0.692	0.76	0.795	0.785	0.741	0.771	0.76
	XGBoost	0.76	0.754	0.74	0.775	0.781	0.79	0.763	0.77	0.751	0.8
	GBM	0.752	0.761	0.754	0.711	0.761	0.753	0.779	0.762	0.772	0.784
	DT	0.622	0.633	0.612	0.654	0.742	0.681	0.713	0.70	0.695	0.72
	KNN	0.701	0.741	0.733	0.75	0.756	0.776	0.782	0.773	0.801	0.81
	DNN	0.76	0.782	0.741	0.776	0.822	0.785	0.773	0.794	0.791	0.84
	BioSurv	0.761	0.821	0.80	0.78	**0.88**	0.843	0.812	0.78	0.84	**0.875**
Specificity	SVM	0.684	0.701	0.7032	0.692	0.735	0.671	0.714	0.703	0.712	0.749
	RF	0.792	0.773	0.797	0.792	0.821	0.792	0.773	0.752	0.74	0.813
	XGBoost	0.751	0.767	0.746	0.764	0.763	0.741	0.751	0.77	0.76	0.79
	GBM	0.681	0.72	0.711	0.731	0.762	0.723	0.741	0.729	0.742	0.75
	DT	0.683	0.718	0.711	0.70	0.721	0.63	0.659	0.681	0.674	0.69
	KNN	0.672	0.710	0.732	0.725	0.743	0.747	0.76	0.754	0.765	0.796
	DNN	0.766	0.791	0.748	0.772	0.821	0.799	0.788	0.763	0.814	0.87
	BioSurv	0.76	0.82	0.793	0.8	**0.891**	0.821	0.80	0.813	0.85	**0.883**
Precision	SVM	0.706	0.691	0.721	0.708	0.74	0.706	0.652	0.721	0.721	0.754
	RF	0.752	0.762	0.721	0.71	0.791	0.62	0.682	0.84	0.757	0.791
	XGBoost	0.74	0.712	0.731	0.751	0.74	0.65	0.62	0.713	0.701	0.75
	GBM	0.698	0.711	0.724	0.723	0.739	0.651	0.623	0.691	0.689	0.701
	DT	0.671	0,732	0.709	0.705	0.743	0.637	0.671	0.673	0.678	0.689
	KNN	0.684	0.724	0.712	0.738	0.761	0.669	0.693	0.699	0.743	0.762
	DNN	0.773	0.794	0.753	0.764	0.804	0.762	0.712	0.795	0.781	0.843
	BioSurv	0.741	0.861	0.768	0.791	**0.9**	0.791	0.781	0.80	0.821	**0.864**
AUC	SVM	0.501	0.505	0.507	0.537	0.556	0.482	0.474	0.515	0.57	0.602
	RF	0.52	0.526	0.51	0.506	0.582	0.506	0.506	0.537	0.545	0.599
	XGBoost	0.591	0.683	0.586	0.645	0.726	0.58	0.498	0.525	0.653	0.703
	GBM	0.491	0.536	0.522	0.530	0.540	0.448	0.485	0.549	0.539	0.557
	DT	0.497	0.551	0.488	0.522	0.573	0.481	0.487	0.514	0.513	0.529
	KNN	0.511	0.582	0.543	0.556	0.642	0.490	0.50	0.559	0.608	0.673
	DNN	0.552	0.772	0.62	0.712	0.79	0.60	0.715	0.64	0.754	0.827
	BioSurv	0.61	0.832	0.653	0.79	**0.9**	0.78	0.70	0.771	0.81	**0.86**

A-mRNA, B-miRNA, C-CNV, D-DM, E-Integrated

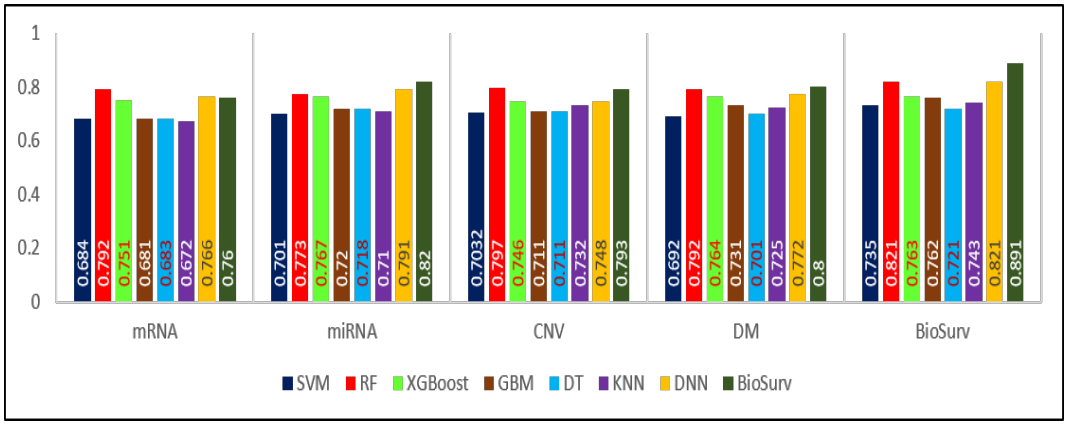
Moreover, the bar plots of accuracy, sensitivity, specificity, and precision have been plotted for BRCA and LUAD patients to test the effectiveness of BioSurv. From the plots, it is visible that BioSurv performed better than the existing models. The bar plots for BRCA are shown in Figure 4.5.



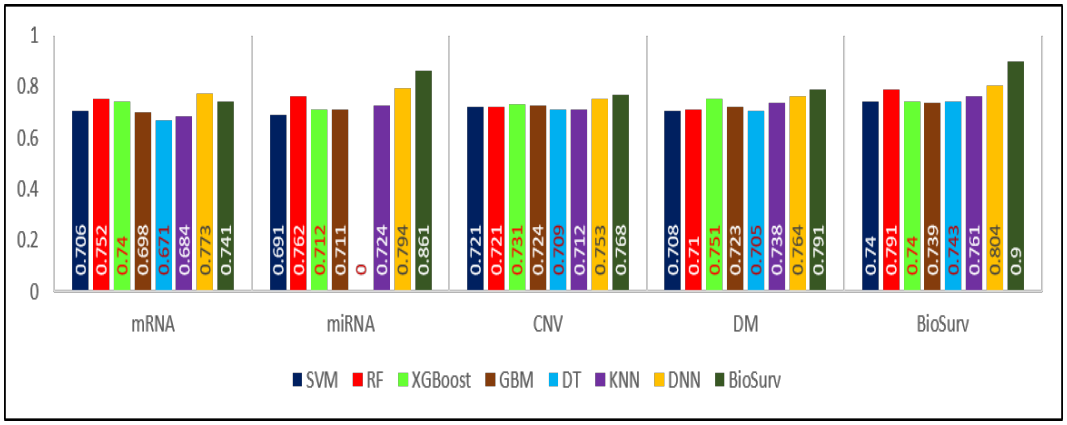
(i) Accuracy



(ii) Sensitivity



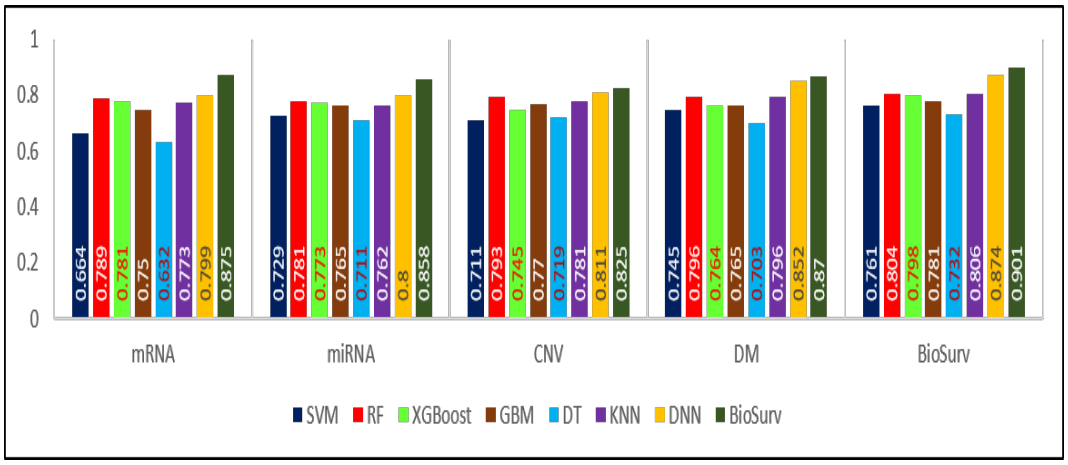
(iii) Specificity



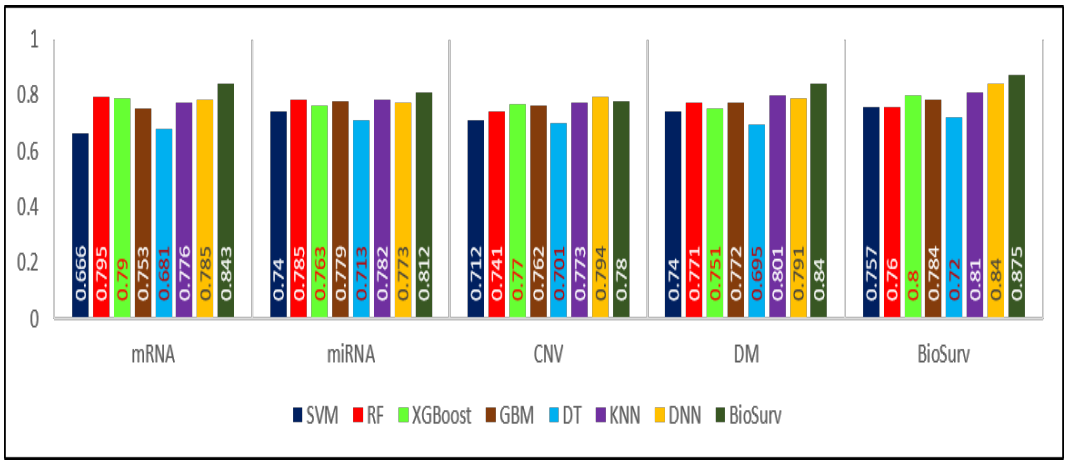
(iv) Precision

Figure 4.5: Bar plots of accuracy, sensitivity, specificity, precision for BRCA [2]

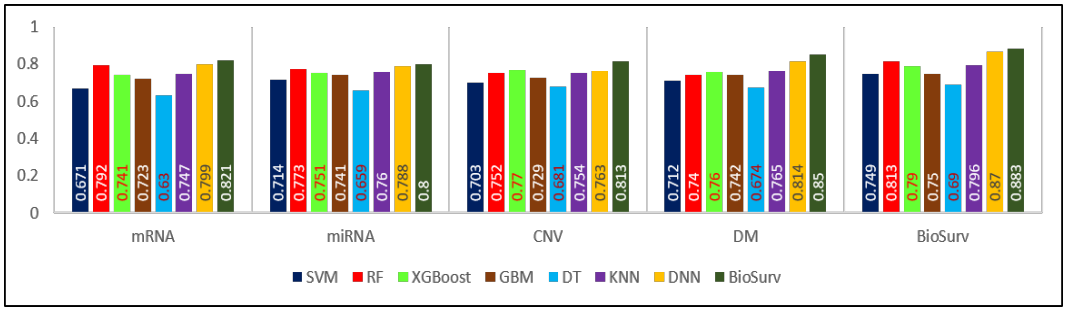
The bar plots for LUAD are shown in Figure 4.6.



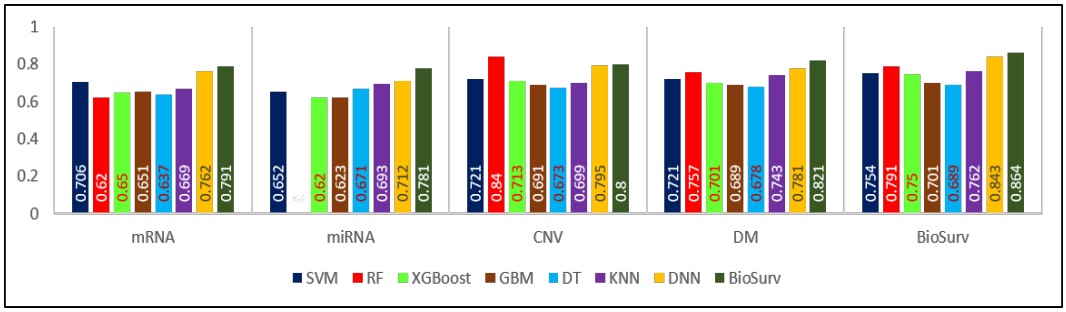
(i) Accuracy



(ii) Sensitivity



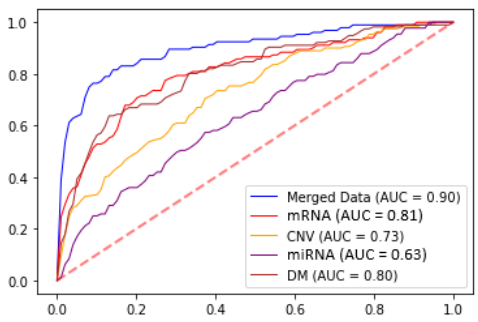
(iii) Specificity



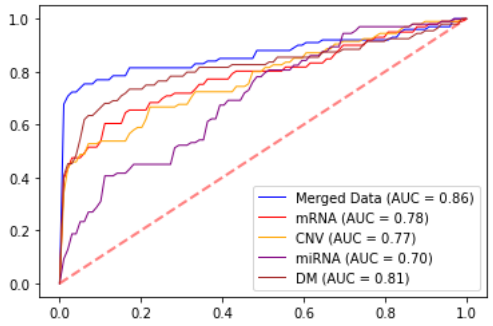
(iv) Precision

Figure 4.6: Bar plots of accuracy, sensitivity, specificity, precision for LUAD [2]

Additionally, the multi-roc curve for AUC has been plotted for BRCA and LUAD patients and shown in Figure 4.7. The comparison is between single omics and multi-omics. The high curve for integrated data shows that BioSurv performed well with an AUC value of 90% and 87% for BRCA and LUAD, respectively.



(i) BRCA

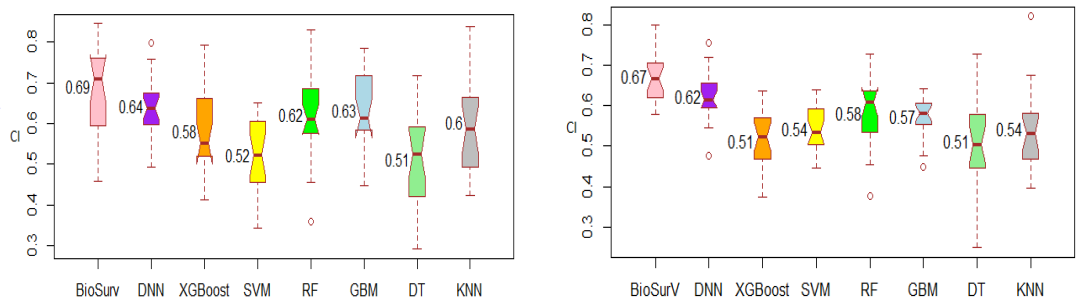


(ii) LUAD

Figure 4.7: Multi-ROC for single and integrated omics for BRCA and LUAD [2]

Similarly, the boxplot of the CI for integrated omics is also plotted and shown in Figure 4.8. The CI is computed using the *concordance.index* function. The code runs for 20 epochs, and an average is calculated as the final CI value. The

CI is compared with the existing models comprising DNN, XGboost, RF, SVM, and GBM, KNN, and DT it is evident from the plots that BioSurv outperformed with a CI value of 0.69 and 0.67 for BRCA and LUAD, respectively.



(i) BRCA (ii) LUAD

Figure 4.8: Comparative analysis of BioSurv CI for integrated omics with existing models for BRCA and LUAD [2]

Furthermore, the BioSurv is compared with the existing works comprising [87], and [77] based on CI value. In these works, 10-fold cross-validation is used. BioSurv also utilized 10-fold cross-validation, and it is found that BioSurv shows an improvement of approx 3% for BRCA and 5% for LUAD patients. The comparison of BioSurv with existing works is shown in Table 4.6.

Table 4.6: Comparison of BioSurv with existing works

Cancer Type	Dataset	Work	CI
BRCA	mRNA+miRNA+CNV+DM	[77]	0.66
		BioSurv	0.69
LUAD	mRNA+miRNA+CNV+DM	[77]	0.62
		[87]	0.65
		BioSurv	0.67

4.2.3.3 Validation of BioSurv on METABRIC Dataset

To validate the performance of the BioSurv, the METABRIC dataset (`https://www.mercuriolab.umassmed.edu/metabric`) of BRCA is used. The complete description of the METABRIC dataset is given in Section 3.2.1.2. The dataset

comprises 1980 mRNA, 2172 CNV, and 1418 DM samples. At first, the common samples from each type are extracted, which returns 1418 for the common samples. The 5-year survivability is used to perform the experiment. The NA values for each type are imputed using the KNNImpute method. The z-score normalization is performed for the mRNA and DM datasets, and the CNV data is utilized as it is. The statistical analysis test is performed, and FDR and $|log_2(FC)|$ values are computed. Those values are selected whose $FDR < 0.05$ and $|log_2(FC)| > 0.5$, and it returns 2332 mRNAs, 2584 CNVs, and 814 DMs, respectively. The extracted features are then passed to RSLBCSO, which optimizes the feature space and returns 153 mRNAs, 101 CNVs, and 120 DMs as features. The extracted features are then passed to KEGG pathway analysis for the identification of biomarkers that are responsible for cell cycle, growth and development, proliferation, and migration, respectively. 6 mRNAs comprising TRAF4 [228], DCTPP1 [229], RRM2 [230], CTTN [231], PKN2 [232] and CAPN5 [232], five CNVs including ELK1 [233], CKS2 [234], CD58 [235], PIM2 [236], and COL4A2 [237], and 5 DMS comprising ARAP3 [238], ABCB4 [239], CLDN15 [240], DSC3 [241], and DHX9 [242] has been identified. Further, the survival analysis of the markers extracted using KEGG analysis is performed, and it identifies 5 mRNAs, 1 CNV, and 3 DMs at poor prognostic markers because of the HR close to 1 and greater than 1. The HR and p-value of extracted markers are given in Table 4.7.

The extracted features/ biomarkers are then trained using a Bayesian-optimized DNN. The performance is evaluated, and it is found that BioSurv performed well with accuracy, sensitivity, specificity, precision, AUC, and CI values of 88.78%, 87.84%, 86.70%, 88.60%, 93%, and 0.70 respectively. The results of BioSurv on single-omics and multi-omics METABRIC data are shown in Table 4.8. Furthermore, the multi-roc curve is plotted in Figure 4.9 to show the effectiveness of BioSurv. It is visible that BioSurv performed effectively on the integrated dataset and shows an improvement with 6%, 17%, and 5% in terms of AUC compared to mRNA, CNV, and DM, respectively. The results of BioSurv reveal that it performed well in terms of accuracy, sensitivity, specificity, precision, AUC, and CI value when compared with existing models and state-of-the-art works. BioSurv outperformed all cancer types, including TCGA-BRCA, TCGA-LUAD, and METABRIC. The reason behind the best performance of BioSurv is that the parameters of Biosurv are tuned with Bayesian optimization, which adapts itself to the observed results during the search process. It updates the posterior distribution of the hyperparameter space based on the evaluated configurations, leading to better performance. Moreover, in the proposed research, RSLBCSO is used

Table 4.7: P-value and HR values of extracted markers

Data	**Marker**	**p-value**	**HR**
mRNA	TRAF4	0.02	1.20
	DCTPP1	0.009	0.76
	RRM2	0.04	0.69
	CTTN	0.001	1.10
	PKN2	0.005	1.02
	CAPN5	0.004	1.03
CNV	ELK1	0.0001	0.84
	CKS2	0.005	0.79
	CD58	0.01	0.80
	PIM2	0.019	0.85
	COL4A2	0.012	1.12
DM	ARAP3	0.03	0.80
	ABCB4	0.03	1.03
	CLDN15	0.006	1.11
	DSC3	0.0002	1.21
	DHX9	0.005	0.78

to select the most optimized features, which in turn is a better strategy with a better balance between exploitation and exploration, allowing it to explore a more prominent search space and potentially find better solutions.

4.3 Statistical Analysis of BioSurv

To ensure the superiority of the BioSurv, two tests comprising Wilcoxon signed-rank test and Friedman tests have been used. Wilcoxon signed rank is presented by Derrac et al. [243] and is used to make a simple pairwise comparison test. On the other side, the Friedman test is given by Zhang et al. [244], which is used to

Table 4.8: Results of BioSurv on single-omics and multi-omics

Parameter	mRNA	CNV	DM	mRNA+CNV+DM
Accuracy	0.861	0.857	0.842	**0.887**
Sensitivity	0.854	0.706	0.804	**0.878**
Specificity	0.841	0.714	0.791	**0.867**
Precision	0.854	0.732	0.86	**0.886**
AUC	0.87	0.76	0.88	**0.931**
CI	0.62	0.58	0.635	**0.70**

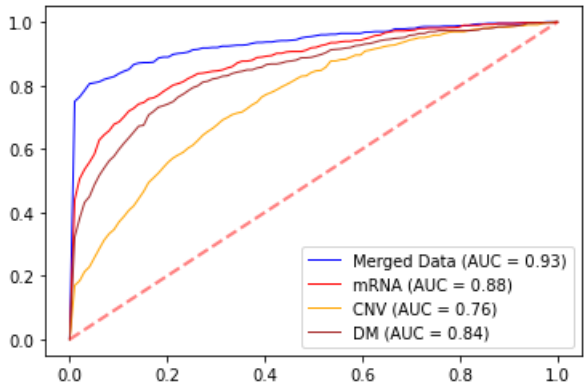


Figure 4.9: Multi-roc plot for single and integrated omics [2]

make multiple comparisons. The original claim in the Wilcoxon signed rank test is that there is no discernible difference between the two models' prediction accuracies. According to the Friedman test's basic hypothesis, there is no discernible difference between all of the models being compared in terms of how accurate their predictions are [245]. The statistical analysis results of both BRCA and LUAD patients comparing the performance of BioSurv with benchmarking models comprising SVM, RF, XGBoost, DNN, GBM, DT, and KNN are given in Table 4.9. The p-value for both tests is computed, and those tests are considered significant, whose p-value is less than 0.05. When the p-value is less than 0.05, then that hypothesis is rejected. From the results of the tests, it has been found that BioSurv performed well in the case of the Fridman test and received a significance than the benchmarking models. In the Wilcoxon signed-rank test, only one case is there where BioSurv, when compared with DNN in LUAD, is not significant.

It achieves a p-value of 0.06, which is violating the test condition. However, it performed well in predicting survival in BRCA patients.

Table 4.9: Statistical analysis results of BioSurv with existing models on BRCA and LUAD

Cancer Type	**Model**	**Wilcoxon signed-rank test (p-value)**	**Friedman test (p-value)**
BRCA	BioSurv vs. DNN	0.0076	$H_0 : e_0 = e_1 = e_2 = e_3 = e_4 = e_5 = e_6 = e_7$ $F = 67.48$ p=$4.7 \times e^{-12}$ (0.0000**) Reject H_0
	BioSurv vs. XG-Boost	0.0039	
	BioSurv vs. SVM	0.00195	
	BioSurv vs. RF	0.0042	
	BioSurv vs. GBM	0.0039	
	BioSurv vs. DT	0.00195	
	BioSurv vs. KNN	0.0039	
LUAD	BioSurv vs. DNN	0.061	$H_0 : e_0 = e_1 = e_2 = e_3 = e_4 = e_5 = e_6 = e_7$ $F = 58.7180$ p=$2.72 \times e^{-10}$ (0.0000**) Reject H_0
	BioSurv vs. XG-Boost	0.0097	
	BioSurv vs. SVM	0.00195	
	BioSurv vs. RF	0.0058	
	BioSurv vs. GBM	0.00585	
	BioSurv vs. DT	0.00195	
	BioSurv vs. KNN	0.00976	

4.4 Conclusion

This chapter discussed the biomarker identification for disease survival prediction in multi-omics datasets by the proposed BioSurv approach. Rigorous statistical analysis and the RSLBCSO algorithm are developed and applied to effectively identify the most promising set of features, enhancing the accuracy and efficiency of subsequent analyses. The KEGG and survival analysis is performed which helped differentiate between good and poor prognostic markers, providing valuable insights into the potential outcomes for cancer patients. A DNN is developed for survival prediction whose parameters are tuned using Bayesian optimization. The results of the developed model are evaluated by taking the TCGA-BRCA, TCGA-LUAD dataset, and METABRIC. Predominantly, BioSurv showcased its effectiveness in biomarker identification and cancer survival prediction. Its robust framework and utilization of advanced computational techniques contributed to

the successful identification of poor prognostic markers and accurate prediction of cancer survival outcomes. The ability of the proposed BioSurv approach to accurately identify the biomarkers and predict cancer survival will help clinicians in guiding more suitable cancer treatment or post-surgical therapeutic decisions for the patient. Also, clinicians can provide special care programs to a person predicted as a short-term survivor.

In the next chapter, biomarker identification for the disease subtype prediction using integrated multimodal variational autoencoder and simplified graph convolutional neural network is discussed. The objective of the chapter is to identify the prognostic biomarkers for disease subtypes by considering multi-omics datasets for three different types of cancers.

Chapter 5

iMVAN: Integrative Multimodal Variational Autoencoders based Biomarker Identification for disease subtype classification

In the previous Chapter, a proposed BioSurv approach for biomarker identification for disease survival prediction is discussed. The biomarker identification is performed using Random Spatial Local Best Cat Swarm Optimization (RSLBCSO) and Bayesian optimized deep neural network (DNN) is employed to perform disease survival prediction. The capability of BioSurv is to accurately identify the biomarkers and to predict a patient as a short-term survivor or long-term survivor. This prediction could be helpful in providing treatment for short-term survivors by focusing on the identified poor prognostic biomarkers.

In this Chapter, a detailed overview of the developed iMVAN approach based on the proposed framework using integrative multimodal variational autoencoder (MVAE) and simplified graph convolutional networks (SGC) are discussed in detail. The design for biomarker identification for disease subtype classification is built using data acquisition, dimensionality reduction, similarity network fusion (SNF), and modeling phases. The MVAE is used as dimensionality reduction for biomarker identification and a SGC is taken as a learning model for disease subtype classification. In the chapter, multi-omics datasets of multiple cancer types are taken to validate the iMVAN approach.

This chapter starts with the discussion of the developed iMVAN approach in Section 5.1. The experimental analysis and results are given in Section 5.2. The computational evaluation of iMVAN is discussed in Section 5.3. At the end, the conclusion is provided in Section 5.4.

5.1 Overview of iMVAN approach

The identification of biomarkers specific to various cancer subtypes plays a vital role in enhancing the accuracy of diagnosis, prognosis, and therapeutic approaches.

This multi-class classification task is challenging when integrating multi-omics using machine learning (ML) approaches. A multi-class network classification model is required for the discovery of biomarkers related to different subtypes of cancer. Therefore, an approach called iMVAN based on multimodal variational autoencoder (MVAE), Similarity Network Fusion (SNF), and Simplified Graph Convolutional Networks (SGC) is developed to identify the biomarkers related to different subtypes of cancer in multi-omics data. The iMVAN is divided into six phases comprising data acquisition, dimensionality reduction, network fusion, biological interpretation, modeling, and results evaluation as shown in Figure 5.1. At first, multi-omics data comprising genomic, transcriptomic, and proteomic datasets are collected, and MVAE is applied to extract the features. The extracted features are then passed to biological interpretation and survival analysis, identifying the top prognostic biomarkers. On the other side, a Patient Similarity Network (PSN) is constructed using SNF. Then the extracted features (vector data) from MVAE and matrix data from SNF are integrated using an SGC to test the extracted biomarker's performance and classify them into their cancer subtype. The six phases are described in detail in the following sections.

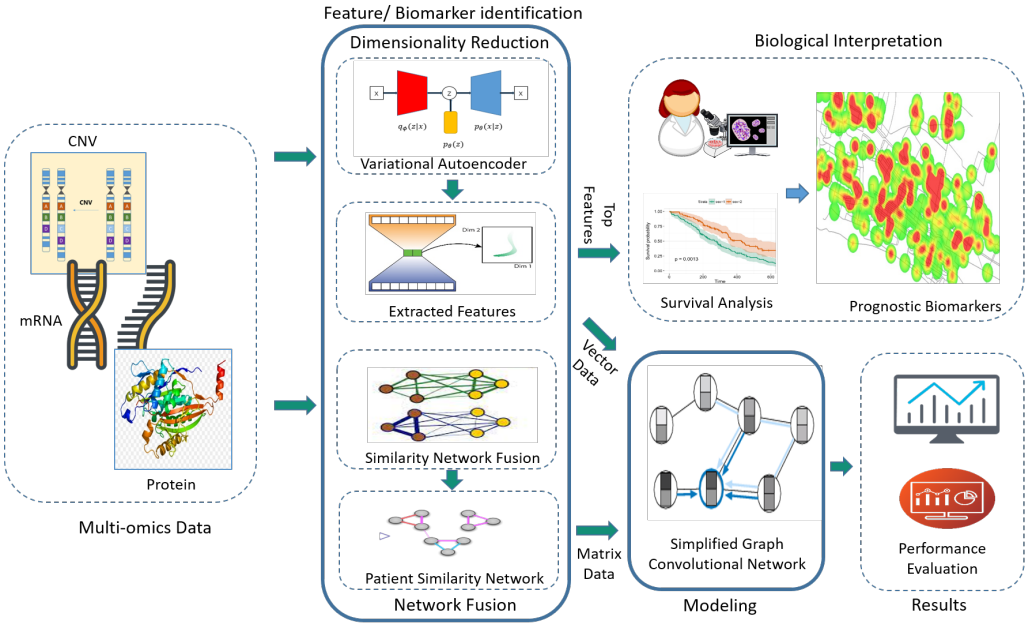


Figure 5.1: Workflow of iMVAN [3]

5.1.1 Data Acquisition

Genomic, transcriptomic, and proteomic play an essential role in cancer progression. Multi-omics data comprising Copy Number Variation (CNV) at the genomic

level, messenger ribonucleic acid (mRNA) data at the transcriptomic level, and reverse phase protein array (rppa) data at the proteomic level of breast cancer (BRCA) patients is collected from The Cancer Genome Atlas Portal (TCGA) portal [246]. The TCGA dataset is discussed in Section 3.2.1.1. Each type contains different samples, for example, in CNV, there are 1080 samples, followed by 1093 samples in mRNA, 887 samples in rppa data, and 1097 samples in clinical data. The common samples are collected using the Venn diagram, which returns 511 samples. The breast tumor patients are divided into four subtypes comprising Basal-like, human epidermal growth factor receptor 2 (Her2) enriched, Luminal A, and Luminal B, respectively. Genes that are normally present in the breast's basal cells display certain characteristics that identify basal-like breast cancer. Triple-negative breast cancer (TNBC) is the term used to describe it because it is characterized by the absence of expression of three specific receptors including Her2, progesterone receptor (PR), and estrogen receptor (ER). Amplification of the Her2 gene or overexpression of the Her2 protein are characteristics of Her2-enriched breast cancer. Genes linked to luminal epithelial cells express certain characteristics that identify luminal A breast cancer. It usually expresses PR and/or ER, and it does not exhibit Her2 amplification or overexpression. The expression of genes linked to luminal epithelial cells is another characteristic of luminal B breast cancer, but it frequently exhibits faster rates of proliferation than luminal A tumors. The complete summary of the dataset is given in Table 5.1.

Table 5.1: Summary of Dataset

Omics Type	**Total Samples**	**Total features**	**BRCA Subtype**	**No. of Samples**
CNV	1080	24776	Basal-like	112
mRNA	1093	20175	Her2-enriched	53
rppa	887	212	Luminal A	248
clinical	1097	-	Luminal B	98
Common Samples	511	45163	**Total**	511

The dataset downloaded from TCGA is in raw form. Therefore, it needs cleaning before further processing. For this, first, the missing values from each type of omics comprising CNV, mRNA, and rppa dataset are imputed using the kNNImputer [148] function. The working of KNNImputer is given in Section 3.2.2. Further, the values in mRNA and rppa are not in a specific range. Some values are quite large and some are close to 0. Therefore, consistency is required for better performance. To do so, normalization is done using z-score normalization. The CNV values are already in the range of [-2,2], hence they are utilized di-

rectly. Once the normalization is done, feature extraction is required because of the large feature space. Therefore, the dimensionality reduction technique MVAE is developed which is discussed in the following subsection.

5.1.2 Dimensionality reduction

The dataset consists of a total of 45,163 features, which is enormous. It is challenging to work with such a large feature space. The analysis of multi-omics data sets may be computationally demanding due to their inherent high dimension. Dimensionality reduction is an all-encompassing method for reducing computational load. Therefore, to reduce the dimension, multimodal variational autoencoder (MVAE) is proposed and is discussed in the following section.

5.1.2.1 Multimodal Variational Autoencoder (MVAE)

Variational Autoencoders (VAEs) [157] are deep generative models that can develop a meaningful data manifold from high-dimensional input data. There are three components to VAE: an encoder, a sampling module, and a decoder. Standard autoencoders encode input ($\mathbf{x}_i$) as a single point, but a VAE encodes an input distribution throughout latent space. Considering an omics dataset D with N samples $\{\mathbf{x}^i\}_{i=1}^{N}$ and d-dimensional omics features, a VAE implies that each sample $\mathbf{x}^i \in \mathbb{R}^d$ is built from a latent vector $\mathbf{z}^i \in \mathbb{R}^{\mathbf{p}}$, where $d \gg \mathbf{p}$.

Each latent variable, $\mathbf{z}^i$, is encoded by an encoder based on a prior distribution or a latent distribution, $\mathbf{p_\Theta}(\mathbf{z})$. Furthermore, the encoder implements a variational distribution known as $\mathbf{q_\psi}(\mathbf{z}|\mathbf{x})$ to estimate the posterior probability and deal with the rebellious nature of the posterior value known as $\mathbf{p_\Theta}(\mathbf{z}|\mathbf{x})$ while calculating the distribution of $\mathbf{X}$ or $\mathbf{p_\Theta}(\mathbf{X})$. The learnable parameters of the encoder are represented by $\boldsymbol{\psi}$. A sampler extracts data from a latent space by taking a representative sample from an encoded or encoding distribution, denoted by $\mathbf{q}_\psi(\mathbf{z}|\mathbf{x})$. Sampled points from the $\mathbf{p_\Theta}(\mathbf{x}|\mathbf{z})$ conditional distribution is decoded by a decoder, which reconstructs the inputs $\mathbf{x}$ using the decoder's learnable parameters. Reconstruction and regularisation terms are used in this stage to determine the loss or error in VAE's loss function. The regularisation term measures the distance between the estimated posterior $\mathbf{q_\psi}(\mathbf{z}|\mathbf{x})$ and true posterior $\mathbf{p_\Theta}(\mathbf{z}|\mathbf{x})$ to regularise the latent space. In contrast, the reconstruction term estimates the reconstruction loss that penalizes the network for providing outputs that are different from the input. Kullback-Leibler (KL) divergence [247] is used as the regularisation term in a standard VAE, which utilizes the following loss function given by Eq. (5.1)

to optimize the encoder and decoder at once.

$$\text{loss} = \underset{\boldsymbol{\psi}}{\text{argmin}}(\mathbf{E}_{\mathbf{q}_{\boldsymbol{\psi}}(\mathbf{z}|\mathbf{x})}\left[\log \mathbf{p}_{\boldsymbol{\Theta}}(\mathbf{x}|\mathbf{z})\right] - D_{\text{KL}}\left(\mathbf{q}_{\boldsymbol{\psi}}(\mathbf{z}|\mathbf{x}) \,||\, \mathbf{p}_{\boldsymbol{\Theta}}(\mathbf{z})\right)) \tag{5.1}$$

The KL divergence is given by D_{KL}. The total loss for all features is calculated and is described by Eq. (5.2):

$$\text{loss} = \underset{\boldsymbol{\psi}}{\text{argmin}}(L(\mathbf{x} - \overline{\mathbf{x}}) + \sum_{j} D_{\text{KL}}\left(\mathbf{q}_{\boldsymbol{\psi}}(\mathbf{z}|\mathbf{x}) \,||\, \mathbf{p}_{\boldsymbol{\Theta}}(\mathbf{z})\right)) \tag{5.2}$$

where the reconstruction loss is determined by L. The loss given in Eq. (5.2) is calculated for a single type of data. In the current study, multi-omics datasets comprising genomic, transcriptomic, and proteomic of BRCA patients are used. Therefore, to combine these different types of datasets, a multi-modal variational autoencoder (MVAE) is proposed. VAE has many inputs and outputs since the input data are characterized by multi-omics data types and represented by several matrices $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ for genome, transcriptome, and proteome respectively. Therefore, a combined loss for all three different types is required to be computed. Hence by using Eq. (5.2), the combined loss for the above-mentioned data types is calculated and is given by the following expression:

$$\begin{aligned}
\text{loss} = \alpha(L_1(\mathbf{x}_1 - \overline{\mathbf{x}}_1) + \sum_{j} D_{\text{KL}}(\mathbf{q}_{\boldsymbol{\psi}}(\mathbf{z}|\mathbf{x}_1)||\mathbf{p}_{\boldsymbol{\Theta}}(\mathbf{z})))+ \\
(\beta(L_2(\mathbf{x}_2 - \overline{\mathbf{x}}_2) + \sum_{j} D_{\text{KL}}(\mathbf{q}_{\boldsymbol{\psi}}(\mathbf{z}|\mathbf{x}_2)||\mathbf{p}_{\boldsymbol{\Theta}}(\mathbf{z})))+ \\
(\gamma(L_3(\mathbf{x}_3 - \overline{\mathbf{x}}_3) + \sum_{j} D_{\text{KL}}(\mathbf{q}_{\boldsymbol{\psi}}(\mathbf{z}|\mathbf{x}_3)||\mathbf{p}_{\boldsymbol{\Theta}}(\mathbf{z})))
\end{aligned}$$

The reconstruction losses are denoted by L_1, L_2, and L_3 for mRNA, CNV, and rppa data types respectively. The weights of each omics type (CNV, mRNA, and rppa) are represented by α, β, and γ, and the sum of α, β, γ should be equal to 1. The multi-omics data extracted from preprocessing stage is passed to the encoder, which encodes the data using a latent distribution. The performance of MVAE is evaluated using Deep Neural Network (DNN). The working of DNN is explained in Section 3.2.5. The learning rate of 0.05 and Adam optimizer are used. The MVAE runs for 100 epochs, returning a feature matrix as an output. The top 100 features are extracted for each type, from which the top 5 features are identified as the prognostic markers. The biological interpretation is done of the identified markers which is given in the following section.

5.1.3 Biological Interpretation

After training the dataset with MVAE for 100 epochs, the top 100 features are extracted for each fold. These features are then validated for biomarker identification using Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis from DAVID function enrichment analysis tool [160]. Along with that, survival analysis is also performed, which identifies those genes whose p-value is less than 0.05. The survival analysis is performed using Kaplan–Meier (KM) plotter [163]. The overall 10-year survival analysis is performed, which identifies the prognostic markers that are responsible for cancer recurrence in different sub-types of breast cancer.

5.1.4 Network Fusion

To uncover breast cancer subtypes by isolating similar participants from each other in multi-omics data, Similarity Network Fusion is proposed which is discussed in the following section.

5.1.4.1 Similarity Network Fusion (SNF)

The SNF [166] algorithm combines many types of omics data, constructing a network for each type of omics data, ultimately producing an all-encompassing perspective of the disease being studied. SNF is superior to other single-data analysis approaches because it can compute and combine Patient Similarity Networks (PSNs) for each data type. This enables the utilization of complementary information from multi-omics data. Specifically, the algorithm creates patient-patient similarity networks and computes patient-patient similarity matrices for each omics type. The next step is undertaking network fusion, which aims to improve strong connections while removing weak ones. In the end, a PSN that is fused is established. Let us assume that there are m patient samples with n omics data types, then for some $v_{\text{th}}(v = 1, \ldots, n)$ data type, an exponential similarity matrix will be calculated using the expression as follows:

$$\mathbf{W}(i,j) = \exp\left(-\frac{\rho^2(\mathbf{x}_i, \mathbf{x}_j)}{\mu\epsilon_{ij}}\right)$$

Where the euclidean distance between two samples $\mathbf{x}_i$ and $\mathbf{x}_j$ is given by $\rho(i,j)$.The similarity matrix $m \times m$ between two patients is denoted by $\mathbf{W}(i,j)$, the hyper-parameter and scaling factor required for eliminating the scaling problem is represented by μ and ϵ respectively. The similarity matrix $\mathbf{P}^{(v)}$ and KNN similarity

matrix $\mathbf{S}^{(v)}$ of all patients are calculated by the following expressions:

$$\mathbf{P}(i,j) = \begin{cases} \frac{\mathbf{W}(i,j)}{2\Sigma_{k\neq i}\mathbf{W}(i,k)}, & j \neq i \\ 12, & j = i \end{cases}$$

$$\mathbf{S}(i,j) = \begin{cases} \frac{\mathbf{W}(i,j)}{\Sigma_{k\in M_i}\mathbf{W}(i,k)}, & j \in M_i \\ 0, & \text{otherwise} \end{cases}$$

Suppose there are two datatypes, then the following process will be followed:

1. Calculate $\mathbf{P}^{(1)}$, $\mathbf{P}^{(2)}$, $\mathbf{S}^{(1)}$, $\mathbf{S}^{(2)}$. Let the initial status matrices at time $t = 0$ is represented by $\mathbf{P}^{(1)}_{t=0} = \mathbf{P}^{(1)}$, and $\mathbf{P}^{(2)}_{t=0} = \mathbf{P}^{(2)}$

2. Update the developed matrix recursively as follows:

$$\mathbf{P}^{(1)}_{t+1} = \mathbf{S}^{(1)} \times \mathbf{P}^{(2)}_{t} \times (\mathbf{S}^{(1)})^T,$$
$$\mathbf{P}^{(2)}_{t+1} = \mathbf{S}^{(2)} \times \mathbf{P}^{(1)}_{t} \times (\mathbf{S}^{(2)})^T$$

3. The complete matrix for t steps will be computed and is given by the following expression:

$$\mathbf{P}^{(t)} = \frac{\mathbf{P}^{(1)}_{t} + \mathbf{P}^{(2)}_{t}}{2}$$

For a larger number of data types, the above process can be easily generalized and is given as follows:

$$\mathbf{P}^{(v)} = \mathbf{S}^{(v)} \times \left(\frac{\Sigma_{k\neq v}\mathbf{P}^{(k)}}{m-1} \right) \times (\mathbf{S}^{(v)})^T$$

where the value of v is $1, 2, \ldots, n$. SNF efficiently merges similarity networks with paired patient similarity measurements into a single fused network called PSNs. In the current study, the extracted CNVs, mRNAs, and rppas from MVAE are used as input, and their corresponding similarity is calculated, which returns the adjacency matrix as an output. The adjacency matrix and feature matrix are passed to the modeling phase for breast cancer subtype classification and are given in the following section.

5.1.5 Modeling

The feature matrix from MVAE and adjacency matrix from SNF are passed as input to a Simplified Graph Convolutional Network (SGC) for cancer subtype

classification. The complete working of SGC is described in the given subsection.

5.1.5.1 Simplified Graph Convolutional Networks

Graph Convolutional Networks (GCNs) [172] accept as input a graph that already has some of its nodes labeled, and then they make label predictions for the rest of the graph's nodes [248]. Consider the graph $\mathcal{G} = (\mathcal{V}, \mathbf{A})$, where $\mathcal{V}$ denotes the vertex set containing nodes $(v_1, \ldots, v_n)$, and an adjacency matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ that is often sparse in which the weight of the edge between two nodes v_i and v_j is represented by a_{ij}. There is an associated d-dimensional feature vector $\mathbf{x}_i \in \mathbb{R}^d$ associated with each of the graph's nodes v_i. The whole feature matrix $\mathbf{x}_i \in \mathbb{R}^{n \times d}$ places n feature vectors one above the other in the form of a stacked list, denoted by the notation $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^\intercal$. Each node is a member of one of the C classes having C-dimensions with values 0 and 1. In this, labels for some of the nodes are known, and the goal of GCN is to predict the labels for the unknown nodes. GCNs, like Multi-layer Perceptron (MLP) [249], learn the new feature representation for each node's feature $\mathbf{x}_i$ over numerous layers and feed it into a linear classifier. For K-layer GCN, the input nodes and output nodes are represented by the matrix $\mathbf{H}^{(K-1)}$ and $\mathbf{H}^K$ respectively. The input to the first layer of GCN is given by the following expression:

$$\mathbf{H}^{(0)} = \mathbf{X}$$

To distinguish the K-layer GCN from MLP, a feature propagation method is used, which smoothes the hidden representations locally along the graph edges and promotes neighboring nodes to make similar predictions and is given as follows:

$$\mathbf{H}^K = \mathbf{S}\mathbf{H}^{(K-1)}$$

Here, the normalized adjacency matrix with added self-loops is denoted by $\mathbf{S}$. The smoothed hidden feature representations are linearly modified, and each layer is connected to a learned weight matrix $\mathbf{\Theta}^{(K)}$ [250]. Finally, a nonlinear activation function Rectified Linear Unit (ReLU (R)) is applied pointwise before outputting feature representation $\mathbf{H}^{(K+1)}$ and is given by the expression as follows:

$$\mathbf{H}^{K+1} = \mathrm{R}(\mathbf{S}\mathbf{H}^K\mathbf{\Theta}^K)$$

For the classification of the n nodes, a softmax classifier is utilized in the final layer of a GCN to predict the labels. The class predictions for n nodes is represented using the notation $\hat{\mathbf{Y}} \in \mathbb{R}^{n \times C}$, where $\hat{\mathbf{y}}_{ic}$ is the probability of node i belonging to

class c. For K-layer GCN, the class prediction $\hat{\mathbf{Y}}$ can be expressed by Eq. (5.3):

$$\hat{\mathbf{Y}}_{\text{GCN}} = \text{softmax}(\mathbf{S}\text{R}(\ldots\mathbf{S}\text{R}(\mathbf{S}\mathbf{X}\boldsymbol{\Theta}^1)\boldsymbol{\Theta}^2\ldots)\boldsymbol{\Theta}^K) \tag{5.3}$$

where

$$\text{softmax(x)} = \frac{exp(\mathbf{x})}{\sum_{c=1}^{C} exp(\mathbf{x}_c)}$$

is a normalizer that works across all classes. The normalized adjacency matrix and the input matrix of the first input node are represented by $\mathbf{S}$ and $\mathbf{X}$, respectively. A learned weight matrix is given by $\boldsymbol{\Theta}^1, \ldots, \boldsymbol{\Theta}^{(K)}$. The structure of GCN is quite complex. As a result, SGC [172] has been proposed to reduce the complexity. This will be accomplished sequentially by reducing nonlinearities and collapsing weight matrices between succeeding layers. It does this by smoothing the node input features by employing powers of the normalized adjacency matrix in conjunction with self-loops. It employs a single softmax layer, simplifying GCN on smoothed node features. The non-linear transition functions, i.e., ReLU (R) between each layer of GCN, is removed, and only the softmax function is considered to obtain the probabilistic output. Therefore, by using Eq (5.3), the final linear model is achieved, and the class prediction $\hat{\mathbf{Y}}$ for SGC is calculated, which is given by the following Eq. (5.4):

$$\hat{\mathbf{Y}}_{\text{SGC}} = \text{softmax}(\mathbf{S}\ldots\mathbf{S}\mathbf{S}\mathbf{X}\boldsymbol{\Theta}^{(1)}\boldsymbol{\Theta}^{(2)}\ldots\boldsymbol{\Theta}^{(K)}) \tag{5.4}$$

This notation given by Eq. (5.4) is simplified by raising the normalized adjacency matrix $\mathbf{S}$ to the K^{th} power, which results in a single matrix called $\mathbf{S}^K$. This action collapses the repeated multiplication with $\mathbf{S}$ into a single matrix. In addition, we can reparameterize our weights ($\boldsymbol{\Theta}^{(1)}\boldsymbol{\Theta}^{(2)}\ldots\boldsymbol{\Theta}^{(K)}$) into a single matrix ($\boldsymbol{\Theta}$) and the resulting SGC is represented by the following expression:

$$\hat{\mathbf{Y}}_{\text{SGC}} = \text{softmax}(\mathbf{S}^{(K)}\mathbf{X}\boldsymbol{\Theta})$$

In the given study, the features extracted from MVAE and adjacency matrix from SNF are passed as nodes and edges to SGC, which runs for 150 epochs with a learning rate of 0.01. The transformation-based integration is used which is explained in Section 1.2.1.1. 10-fold cross-validation is used, and the performance is evaluated by taking the average of each fold. The working of cross-validation is discussed in Section 3.2.5.8. The fused CNV, mRNAs, and rppas are divided into 10-fold cross-validation where one fold is used as a test set and 9 folds are

used as a train set. The performance is evaluated and recorded. This is done until all the datasets have been explored through train and test fold. The average of all the folds is computed as the final output. Five parameters comprising Accuracy, Sensitivity, specificity, precision, and F1-score are computed to evaluate the performance of the proposed iMVAN model. The complete working of iMVAN is given in Algorithm 5.1.

Algorithm 5.1 Pseudocode of iMVAN

Input: Dataset D (CNV, mRNA, rppa), $D = \mathbf{x}^1, \ldots, \mathbf{x}^n$, Class Prediction $\mathbf{Y}$, Initial Parameters ($\boldsymbol{\Theta}$,$\boldsymbol{\psi}$), $\mathbf{z}$

Output: Top Biomarkers N, Performance Parameters

Begin:

Divide D in training and testing set
for $i \leftarrow 1$ to 100 **do**
    Pass input ($\mathbf{x}_i$) to probabilistic encoder ($\mathbf{q}_{\boldsymbol{\psi}}(\mathbf{z}|\mathbf{x})$)
    Extract Latent Data L from $\epsilon \sim \boldsymbol{\eta}(0, 1)$
    Decode input using probabilistic decoder ($\mathbf{p}_{\boldsymbol{\Theta}}(\mathbf{z}|\mathbf{x})$)
    Update KL divergence Loss
    Return Features F_i
end for
Calculate p-value and survival analysis
if p-value$_{F_i} < 0.05$ **then**
    $N \leftarrow$Top Prognostic Markers
else
    Normal Feature
end if
for $j \leftarrow 1$ to D **do**
    Compute Similarity between D_j
    Return Adjacency Matrix that is $\mathbf{S}_j \leftarrow$ Adjacency Matrix
end for
for all nodes F_i and edges $\mathbf{S}_j$ **do**
    $\mathbf{Y} \leftarrow$ SGC(F_i,$\mathbf{S}_j$) ▷ pass F_i,$\mathbf{S}_j$ to SGC for training
    Compute Performance Parameters
end for

End

Moreover, the superiority of SGC against GCN is checked through network

parameters and floating point operations (FLOPs) which is discussed in the given section.

5.1.5.2 Network Parameter and Floating Point Operations (FLOPs) Calculation

Network parameters are the variables that are acquired by a neural network through the process of training. The objective of calculating network parameters is to minimize the discrepancy between the network's predictions and the desired output. These are computed as follows:
In GCN, there are three layers, and each has its own weight matrix and bias vector [251]. Given N nodes, F features, and H hidden layer units, the total number of network parameters are:

- Input layer weight matrix, with $F \times H$ parameters
- Input layer bias vector, with H parameters
- Hidden layer weight matrix, with $H \times H$ parameters
- Hidden layer bias vector, with H parameters
- Output layer weight matrix, with $H \times K$ parameters
- Output layer bias vector, with K parameters

The final result is determined by computing the total of all the above parameters.

SGC, on the other hand, consists of two sets of weight matrices, one set for each layer and one bias vector for each layer [172]. Assuming that the input graph includes N nodes (samples), F features, and H units in the hidden layer, the total number of network parameters are as follows:

- Input layer weight matrix, with $F \times H$ parameters
- Input layer bias vector, with H parameters
- hidden layer weight matrix, with $H \times K$ parameters
- hidden layer bias vector, with K parameters

The outcome is calculated by adding up all of these parameters. In these network parameters, the bias vector value is assumed to be equal to the number of hidden layers in each layer.

FLOPs Calculations: FLOPs serve as a metric for quantifying the computing burden associated with executing mathematical computations involving real numbers, specifically those represented as floating-point integers. Within the domain of DL, the utilization of FLOPs serves as a means to approximate the computing expenditure associated with executing a neural network. FLOPs are commonly quantified in terms of arithmetic operations involving floating-point numbers, such as additions, subtractions, multiplications, and divisions. The quantity of FLOPs varies according to the size of the input graph and the number of output classes. Mathematically, it is expressed as:
In GCN, if the input graph contains N nodes and F features, and the output is a K-dimensional, one-hot vector, then the total number of FLOPs is:

- Graph input: $N \times F$ FLOPs
- Activations of the hidden layer: $N \times H$ FLOPs
- Activations of the output layer: $N \times K$ FLOPs

The sum of the parameters mentioned above is computed as the final result.
While in SGC, the number of FLOPs is:

- Input graph: $N \times F$ FLOPs
- Output Classes: $N \times K$ FLOPs

The end result is the sum of the above parameters. In the current study, the network operations and FLOPs are computed to prove the superiority of SGC over GCN.

5.2 Experimental Setup and Results

5.2.1 Experimental Steps

The minimum hardware requirement to implement the work is 8 GB RAM with an i5 processor. Python 3.9.0 is used to implement the work. DAVID functional Analysis [160] is used to perform pathway analysis. KM plotter [163] is used to perform survival analysis. The libraries used to perform the implementation are NumPy, pandas, torch, SNF, sklearn, and matplotlib. The scikit-learn library (https://scikit-learn.org/stable/) is used to implement the DT, RF, and DNN.

5.2.2 Experimental Steps

The following steps have been implemented for iMVAN based biomarker identification for disease subtype classification.

- Multi-omics data comprising mRNA, CNV, and rppa dataset of BRCA patients have been utilized with BRCA divided into four subtypes. The data preprocessing is performed using KNNimputer and z-scale normalization.
- Secondly, MVAE is presented to reduce the dimensionality of high-dimensional feature set. The extracted features are validated using KEGG and survival analysis test which identifies the prognostic markers for the corresponding BRCA subtype.
- SNF is presented to integrate the latent features from MVAE which returns a fused adjacency matrix.
- Ultimately, the MVAE and SNF output is passed to SGC for model training and testing and the performance is evaluated using five parameters comprising accuracy, recall, specificity, precision, and F1-score which are discussed in Section 3.2.6.

5.2.3 Experimental Data

The identified markers and features from MVAE and adjacency matrix from SNF were passed to the SGC for training, as described in Section 5.1. The integrated dataset, including mRNA, CNV, and rppa, is divided into 70:30 ratios, meaning 70% is used for training and 30% for testing. A 10-fold cross-validation experiment is performed on the training data to find the optimal model parameters. Then, the pre-trained model is utilized to make predictions on the testing set [252, 253]. In the present research, there are 511 patients, in which 358 samples are used for training, and 153 samples are used for testing the model. 10-fold cross-validation is applied on 358 samples, and the optimal model parameters are obtained. This pre-trained model is then applied to 153 testing samples to make the final predictions. Table 5.2 shows the division of samples with total patients.

5.2.4 Results

The results are divided into two parts, MVAE results and SGC results. At first, the results of MVAE are computed on the TCGA BRCA multi-omics dataset using DNN. The performance parameters comprising accuracy and F1-score are

Table 5.2: Division of Samples in training and testing

Sample	**No. of Patients**
Total Samples	511
Training Samples	358
Testing Samples	153

used to calculate the performance of MVAE. MVAE is evaluated on single omics, that is, CNV, rppa, and mRNA alone, and on integrated omics (CNV + mRNA + rppa) data. The results are compared with some existing dimensionality reduction techniques, including principal component analysis (PCA), independent component analysis (ICA), factor analysis (FA), and autoencoders (AE), which are implemented using the Python scikit-learn library. The results are shown in Table 5.3.

Table 5.3: Performance of MVAE in comparison with existing dimensionality reduction methods

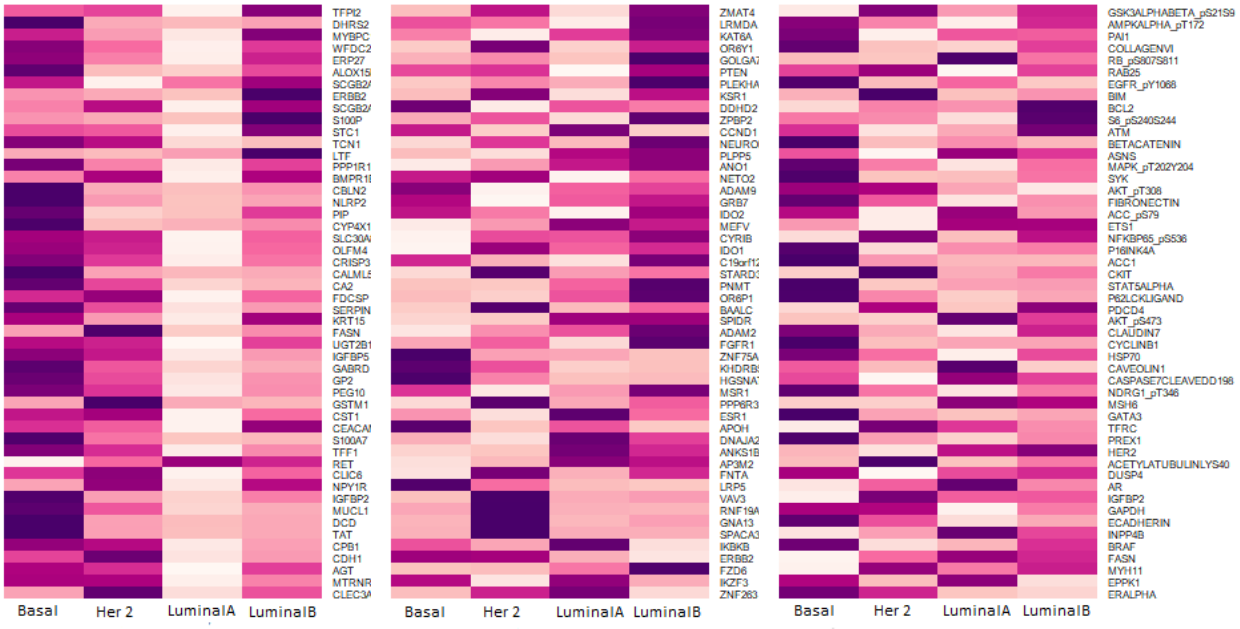
Method	**Parameters**	**CNV**	**mRNA**	**rppa**	**CNV + mRNA + rppa**
PCA	**Accuracy**	0.60 ± 0.04	0.83 ± 0.04	0.77 ± 0.01	0.82 ± 0.04
	F1-score	0.55 ± 0.05	0.80 ± 0.05	0.73 ± 0.03	0.80 ± 0.04
ICA	**Accuracy**	0.50 ± 0.03	0.68 ± 0.03	0.54 ± 0.02	0.62 ± 0.04
	F1-score	0.42 ± 0.04	0.62 ± 0.03	0.44 ± 0.04	0.56 ± 0.05
FA	**Accuracy**	0.59 ± 0.02	0.78 ± 0.03	0.74 ± 0.04	0.70 ± 0.04
	F1-score	0.54 ± 0.02	0.74 ± 0.04	0.70 ± 0.05	0.65 ± 0.04
AE	**Accuracy**	0.56 ± 0.04	0.83 ± 0.03	0.80 ± 0.04	0.85 ± 0.04
	F1-score	0.52 ± 0.05	0.81 ± 0.05	0.79 ± 0.04	0.85 ± 0.05
MVAE	**Accuracy**	$\mathbf{0.62 \pm 0.03}$	$\mathbf{0.85 \pm 0.01}$	$\mathbf{0.82 \pm 0.05}$	$\mathbf{0.88 \pm 0.03}$
	F1-score	$\mathbf{0.58 \pm 0.04}$	$\mathbf{0.83 \pm 0.04}$	$\mathbf{0.81 \pm 0.02}$	$\mathbf{0.87 \pm 0.03}$

10 fold cross-validation (mean ± standard deviation)

The results evidence that the MVAE reduces the dimensionality and extracts features with an accuracy of 88% and an F1-score of 87%, respectively, on the integrated dataset. The reason behind the best performance of MVAE is that multi-omics data include a considerable amount of noise; therefore, the relative information density is poor, which hinders the performance of conventional al-

gorithms. In addition, standard algorithms are linear approaches that cannot identify possible nonlinear correlations in complex biological data.

Additionally, to show the importance of top markers extracted by MVAE, heatmaps are designed for CNV, mRNA, and rppa features as shown in Figure 5.2. The intensity of the color used in heatmaps shows the relative significance of



(i) CNV Heatmap (ii) mRNA Heatmap (iii) rppa Heatmap

Figure 5.2: Heatmaps for extracted CNV, mRNA, and rppa genes [3]

each feature compared to the other features. It is visible from the heatmaps that most CNVs are significant in the basal and HER2 subtypes. mRNA transcripts are essential in the Luminal B subtype, and rppa proteins are equally important in each subtype. Moreover, to identify the top genes, that is, genes with the highest priority, KEGG analysis, and survival analysis is performed, which identify the prognostic markers.

5.2.4.1 Biological Interpretation of identified biomarkers

The identified features from MVAE are validated using KEGG and survival analysis tests. The complete working of KEGG analysis and survival analysis is given in Section 3.2.4. In KEGG, the pathway analysis is performed, which identified five genes comprising GSTM1, AGT, CDH1, RET, and CALML5 with a p-value less than 0.05. GSTM1 plays a vital role in the Her2 subtype of breast cancer.

GSTM1 is present in locus with 3 KB in size, and it can lead to an increased breast cancer risk [163]. AGT is highly important in the basal subtype of breast cancer and is identified as a prognostic marker [254] in colorectal cancer. CDH1 is present in the Her2 subtype of breast cancer, and it is identified as a driver gene by [255], which can lead to tumor progression. RET is highly important in Luminal A and B and is identified as a candidate marker in the advancement of breast cancer [256]. CALML5 played an essential role in the basal subtype and is identified as a poor prognostic marker in breast cancer [257].

Moving ahead, the pathway analysis of mRNA's has been done, and seven markers comprising ERBB2, PTEN, ESR1, CCND1, FZD6, LRP5, and FZFR1 are identified. ERBB2 is highly important in Her2 and Basal subtypes of breast cancer. ERBB2 is recognized as a poor prognostic marker and plays a vital role in tumor aggressiveness [258]. PTEN plays a crucial role in basal, Her2, and Luminal B subtypes followed by CCND1, which is essential in all subtypes and identified as a molecular marker of hormone responsiveness [259]. PTEN is identified by [260] as an excellent prognostic marker and has a vital role in the prevention of breast cancer. ESR1 is found in the Luminal A subtype and is identified as a prognostic marker in metastatic breast cancer [261]. Moving ahead are the FGFR1 and FZD6, which play an essential role in the Luminal B subtype. FGFR1 is identified as a prognostic and predictive marker in Luminal B breast cancer [262], and FZD6 leads to an aggressive increase in neuroblastoma cells [263]. It is identified as a new surface marker. LPR5 is found in the basal subtype. It is recognized as a tumor suppressor gene in metastatic breast cancer [264].

Similarly, pathway analysis of rppa data is performed, which identified five markers comprising BRAF, BCL2, AR, ETS1, and MSH6 with a p-value less than 0.05. AR and ETS1 are key markers in the Luminal A subtype followed by BCL2, which plays a significant role in the Luminal B subtype. BRAF is found in the basal subtype, and MSH6 is almost equally important in the Luminal A and B subtypes. AR is identified as a unique prognostic marker for targeted therapy [265]. BCL2 protein expression acts as an indicator of a favorable prognosis in Breast Cancer. The link between BCL2+ and Estrogen Receptor (ER+) has been used to explain why BCL2+ patients have a higher survival rate [266]. MSH6 is identified as a risk biomarker with an increased risk of breast cancer [267]. BRAF is spotted as a driver gene in the Her2 subtype of breast cancer [268]. ETS1 is recognized as a prognostic marker for relapse-free survival in breast cancer [269].

Furthermore, the genes are then passed to the survival analysis test. The 10-year overall survival is selected. The Hazard Ratio (HR) is computed, which

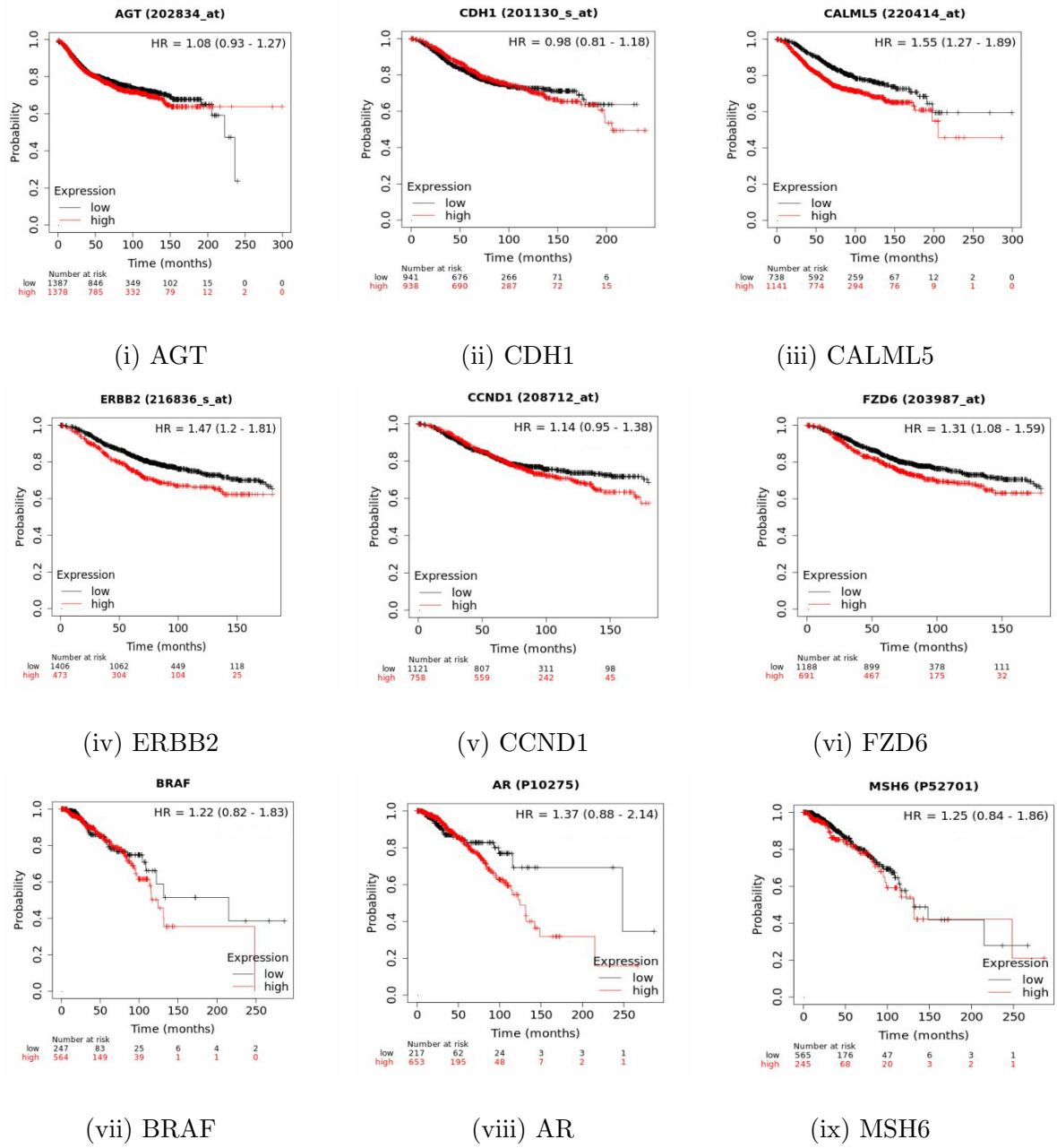
depicts that the patients with higher hazard, that is, close to one or greater than one are at higher risk of death.

The p-value and hazard ratio are calculated for each identified gene and are shown in Table 5.4.

Table 5.4: Survival Analysis of Identified Markers

Data Type	**Biomarker**	**P-value**	**Hazard Ratio**	**Breast Cancer Subtype**
CNV	GSTM1	0.008	0.69 (0.62-0.76)	Her2
	AGT	0.0065	1.08 (0.93-1.27)	Basal
	CDH1	0.0004	0.98 (0.81-1.18)	Her2
	RET	0.04	0.82 (0.68-0.99)	Luminal A and Luminal B
	CALML5	0.01	1.55 (1.27-1.89)	Basal
mRNA	ERBB2	0.00021	1.47 (1.2-1.81)	Basal and Her2
	PTEN	0.018	0.72 (0.55-0.95)	Basal, Her2 and Luminal B
	ESR1	0.0014	0.58 (0.48-0.71)	Luminal A
	CCND1	0.0017	1.14 (0.95-1.38)	Basal, Her2, Luminal A, and B
	FZD6	0.0056	1.55 (1.08-1.59)	Luminal B
	LRP5	0.023	0.79 (0.65-0.97)	Basal
RPPA	BRAF	0.033	1.22 (0.82-1.83)	Basal
	BCL2	0.0017	0.56 (0.39-0.81)	Luminal B
	AR	0.016	1.37 (0.88-2.14)	Luminal A
	ETS1	0.032	0.65 (0.44-0.97)	Luminal A
	MSH6	0.026	1.25 (0.84-1.86)	Luminal A and Luminal B

From survival analysis, it is found that out of a total of 17 markers, nine markers comprising AGT, CDH1, CALML5, ERBB2, CCND1, FZD6, BRAF, AR, and MSH6 have hazard ratios close to one and greater than one, which shows that these markers are identified as poor prognostic markers. The survival analysis plots of identified markers have been shown in Figure 5.3.



(i) AGT (ii) CDH1 (iii) CALML5

(iv) ERBB2 (v) CCND1 (vi) FZD6

(vii) BRAF (viii) AR (ix) MSH6

Figure 5.3: Survival analysis plots for poor prognostic markers [3]

5.2.4.2 Disease subtype classification results

To classify cancer data into its subtypes, the extracted latent data from MVAE and similarity matrix from SNF have been utilized and passed to SGC. The dataset is divided into 70:30 training and testing, which means 70% is used for training data, and 30% is used for testing data. 10-fold cross-validation is used on training data to optimize the model parameter, and the mean accuracy of 10-fold is the

estimated result. The pre-trained model is then used to test the performance of testing data. The performance is evaluated using five parameters comprising accuracy, recall, specificity, precision, and F1-score. The results of iMVAN have been compared with DT, RF, DNN, SGC + MVAE, and SGC + SNF, and it is evident from the results that iMVAN performed better with an accuracy of 88% and an F1 score of 89%. The results of iMVAN are shown in Table 5.5 below.

Table 5.5: Performance Evaluation of iMVAN

Work	**Accuracy**	**Recall**	**Specificity**	**Precision**	**F1-score**
DT	0.74 ± 0.04	0.78 ± 0.03	0.80 ± 0.05	0.75 ± 0.01	0.69 ± 0.02
RF	0.85 ± 0.03	0.84 ± 0.01	0.86 ± 0.04	0.83 ± 0.02	0.84 ± 0.01
DNN	0.85 ± 0.02	0.85 ± 0.03	0.86 ± 0.02	0.84 ± 0.02	0.85 ± 0.03
SGC + SNF	0.79 ± 0.03	0.81 ± 0.02	0.79 ± 0.03	0.76 ± 0.05	0.80 ± 0.04
SGC + MVAE	0.85 ± 0.05	0.89 ± 0.03	0.90 ± 0.03	0.86 ± 0.02	0.86 ± 0.03
iMVAN	0.87 ± 0.04	0.90 ± 0.05	0.91 ± 0.03	0.88 ± 0.03	0.88 ± 0.03

Moreover, the bar plots have been plotted for accuracy, recall, specificity, precision, and F1-score to show the effectiveness of the iMVAN and are shown in Figure 5.4. It is visible from the plots that the iMVAN performed well in terms of accuracy, recall, specificity, precision, and F1-score values of 87%, 90%, 91%, 88%, and 88%, respectively.

5.2.4.3 Validation of iMVAN on KIPAN and CESC

To validate the proposed iMVAN, two cancers comprising TCGA KIPAN (Pan Kidney Cohort) and TCGA Cervical and Endocervical Cancer (CESC) have been used. The total number of patients in KIPAN and CESC is 746 and 163, respectively. The KIPAN is divided into three subtypes comprising Kidney Clear Cell Carcinoma (KIRC), Kidney Chromophobe (KIHC), and Kidney Papillary Cell Carcinoma (KIRP). Similarly, Cervical Cancer is composed of two subtypes, including Cervical Squamous Cell Carcinoma (CSCC) and Endo Cervical Adenocarcinoma (ECA). The CNV, mRNA, and rppa datasets from TCGA [246] for TCGA KIPAN and TCGA CESC are used and preprocessed using KNNimputer and z-scale normalization. Then MVAE is applied and, results have been computed, which are shown in Table 5.6. It is visible from the results that MVAE

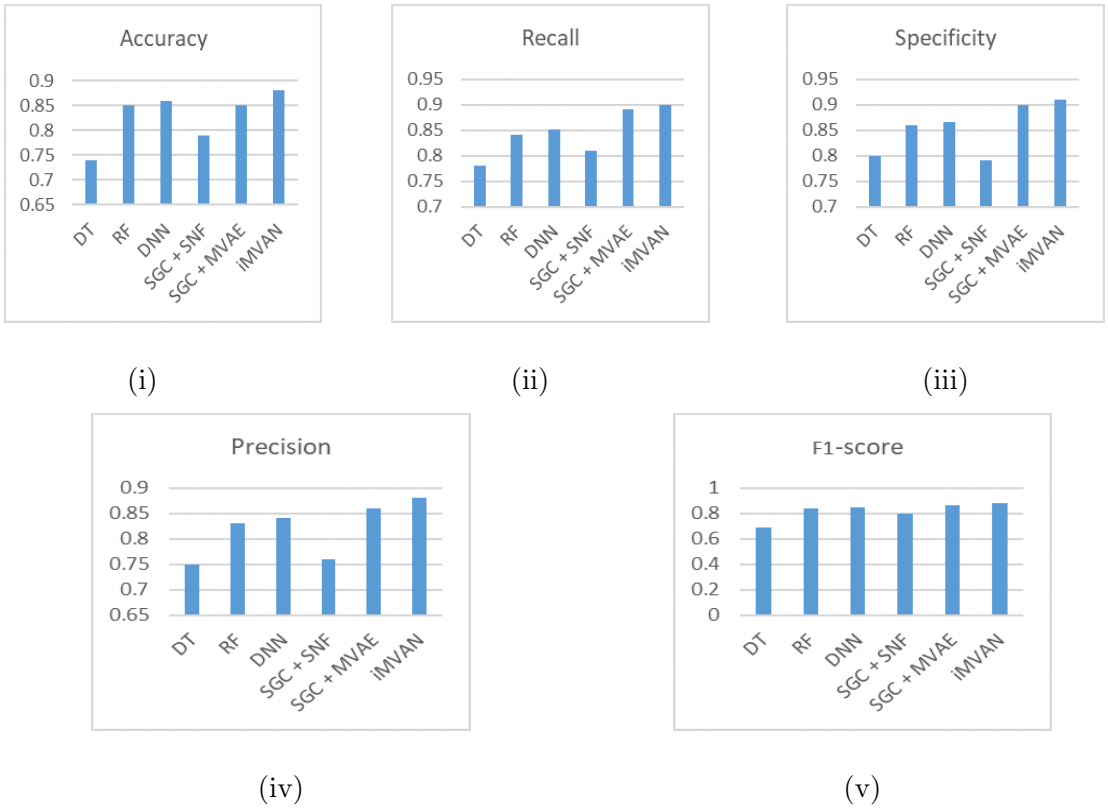


Figure 5.4: Bar plots for accuracy, recall, specificity, precision, and F1-score [3]

performed well on KIPAN and CESC with an accuracy of 92% and 86%, respectively. Table 5.6 shows the results of MVAE for TCGA KIPAN and TCGA CESC.

Table 5.6: Performance of MVAE on TCGA KIPAN and TCGA CESC

Dataset	**TCGA KIPAN**		**TCGA CESC**	
	Accuracy	**F1-score**	**Accuracy**	**F1-score**
CNV	0.68 ± 0.03	0.62 ± 0.04	0.61 ± 0.04	0.60 ± 0.04
mRNA	0.87 ± 0.02	0.85 ± 0.03	0.85 ± 0.02	0.84 ± 0.03
rppa	0.85 ± 0.02	0.83 ± 0.03	0.80 ± 0.04	0.79 ± 0.04
CNV + mRNA + rppa	0.92 ± 0.02	0.90 ± 0.03	0.86 ± 0.03	0.85 ± 0.02

Further, the KEGG and survival analyses are also applied to TCGA KIPAN and TCGA CESC patients. The pathway analysis identified four CNV markers comprising FGB, C6, FGG, and PLG, four mRNA markers including EFNA5, EFNA5, ROBO1, SEMA3E, and five rppa markers consisting of BRAF, MAPK9, CDKN1A, BAK1, and EGFR for TCGA KIPAN and four CNV markers including MMP13, CXCL6, CXCL5, MMP1 and three mRNA markers including BIRC2, BIRC3, MAPK10 and five rppa markers including RAD51, MYC, DVL3, ETS1,

VHL for TCGA CESC patients. Additionally, survival analysis is also performed on TCGA KIPAN and TCGA CESC, which identified six poor survival markers comprising EFNA5, CADM2, SEMA3E, FGB, FGG, and BAK1 for TCGA KIPAN and six poor survival markers including MMP13, CXCL6, CXCL5, MMP1, BIRC2, and MYC for TCGA CESC patients respectively and is shown in Table 5.7.

Table 5.7: Survival Analysis Results on TCGA KIPAN and TCGA CESC

Data Type	**TCGA KIPAN**				**TCGA CESC**			
	Marker	**P-value**	**HR**	**Sub-type**	**Marker**	**P-value**	**HR**	**Sub-type**
CNV	EFNA5	0.0008	2.28 (1.69-3.08)	KIRP	MMP13	0.014	1.44 (0.88-2.34)	CSCC
	CADM2	0.0075	1.31 (0.97-1.77)	KIRC	CXCL6	0.038	1.64 (1.02-2.64)	ECA
	ROBO1	0.0007	0.59 (0.44-0.81)	KIRP	CXCL5	0.0028	2.34 (1.46-3.74)	CSCC
	SEMA3E	0.0001	2.28 (1.69-3.08)	KIHC	MMP1	0.0008	2.31 (1.39-3.83)	CSCC
mRNA	FGB	0.011	1.34 (0.93-1.93)	KIHC	BIRC2	0.037	1.26 (0.76-2.08)	ECA
	C6	0.0029	0.63 (0.47-0.86)	KIRC	BIRC3	0.038	0.81 (0.51-1.3)	ECA
	FGG	0.008	1.33 (0.97-1.83)	KIHC	MAPK10	0.017	0.69 (0.41-1.17)	CSCC
	PLG	0.0006	0.4 (0.29-0.55)	KIRP	-	-	-	-
RPPA	EGFR	0.0015	0.6 (0.43-0.82)	KIRP	RAD51	0.0082	0.43 (0.23-0.82)	ESA
	BRAF	0.0001	0.56 (0.42-0.76)	KIRC	MYC	0.005	1.98 (1.21-3.23)	CSCC
	MAPK9	0.00011	0.54 (0.4-0.74)	KIRP	DVL3	0.013	0.65 (0.37-1.15)	ESA
	CDKN1A	0.0031	0.62 (0.45-0.85)	KIHC	ETS1	0.029	0.75 (0.44-1.27)	CSCC
	BAK1	0.028	1.87 (1.38-2.54)	KIRP	VHL	0.085	0.66 (0.41-1.06)	ECA

The extracted features from MVAE are combined with the results of SNF, which are then trained using SGC, and final predictions for iMVAN have been made. From the findings, it has been found that iMVAN performed well on TCGA KIPAN and TCGA CESC with 0.999 ± 0.1 and 0.87 ± 0.02 respectively. Table 5.8 showed the results of iMVAN on TCGA KIPAN and TCGA CESC respectively. The sensitivity, specificity, precision, and F1 score is also computed, which is $(0.99 \pm 0.2, 0.99 \pm 0.009, 0.98 \pm 0.02,$ and $0.99 \pm 0.01)$ and $(0.88 \pm 0.2, 0.89 \pm$

0.04, 0.88±0.4, and 0.90±0.01) respectively for TCGA KIPAN and TCGA CESC.

Table 5.8: iMVAN results on TCGA KIPAN and TCGA CESC

Parameter	**TCGA KIPAN**	**TCGA CESC**
Accuracy	0.999 ± 0.1	0.87 ± 0.02
Sensitivity	0.99 ± 0.2	0.88 ± 0.2
Specificity	0.99 ± 0.009	0.89 ± 0.04
Precision	0.98 ± 0.02	0.88 ± 0.4
F1-score	0.99 ± 0.01	0.90 ± 0.01

5.3 Computational Effectiveness of iMVAN

To show the effectiveness of presented SGC over GCN, network parameters and FLOPs are calculated. The calculations of network parameters and FLOPs for SGC and GCN are described in Section 5.1.5.2. In the current research, MVAE extracts 100 features for each data type comprising CNV, mRNA, and rppa. Therefore, a total of 300 features (F) are there, with 511 samples (N) and four classes (K) for the TCGA BRCA dataset. A total of 64 hidden layers (H) have been used in the proposed model. The total parameters for SGC are computed as follows:

Input Layer parameters = 300×64

Input Layer bias = 64

Hidden Layer parameters = 64×3

Hidden Layer bias = 3

So, by computing the total, 19,459 network parameters have been achieved. On the other hand, GCN has a total of 23,555 parameters, much more than the SGC parameters. Hence, it shows that SGC is far more effective than GCN.
Similarly, FLOPs are also calculated for SGC and GCN. For SGC, the total FLOPs for the input graph and output classes are (511×300) and (511×3), respectively. Hence, 1,54,833 FLOPs have been achieved. For GCN, total FLOPs are 1,87,537, clearly showing that GCN is more complex than SGC.

Additionally, the results of iMVAN have been compared with existing works comprising [270] and [271]. In these existing works, a 10-fold cross-validation method is used as a verification method. iMVAN also utilized a 10-fold cross-validation method to tune the model's parameters. The results are computed, and it proves that the iMVAN performed well with an improvement of 6% and 4%, respectively in terms of accuracy. The results for TCGA KIPAN and TCGA CESC have also been computed, proving that the iMVAN performed efficiently with an accuracy of 0.999 ± 0.1 and 0.87 ± 0.02 respectively. The results of the comparison of iMVAN with existing work are shown in Table 5.9. The bar plot has also been plotted to

Table 5.9: Comparison of iMVAN with Existing Work

Dataset	**Work**	**Accuracy**
TCGA BRCA	[270]	82%
	[271]	84%
	iMVAN	87%
TCGA KIPAN	[270]	99%
	iMVAN	99.99%
TCGA CESC	[272]	84%
	iMVAN	87%

depict the efficacy of iMVAN and is shown in Figure 5.5.

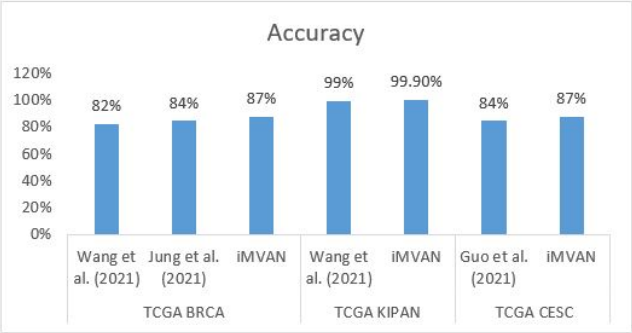


Figure 5.5: Bar plot for comparison of existing work with iMVAN [3]

5.4 Conclusion

This chapter presented the biomarker identification and disease subtype classification using the developed iMVAN approach. The iMVAN is developed using MVAE, SNF, and SGC in which MVAE is used to reduce the dimensionality of high-dimensional multi-omics datasets comprising CNV, mRNA, and rppa for TCGA BRCA patients. KEGG and survival analysis on reduced features is performed to identify the prognostic markers. The SNF is employed to integrate the multi-omics datasets which is a transformer-based integration technique. The output from MVAE and SNF is passed to SGC for subtype classification of BRCA patients. The iMVAN is validated on TCGA KIPAN and TCGA CESC subtypes. The enhanced and accurate performance of biomarker identification and disease subtype classification using iMVAN can be helpful in recommending treatment to a patient.

In the next chapter, the HBS-STACK approach for biomarker identification and disease prediction in multi-omics datasets is discussed.

Chapter 6

HBS-STACK: Hierarchical Biomarker Selection and Stacked Ensemble for Disease Prediction

In the previous Chapter, a proposed iMVAN based integrative multimodal variational autoencoder approach for multi-omic biomarker identification for disease subtype classification is discussed. The iMVAN demonstrates remarkable efficiency to significantly contribute to patient care through accurate identification of biomarkers in multi-omics data for disease subtype classification of Breast, Pan kidney, and Cervical Cancer.

In this Chapter, a detailed overview of the developed HBS-STACK approach based on the proposed framework using a hierarchical biomarker selection and stacked ensemble for disease prediction in multi-omics data is discussed in detail. The design of HBS-STACK is built using data acquisition, data preprocessing, hierarchical feature selection, biological significance, modeling using a stacked ensemble, and performance evaluation. In the chapter, multi-omics datasets of multiple diseases are taken to validate the HBS-STACK approach.

This chapter starts with the discussion of the proposed HBS-STACK approach in Section 6.1. The experimental analysis and results are given in Section 6.2. The results are discussed in Section 6.3. At the end, conclusion is provided in Section 6.4

6.1 Overview of HBS-STACK approach

This section introduced and discussed in detail the HBS-STACK approach based on hierarchical biomarker selection and stacked ensemble model for disease prediction in multi-omics data. The HBS-STACK approach is designed using six phases, involving Multi-Omics Data Collection, Data Preprocessing, Feature/ Biomarker Selection, Biological Interpretation, Modelling, and Performance Evaluation phase as shown in Figure 6.1 and are explained in detail below.

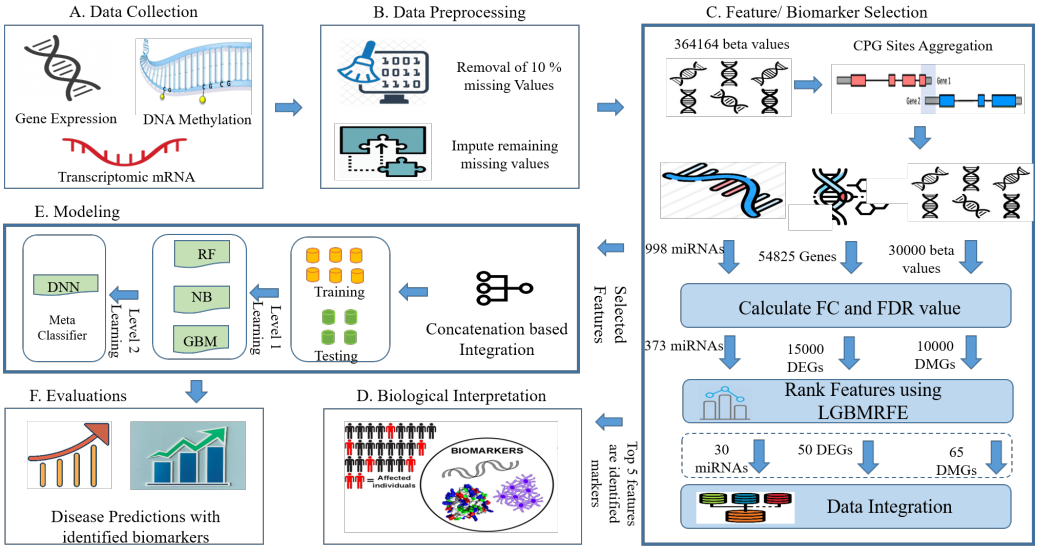


Figure 6.1: Workflow of HBS-STACK [4]

6.1.1 Data Collection

Multi-omics dataset including Gene Expression and miRNA at the transcriptomic level, and DNA Methylation (DM) at the genomic level for the identification of the biomarkers related to breast cancer prediction. First, the publicly available portal TCGA (The Cancer Genome Atlas Portal) is used to download the data. The complete description of TCGA is given in Section 3.2.1.1. This downloaded dataset consists of a different number of samples. The gene expression, DM, and miRNA comprehend 1222, 895, and 1207 samples. Therefore, to find the common patients, the Venn diagram is used, and finally, 841 patients have been obtained with 766 patients having solid tumors and 75 patients with non-tumors. The HTSeq-Counts value is used for gene expression, and the total number of features is 56602. For DM, Illumina Human Methylation 450 array methylation chip data is used in which 4,85,577 probes are there, which covers almost 96% of CpG islands. Similarly, for RNA, miRNA expression values are used with 1372 features. To download the dataset, the TCGABiolinks package has been used. The downloaded dataset is present in the raw form. It contains thousands of missing values which need to be removed for further processing. Therefore, data preprocessing is required which is discussed in the following section.

6.1.2 Data Preprocessing

Data Preprocessing is the process of data cleaning, handling the missing values, and removal of redundant features. The complete description of data preprocess-

ing is given in Section 3.2.2. Firstly, the 10% NA values for gene expression, DM, and miRNA have been removed. This is done using na.omit function. This will search for those rows that have more than 10% NAs and directly delete them from the dataset. Then, the remaining values are imputed using the KNNimputer [273] function which imputes the smallest distance in place of the remaining NA values. After imputing the missing values, the remaining features in gene expression, DM, and miRNA are 54825, 364051, and 999, respectively. The values in gene expression, miRNA, and DM are inconsistent. Therefore, to provide consistency between data elements, normalization is done using normalize between array technique. Normalize between arrays considers information from all samples within a dataset and establishes a measure to normalize them collectively. This is mostly used for multi-omics data analysis. The Limma package of R is used to perform the normalize between array operation on gene expression, miRNA, and DM dataset of BRCA patients. It is a challenging task to identify features with great importance for disease prediction. Hence, feature/ biomarker selection techniques are required which are discussed in the following section.

6.1.3 Feature/ Biomarker Selection

Feature selection is selecting the features or markers with utmost importance. Developing models with datasets having more features compared to samples leads to a problem of overfitting and poor performance of prediction. To tackle this challenge of overfitting, feature selection is adopted as it reduces the feature space by selecting appropriate subset/s. The complete description of the feature selection technique is given in Section 3.2.3.1. To solve the overfitting problem, a hierarchical feature/ biomarker selection approach is provided for biomarker selection in multi-omics data required for disease prediction. First, the DM dataset contains the highest number of features, i.e., approximately 3 lakh features. In these DMs, several CPG sites annotate to a single gene. Therefore, only one gene for several CPGs can be used for processing. Therefore, to select only one gene, Aggregate information between CPG site methods is proposed. Still, it contains thousands of features. Similarly, gene expression and miRNA value contain features in thousands. Therefore, two other feature selection techniques including statistical analysis tests (Fold Change (FC) and False Discovery Rate (FDR)) and a wrapper method approach named light gradient boosting machine with recursive feature elimination (LGBMRFE) have been employed. This three-phase feature selection is applied in a hierarchical manner and is described in detail in the following subsections.

6.1.3.1 Aggregate information between CpG sites and genes

DM data is high dimensional, so working with such a vast dataset is very difficult. Furthermore, a gene can be associated with multiple CpG sites, leading to redundant features. Therefore, to reduce DM's dimension and remove redundancy, methylated genes have been identified by finding the complex biological relations between the genes and the CpG sites. Mathematically, it worked as follows: Let CpG is denoted by c and the value associated is given by v, then the average of v for each c is given by Eq. (6.1) as follows:

$$\text{Average} = \frac{1}{n}\sum_{i=1}^{n} v_i^c \tag{6.1}$$

Where n is the total CpG sites. Once the average is computed, the methylated genes (MG) are calculated. For this, a threshold is selected and if the average of similar probes is greater than the threshold, then that gene is considered a methylated gene. The methylated genes are calculated by fusing the avereps function of the limma package, which calculates the aggregated value for each CpG site. This reduces the dimension of the DM to 30000 features.

6.1.3.2 Fold Change (FC) and False Discovery Rate (FDR)

FC and FDR are the statistical tests employed to reduce the dimensionality of high-dimensional multi-omics datasets. The complete detail of FDR and $log_2(FC)$ is given in Section 3.2.3.3. The tests are applied to miRNAs, DMs, and gene expression datasets of BRCA features, and only those features have been selected where $|log_2(FC) > 0.5|$ and $FDR < 0.05$. This step returns 15924, 10000, and 373 features of gene expression, DM, and miRNA expression, respectively. This is still a huge feature set. Selecting only those features that are responsible for disease diagnosis will help doctors guide treatment therapies by focusing on those selected biomarkers. Therefore, LGBMRFE is presented and is discussed in the following section.

6.1.3.3 LGBMRFE

LGBMRFE (Light Gradient Boosting Machine with Recursive Feature Elimination) is a wrapper technique employed to identify the markers with high significance. The description of LGBMRGE is given in Section 3.2.3.1. RFE works by successfully eliminating features from the training dataset until the target number remains. LGBM calculates the gradient score for each feature and selects

the top features, which are then passed to RFE. RFE selects the best features and ranks them according to their importance. The reason behind choosing this method is that the computing power of LGBM is quite good, which shows that it can efficiently work well with large datasets [154]. Moreover, the LBGMRFE supports parallel processing means it trains the trees parallelly thus reducing the computational time. LGBMRFE returns a total of 30 miRNAs, 50 DEGs, and 65 DMs, respectively. The top 5 features have been selected from each datatype as the identified markers. The complete description of the dataset before and after preprocessing and feature extraction is shown in Table 6.1. Further, the combined top 5 markers from all the types are validated using DAVID functional analysis which is discussed in the following section.

Table 6.1: Description of Dataset

Data	No. of samples	No. of features	Features after Pre-processing	Features after extraction		
miRNA	1207	1372	999	-	373	30
mRNA	1222	56602	54825	-	15000	50
DM	895	485577	364051	30000	10000	65

6.1.4 Biological Interpretation

To validate the identified markers, i.e., the top 5 markers from the integrated multi-omics data, DAVID functional analysis has been performed. The complete description of DAVID functional analysis is given in Section 3.2.4. This proves the validity of identified markers and shows their significance in the pathway analysis, cell adhesion, cell growth, and signal transduction respectively. Further, identified markers along with extracted features are passed to the modeling phase for disease prediction which is detailed in the following section.

6.1.5 Modeling

A Stacked ensemble of Gradient Boosting Machine (GBM), Random Forest (RF), Naive Bayes (NB), and Deep Neural Network (DNN) is developed for disease prediction based on the identified multi-omics features/ biomarkers. Stacking works by using multiple heterogeneous weak models at first-level training to train only

a portion of the problem but not the whole problem. The training of the weak learner is done in parallel [274]. Hence, different base learners have been built, which can be used to make first-level or intermediate predictions [275]. Afterward, a new model called meta-model or meta learner is added, which will make predictions on the same class variable by considering the intermediate predictions as features. Stacking is also known as Stacked Generalization. It is called stacking because the meta-model is trained on the top of the base models just like the stack. The structure of stacking is shown in Figure 6.2.

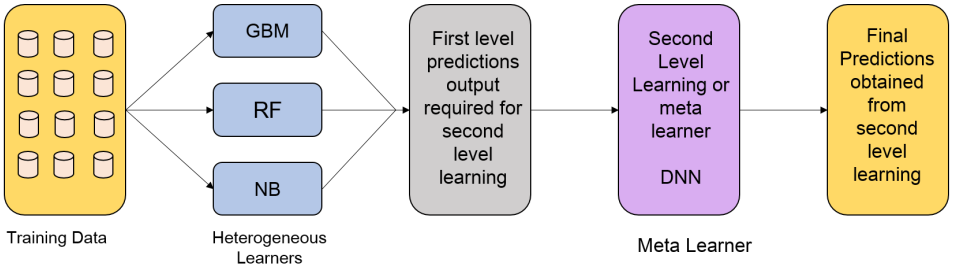


Figure 6.2: Structure of Stacking [4]

Suppose there are M heterogeneous weak learners and one-second-level learners, i.e., meta-learners. The training is performed on the training set by applying k-fold cross-validation using M learners, and predictions are made as p_1, p_2, ..., and p_m at the first level. The predictions from the first level are treated as features, i.e., N features are formed, and then a new dataset $[]_{N \times M}$ and $[]_y$ is made using these N features and y target variable for second-level training. Where $N \times M$ is a feature matrix. This new dataset is trained using a meta-learner, and final predictions are made at second-level training [276].

The extracted gene expression, DM, and miRNA features along with the identified markers are integrated using concatenation-based integration. Concatenation-based integration is discussed in Section 1.2.1.1. Three heterogeneous based learners comprising NB, RF, and GBM are used as base learners to make intermediate predictions, i.e., first-level predictions. The selected algorithms NB, RF, and GBM, each embody distinct methodologies rooted in specific principles. NB operates on Bayes' theorem within a probabilistic framework. RF employs an ensemble approach utilizing decision trees. GBM integrates boosting with ensemble learning techniques. The inherent diversity within the data facilitates the inclusion of numerous features, thereby enhancing the generalization capability of the models [277]. RF and GBM demonstrate resilience towards noisy data and excel at handling outliers. This is accomplished through their respective mechanisms: RF aggregates predictions from multiple trees, while GBM prioritizes the

most informative features. Furthermore, NB exhibits robust performance in high-dimensional datasets characterized by noisy features, owing to its probabilistic nature and independence assumptions. Despite experimentation with alternative methods such as SVM, XGBoost, and DT, satisfactory results were not obtained. Therefore, NB, RF, and GBM were selected as the base models for this study. The performance is evaluated using these base models. Then, these intermediate predictions have been considered as features that are then passed to a meta learner, i.e., DNN for second-level predictions. The complete description of the learning models is given in Section 3.2.5. The experiment is validated using 10-fold cross-validation meaning one fold is used for testing and the remaining 9 folds for training. The average is computed for each fold which is used as a final performance. The algorithm of the HBS-STACK is given in the Algorithm 6.1.

6.2 Experimental Setup and results

6.2.1 Experimental Data

The integrated dataset, including miRNA, gene expression, and DM, is divided into 70:30 ratios, meaning 70% is used for training and 30% for testing. There are a total of 842 patients in BRCA of which 588 samples are used for training, and 254 samples are used for testing the model. The evaluation of model performance solely on training data can lead to potential misinterpretation. In order to effectively assess a model's performance and to remove the overfitting, it is imperative to employ methodologies such as cross-validation as discussed in Section 3.2.5.8. This technique involves partitioning the dataset into various subsets, enabling the model to be trained and tested on distinct portions of the data. This approach offers a more precise estimation of the model's performance on data that has not been previously observed.

6.2.2 Experimental Setup

The minimum hardware requirement required to implement the work is 8 GB RAM with an i5 processor. RStudio 3.5.1 is used for training and testing, Biomart converts the gene IDs and probe IDs to gene symbols, and Python 3.9.0 is used to implement LGBMRFE because of the large size of the dataset. The description of tools used is given in Section 3.1. Moving ahead, the packages used to implement the research are TCGAbiolinks, dplyr, SummarizedExperiment, limma, and H2o. TCGABiolinks package is used to retrieve, download, prepare, analyze, and

Algorithm 6.1 Pseudocode of HBS-STACK

Input: Training Dataset D $\{D = \{X_i, Y_i\}\}$, i=1 to m

Output: Top Biomarkers N, Performance Parameters

Begin:

Step 1- Preprocess Dataset
for each sample S in D **do**
    D <− na.omit(D) ▷ omit 20% missing values
    D <− KNNimpute(D) ▷ impute remaining missing values
end for
Step 2- Extract Features
for each Feature F_i in DM **do**
    F_i <− avereps(F_i)
end for
for each Feature F_i in MI, G, DM **do**
    if $(|log_2(FC) > 0.5 | and FDR < 0.05)$ **then**
        Select F_i Else discard feature
    end if
end for
Set R={}
for selected features, F_i in D **do**
    Train LGBMRFE on each F_i Calculate gradient score
    Select F_i with large gradients
end for
Select the top 5 genes for each D
Integrate Selected Features for each D
Step 3- Learning ▷ Divide dataset into Training and Testing
Impose k-fold Cross Validation
for k=1 to K **do**
    for for n=1 to N **do** ▷ N is no. of learners
        Train base Learners h_{kn} ▷ h is a base learner
    end for
    Create new training set from base predictions
    Train DNN on new dataset ▷ DNN is a meta learner
    Test the performance on the testing dataset
    Compute Performance Parameters
end for

End

visualize the TCGA dataset by accessing the Genomics Data Common Data (GDC) portal. SummarizedExperiment is used to access the assay values represented in the form of matrix-like objects. The limma package is used to normalize the data using the normalizeBetweenArrays function. H2o is an open-source platform used to implement multiple supervised, and unsupervised ML algorithms run in parallel.

6.2.3 Experimental Steps

The HBS-STACK based biomarker selection for disease prediction discussed in this Chapter is designed using the following steps.

- The publicly available multi-omics dataset comprising gene expression, miRNA, and DM of BRCA patients is downloaded from the TCGA portal. The missing values have been removed and imputed and normalization is performed to maintain consistency.
- A hierarchical feature/ biomarker selection method is developed which consists of aggregate information between CPG sites and genes, statistical tests, and LGBMRFE which returns the top features.
- The extracted top 5 features/ biomarkers are validated using DAVID analysis.
- Stacking of four learners comprising NB, RF, GBM, and DNN is proposed on the integrated dataset for disease prediction. The performance of the proposed HBS-STACK approach is evaluated using 6 parameters comprising accuracy, sensitivity, specificity, precision, the area under curve (AUC), and Mathews correlation coefficient (MCC) which are discussed in Section 3.2.6.

6.2.4 Results

The LGBMRFE selects the features and assigns a rank to the selected features. The top 15 markers from each type considering miRNA, gene expr. and DM are selected as markers and are discussed in the following subsection.

6.2.4.1 Identified Markers

The feature importance plot for identified markers has been plotted and shown in Figure 6.3. Out of these 15 markers combined top 5 markers comprising IQSEC1, LHX5-AS1, ADAMTS5, VEGFD, and LIMA1 have been identified as diagnostic

biomarkers for BRCA. The identified genes are validated using functional analysis

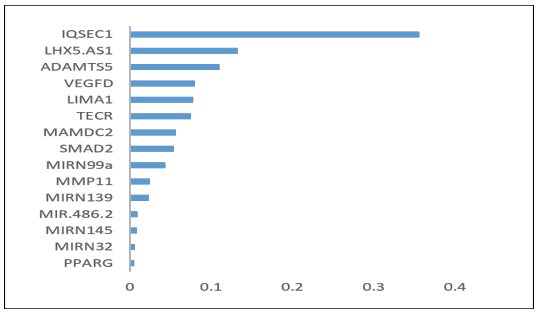


Figure 6.3: Importance plot for top 15 BRCA features [4]

and Kyoto Encyclopedia of Genes and Genomes (KEGG) Pathway analysis with the help of David functional analysis tool. From the analysis, it has been found that IQSEQ1 and VEGFD are involved in endocytosis and MAPK signaling pathways respectively. LHX5-AS1 and LIMA1 play an important role in cell division, immune response, and signal transduction. VEGFD is involved in biological processes and molecular functions. From the literature, Ghazala et al. [278] identified the ADAMTS5 marker, which is under-expressed in invasive ductal breast carcinoma patients. Stephanie et al. [279] placed the VEGFD marker, which shows a higher impact on the growth of tumors in breast cancer patients. Lucas et al. [280] identified LHX5-AS1 as a diagnostic biomarker found in pan-cancer, lung cancer, and KIRC patients. Similarly, LIMA1 has been recognized by Yan et al. [281] as the tumor progressor marker of breast cancer. IQSEC1 marker is discovered for the first time in breast cancer.

Furthermore, along with the identified markers the selected features are integrated and passed to the Stacked ensemble for disease prediction. The proposed approach is applied to the TCGA BRCA dataset by considering 6 performance parameters: accuracy, sensitivity, specificity, precision, MCC, and AUC. The experiments have been done in single omics data and integrated omics to show the variability in results obtained by applying individual heterogeneous models comprising GBM, RF, and NB and stacked ensemble with DNN as a meta learner which are discussed in the following section.

6.2.4.2 Comparison of results in single omics

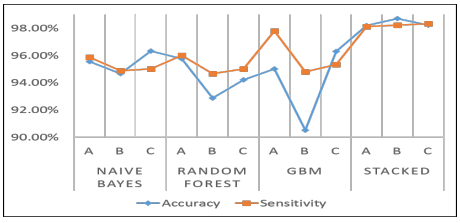
Table 6.2 shows the result of the HBS-STACK on a single omics dataset. The individual heterogeneous models have been applied to each type of omics data, i.e., miRNA, Gene Expression, and DNA data, and compared the results with the proposed stacked ensemble with DNN as a meta learner. It is visible that the

proposed stacked method works well with an accuracy of 98.20%, 98.70%, and 98.21% for miRNA, Gene Expression, and DM data, respectively. Compared with each model, i.e., NB, GBM, and RF, the Stacked ensemble shows an improvement of 3% for each type. Further, other parameters, including sensitivity, specificity, precision, AUC, and MCC have been calculated, which shows that the stacked model outperformed for each type of omics data.

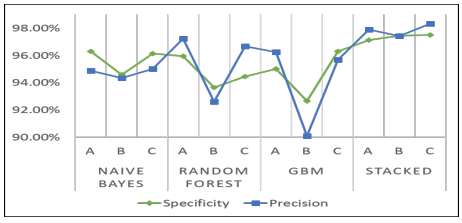
Table 6.2: Model Performance in Single Omics

Model	Data	Accuracy	Sensitivity	Specificity	Precision	AUC	MCC
NB	miRNA	95.54%	95.87%	96.29%	94.87%	97.68%	92.02%
	Gene Expr.	94.65%	94.88%	94.59%	94.33%	96.92%	94.07%
	DM	96.33%	95%	96.12%	95%	97.76%	94.15%
RF	miRNA	95.76%	96%	95.95%	97.22%	97.45%	95.06%
	Gene Expr.	92.87%	94.66%	93.65%	92.59%	96.89%	94.51%
	DM	94.21%	95%	94.45%	96.66%	95.10%	91.82%
GBM	miRNA	95%	97.79%	95%	96.25%	97.95%	93.90%
	Gene Expr.	90.50%	94.79%	92.65%	90.09%	95.89%	92.51%
	DM	96.30%	95.33%	96.29%	95.69%	97.43%	91.45%
Stacked	miRNA	98.20%	98.12%	97.12%	97.90%	98.60%	96%
	Gene Expr.	98.70%	98.22%	97.45%	97.43%	98.12%	95.51%
	DM	98.21%	98.32%	97.50%	98.33%	98.64%	91.88%

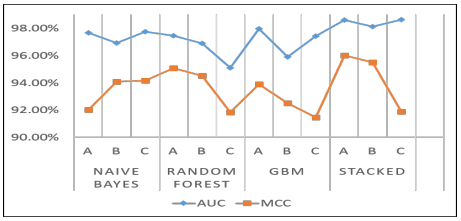
Moreover, to show the variation in performance parameters in single omics, line plots have been plotted and shown in Figure 6.4. The line plots depict that the stacked ensemble performed efficiently for each parameter.



(i) Accuracy, Sensitivity



(ii) Specificity, Precision



(iii) AUC, MCC

Figure 6.4: Line plots for performance parameters [4]

6.2.4.3 Comparison of Proposed Work in Integrated Omics

The HBS-STACK is applied to the integrated dataset and the performance of the individual model and the stacked ensemble is evaluated. Table 6.3 shows the performance parameters obtained using heterogeneous and stacked ensemble models. It is shown from the results that the stacked ensemble worked efficiently with an accuracy of 99.60%. Furthermore, it shows an improvement of 2.19%, 3.35%, and 2.37% compared with NB, RF, and GB. The HBS-STACK has a high sensitivity of 99.98%, a high specificity of 99.95%, and a high AUC of 99.88%, making it an effective breast cancer prediction model. The presented model's high accuracy for breast cancer prediction would aid physicians in making more accurate decisions, which will help give the patient the most suitable therapy. Further, bar plots have been plotted to depict the results in Figure 6.5 visually.

Table 6.3: Results obtained using Integrated Omics

Parameter	**NB**	**RF**	**GBM**	**Stacked Ensemble**
Accuracy	97.41%	96.25%	97.23%	99.60%
Sensitivity	96.99%	97.69%	97.78%	99.98%
Specificity	97.59%	97.79%	97.59%	99.95%
Precision	98.68%	98.11%	97.10%	99.99%
AUC	98.75%	98.81%	98.57%	99.88%
MCC	95.85%	96.90%	95.51%	97.98%

The higher bars show that the stacked ensemble outperformed by approximately 3%, 3%, 3%, 2%, 2%, and 2%, in accuracy, sensitivity, specificity, precision, AUC, and MCC, respectively.

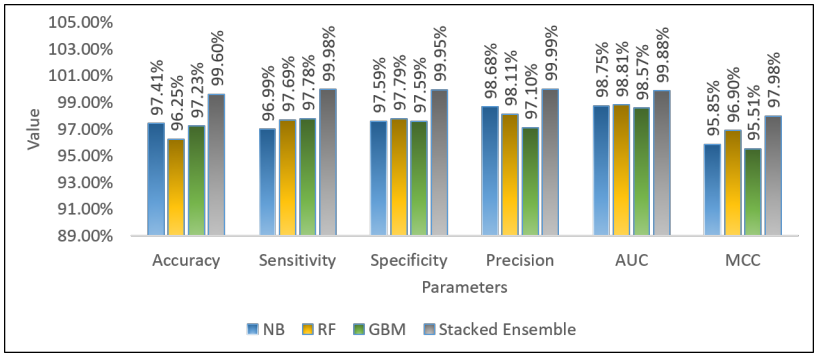
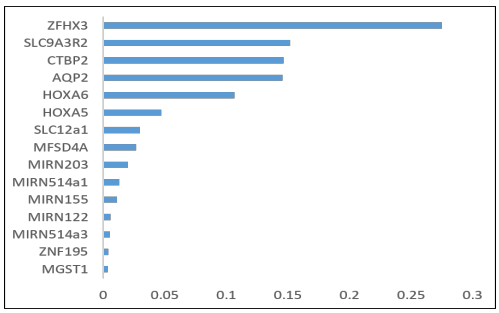


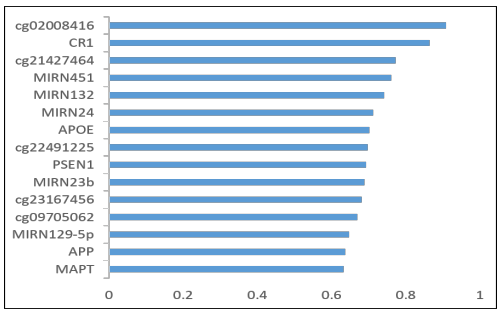
Figure 6.5: Bar plots for performance parameters[4]

6.2.4.4 Validation on KIRC and Alzheimer Disease

The validation of HBS-STACK is done using TCGA KIRC and Alzheimer's Disease dataset. TCGA KIRC dataset is also downloaded from the GDC portal, in which a total of 346 samples are there, with 322 tumor patients and 24 non-tumor patients. Similarly for Alzheimers, the ROSMAP [76] dataset has been used, which contains Alzheimer patients and non-Alzheimer patients, respectively. The description of ROSMAP is given in Section 3.2.1.1. There are 351 samples in Alzheimer's disease (AD) data, with 169 non-AD and 182 AD patients. The total number of features for each type of omics data comprising mRNA, miRNA, and DM is 200. The dataset is preprocessed, and feature extraction is performed. The top 15 features for miRNA, gene expr, and DM are identified for KIRC and Alzheimer's disease whose importance plots are shown in the Figure 6.6.



(i) KIRC



(ii) Alzheimer

Figure 6.6: Top 15 features for KIRC and Alzheimer [4]

Similarly, the combined top 5 markers identified are ZFHX3, SLC9A3R2, CTBP2, AQP2, and HOXA6 and cg02008416, CR1, cg21427464, MIRN451, MIRN-132, for KIRC and Alzheimer, respectively. From David Analysis it has been found that ZFHX, CTBP2, AQP2, CR1, and cg21427464 are involved in signaling path-

ways, and pathways in cancer resulting in cell growth. HOAX6 is involved in biological processes. SLC9A3R2 and TMEM61 play an important role in molecular function resulting in signal transduction. MIRN451 and MIRN132 are involved in cellular processes and gene regulations. From the literature, Hehuan et al. [282] identified AQP2 markers as down-regulated and marked as prognostic markers. Xiaofang et al. [283] identified HOXA6 as a prognostic markers, and it has been identified as low expressed than normal samples. Three new markers, including ZFHX3, CTBP2, and SLC9A3R2, have been placed in KIRC in which ZFH3 and CTBP2 were identified as prognostic markers in glioblastoma patients [284] and hepatocellular carcinoma patients [285]. SLC9A3R2 is recognized as a good prognosis marker [286]. CR1 is identified by Ali et al. [287] as a diagnostic marker for the early detection of Alzheimer's disease. Aparna et al. [288] identified cg21427464 (BIN1) as a diagnostic biomarker for early identification of Alzheimer's disease. The role of hsa-miR-451, and has-miR-132, described by Gustavo et al. [289] and it has been found that the has-miR-451, has-miR-132 are highly upregulated genes in Alzheimer disease. cg02008416 (TMEM61) has been identified for the first time in Alzheimer's patients.

Additionally, the extracted features from miRNA, gene expr. and DM along with identified markers for KIRC and Alzheimer are integrated and passed to HBS-STACK for disease prediction. The performance is evaluated using 5 parameters comprising accuracy, sensitivity, specificity, AUC, and MCC and is shown in Table 6.4. From the results, it has been found that the stacked ensemble performed well for KIRC and Alzheimer with an accuracy of 99.03% and 92.05%, respectively.

6.2.4.5 Comparison of HBS-STACK with Existing Work

The comparative analysis of HBS-STACK with existing works is performed to prove the superiority of the proposed HBS-STACK. Table 6.5 shows the result of the proposed work with the existing work. We compared the results of the proposed work with different authors using single omics and integrated omics. The proposed work is compared with Srinivasulu et al. [290] using miRNA data. It shows a higher accuracy of 98.99%, higher sensitivity of 98.12%, and higher specificity of 97.12%, with an improvement of 18%, 19%, and 17%, respectively. Moving ahead, we compare the results with Wang et al.[38] using gene expression data, and it shows an enhancement of 10%, 3%, and 30% in accuracy, sensitivity, and specificity, respectively. Similarly, the proposed iMVAN outperformed Raweh et al. [291], Al-beity et al. [292] and Joung et al. [293] on DM and Integrated Data with an accuracy of 98.91% for DM and 99.6% for integrated omics corre-

Table 6.4: Results of HBS-STACK on integrated KIRC and Alzheimer features

Disease	Data type	Accuracy	Sensitivity	Specificity	AUC	MCC
KIRC	NB	96.32%	97.24%	96.64%	96.35%	97.48%
	RF	95.55%	96.87%	96.99%	96.32%	95.37%
	GBM	97%	97.86%	98.21%	97.56%	96.12%
	Stacked	99.03%	99.85%	99.99%	98.75%	98.12%
Alzheimer	NB	88.32%	87.61%	86.45%	87.23%	86.54%
	RF	87.39%	86.89%	85.97%	86.23%	86.64%
	GBM	86.12%	85.59%	87.21%	86.78%	86.11%
	Stacked	92.05%	89%	88.14%	88.95%	87.45%

spondingly. The results on KIRC and Alzheimer patients are also compared with the result of Baoshan et al. [294] and Wang et al. [76] respectively. It is found that the proposed work works well with an accuracy, sensitivity, and specificity value of 99.03%, 99.85%, and 99.99%, and 92.05%, 89%, and 88.14% for KIRC and Alzheimer respectively.

Moreover, to compare the proposed work with the existing work, a bar plot has been plotted for miRNA, Gene Expression, DM, Integrated data for BRCA, and integrated data for KIRC and Alzheimer patients, as shown in Figure 6.7.

6.2.4.6 Statistical Analysis to validate significance

The HBS-STACK is validated using statistical tests to prove its significance. Two statistical tests comprising the Wilcoxon signed rank test for making simple pairwise comparisons and the Friedman test for making multiple comparisons. The hypothesis of the Wilcoxon signed rank test is that the two prediction accuracies are equivalent. The fundamental assumption of the Friedman test is that there is no significant dissimilarity in the predictive abilities of the various models under consideration [245]. The statistical analysis is performed for TCGA BRCA, KIRC, and Alzheimer datasets in which the performance of the proposed HBS-STACK is compared with individual GBM, RF, and NB models. The hypothesis is that if the p-value is less than 0.05, then the test is considered to be significant.

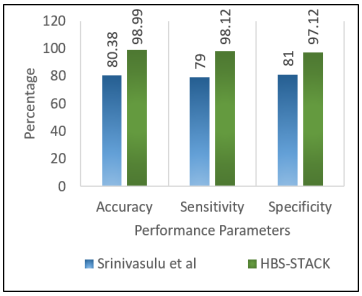
Table 6.5: Comparison of HBS-STACK with Existing Works

Disease	Dataset	Methods	Accuracy	Sensitivity	Specificity
BRCA	miRNA	Srinivasulu et al. [290]	80.38%	79%	81%
		HBS-STACK	98.99%	98.12%	97.12%
	Gene Expr.	Wang et al.[76]	89%	95%	68%
		HBS-STACK	99%	98.22%	97%
	DM	Raweh et al [291]	98.33%	97%	95%
		HBS-STACK	98.91%	98.32%	97.21%
	miRNA, gene expr, DM	Al-Baity et al. [292]	97.33%	96.82%	97.98%
		Joung et al. [293]	89.1%	93.56%	92.49%
		HBS-STACK	99.6%	99.98%	99.99%
KIRC	miRNA, gene expr, DM	Baoshan et al. [294]	73%	80%	78%
		HBS-STACK	99.03%	99.85%	99.99%
Alzheimer	miRNA, gene expr, DM	Wang et al. [76]	82%	81.12%	82.44%
		HBS-STACK	92.05%	89%	88.14%

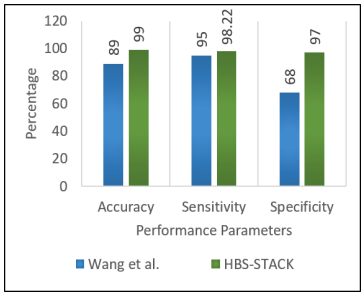
Table 6.6: Significance Analysis of proposed HBS-STACK

Type	Model	Wilcoxon Test	Friedman Test
BRCA	HBS-STACK vs NB	0.0041	H_0 = e1 = e2 = e3 F=48.62 p-value=3.7e-10 (Reject H_0)
	HBS-STACK vs GBM	0.0067	
	HBS-STACK vs RF	0.0034	
KIRC	HBS-STACK vs NB	0.0041	H_0 = e1 = e2 = e3 F=34.23 p-value=1.9e-8 (Reject H_0)
	HBS-STACK vs GBM	0.077	
	HBS-STACK vs RF	0.0034	
Alzheimer	HBS-STACK vs NB	0.021	H_0 = e1 = e2 = e3 F=39.18 p-value=2.5 e-10 (Reject H_0)
	HBS-STACK vs GBM	0.015	
	HBS-STACK vs RF	0.0028	

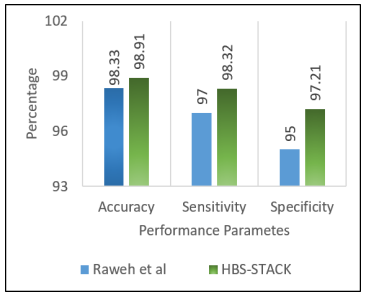
The Wilcoxon and Friedman test results for TCGA BRCA, TCGA KIRC, and Alzheimer's are given in Table 6.6. From the results, it has been found that the HBS-STACK is significant in the case of BRCA and Alzheimer's. On the other side, in KIRC, one case is there in HBS-STACK vs. GBM where the p-value is >



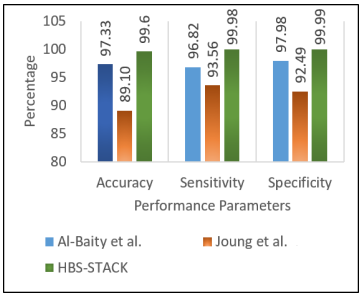
(i) miRNA



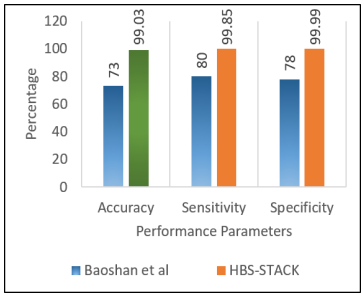
(ii) Gene Expression



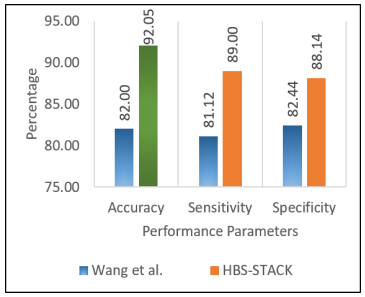
(iii) DM



(iv) Integrated (BRCA)



(v) Integrated (KIRC)



(vi) Integrated (Alzheimer)

Figure 6.7: Bar plots for comparison of HBS-STACK with existing works [4]

0.05. In that case, the value is considered to be non-significant and is violating the test condition. In the Friedman test, HBS-STACK performed best with a p-value < 0.05, thus showing the effectiveness in predicting the disease outcome.

6.3 Discussion of Results

In the current research, biomarkers have been identified with hierarchical biomarker selection comprising aggregate information between CpG and genes, Fold change, FDR value, and LGBMRFE. Results are evaluated using a proposed stacked ensemble on a multi-omics dataset from TCGA. Figures 6.3 and 6.6 show the identified markers in which the markers with their importance in disease are shown. The plots indicate that IQSEC1 in BRCA, ZFHX3 in KIRC, and TMEM61 in Alzheimer's are highly important in diagnosing disease. The three-stage feature selection is proposed due to the high dimensionality of the multi-omics dataset. The main contribution of this research is the removal of redundant genes present in the DM dataset. Multiple CpG probes are associated with one common gene. To solve this, the average of each gene is calculated which is then compared with some threshold value. Additionally, the extracted features are passed to statistical tests and LGBMRFE. The second contribution is to reduce the time while selecting the

features. Traditional ML algorithms with RFE take a high computational time while processing the features, LGBMRFE, on the other side, solves this problem. LGMMRFE works on gradient boosting and supports parallel processing, which means it trains the trees parallel, thus reducing the computational time. Further, the identified markers are integrated, and stacking is performed, shown in Tables 6.2, 6.3, 6.4, 6.5, and 6.6. It is visible from the results that the stacked ensemble with DNN as a meta-learner gives the best results as compared to individual GBM, RF, and NB. This is because the heterogeneous base models provide the result in probabilities instead of actual class labels, which are further combined to make a new dataset that provides more context to a meta-learner. RF is an ensemble method that works by building multiple decision trees and is used to solve the problem of overfitting. However, it will lead to poor performance if the dataset used in each decision tree is unbalanced. Also, it is computationally slow as it makes multiple decision trees to make predictions [295]. GBM is a boosting algorithm that works by combining various decision trees, and the objective is to minimize the loss function. It gives accurate results compared to RF because it tries to reduce the error from each decision tree. Still, it is computationally high on large datasets and sometimes leads to overfitting [170]. NB is fast, easy to implement, and gives good results, but it can cause zero probability problems if test data for some label is not present in the training dataset [168]. DNNs can learn from the data and make their own decisions like the human brain. They also offer parallel processing, making it faster for larger datasets [171]. Therefore, DNN is used as a meta-learner in the stacked model, which performed better with an accuracy of 99.60%. Moreover, to reduce the overfitting problem, cross-validation is used which works by dividing data into train and test sets. Every fold offers a distinct viewpoint into the potential performance of the model on data that has not been previously observed. Additionally, the validation of the proposed HBS-STACK is done on the Alzheimer dataset, showing that the proposed HBS-STACK performed well not only on TCGA BRCA and TCGA KIRC dataset but on other diseases also. Moving ahead, HBS-STACK is compared with Al-Baity et al. [292], Baoshan et al. [294], and Wang et al. [76], and it shows an improvement of 2.27%, 26.03% and 10.05% in accuracy for BRCA, KIRC and Alzheimer, respectively, as shown in Figure 6.7.

6.4 Conclusion

This chapter discussed an HBS-STACK based biomarker selection for disease prediction in a multi-omics dataset. A hierarchical biomarker selection approach is

presented which includes three feature/ biomarker selection techniques including aggregate information between CPG sites and genes, statistical tests (FC and FDR), and LGBMRFE. A three-phase is provided because of the high dimensionality of the multi-omics dataset. Further, the biological interpretation is performed to validate the identified markers. The extracted features and biomarkers are integrated and passed to a proposed stacked ensemble approach for disease prediction. The proposed HBS-STACK is applied to the TCGA BRCA multi-omics dataset and validated on the TCGA KIRC and Alzheimer's disease dataset. The ability of the proposed HBS-STACK is to identify the most important biomarkers for disease prediction that will help clinicians guide treatment therapies by focusing only on the identified markers.

The next chapter will delve into the conclusions drawn from this research work and discuss potential future directions for biomarker identification in multi-omics datasets for accurate diagnosis and prognosis.

Chapter 7
Conclusions and Future Work

This chapter summarizes the thesis by providing the conclusions of the research work done and directions for future work.

This work provides an in-depth exploration of computational technologies such as machine learning and deep learning for biomarker identification in multi-omics data for disease diagnosis and prognosis. The literature review highlights the necessity of a framework for biomarker identification in multi-omics for survival prediction, disease subtype classification, and disease prediction. It emphasizes the need for an exploration of data preparation (data pre-processing, feature/ biomarker selection), validation of identified markers, and learning model options for healthcare researchers. To bridge this gap, three frameworks comprising BioSurv, iMVAN, and HBS-STACK are proposed for biomarker identification in multi-omics data required for survival prediction, disease subtype classification, and disease prediction. BioSurv is proposed for biomarker identification in multi-omics for survival prediction in breast cancer and lung cancer patients and is validated using the METABRIC dataset. iMVAN is presented for biomarker identification in multi-omics datasets for disease subtype classification of breast cancer and is validated using Pan-kidney and cervical cancer datasets. For biomarker selection in multi-omics required for disease prediction, HBS-STACK is presented for breast cancer patients and is validated on kidney cancer and Alzheimer's patients.

The conclusions drawn from this work are presented in Section 7.1, summarizing the key findings and contributions. Additionally, Section 7.2 suggests potential avenues for future research to further enhance the field of biomarker identification in multi-omics data and advance the capabilities of the proposed framework.

7.1 Conclusion

The thesis makes use of various computational technologies such as machine learning (ML) and deep learning (DL) to develop a framework for biomarker identification in multi-omics data for disease diagnosis and prognosis which will contribute

in the existing research towards the improvement of patient care. The conclusions of this research work are described as below:

- The comprehensive review of computationally intelligent approaches is conducted to understand the current research of biomarker identification in multi-omics data. This thorough examination involved investigating, comparing, and categorizing various technologies and tools utilized for biomarker identification in single and multi-omics for disease prediction, survival prediction, disease subtype classification, and treatment/ response. Through this critical analysis, it became evident that there is a need to develop an effective framework specifically tailored for biomarker identification in multi-omics for disease prediction, survival prediction, and disease subtype classification.

- The research methodology is presented for the biomarker identification which is followed by the developed framework for disease prediction, survival prediction, and disease subtype classification. The methodology provides phases of data acquisition, data preparation, feature/ biomarker identification, validation of identified markers, modeling, and performance evaluation. By using the presented methodology, three frameworks comprising BioSurv, iMVAN, and HBS-STACK are developed.

- The BioSurv framework is proposed which uses ML and DL techniques for biomarker identification in multi-omics for survival prediction. The BioSurv is designed using the phases discussed in the research methodology by implementing Random Spatial Local Best Cat Swarm Optimization (RSLBCSO), statistical methods, and Bayesian optimized deep neural network (DNN). To validate the effectiveness of the BioSurv framework, its performance is assessed using multiple cancers (Breast, Lung) and datasets (TCGA, METABRIC). The ability of the BioSurv framework to accurately identify biomarkers and predict patients as either short-term or long-term survivors can significantly assist clinicians in suggesting appropriate treatment recommendations for individual patients.

- The iMVAN framework is developed by utilizing the DL technique for biomarker identification in multi-omics for disease subtype classification. The iMVAN is designed using the phases discussed in the research methodology by incorporating multi-modal variational autoencoder (MVAE), similarity network fusion (SNF), and simplified graph convolutional network (SGC) for biomarker identification in breast cancer. The iMVAN framework is vali-

dated using Pan kidney and cervical cancer. The iMVAN framework provides effective and accurate identification of biomarkers in multi-omics by classifying the disease into its subtypes, thereby contributing to the enhancement and personalization of treatment for affected individuals.

- The HBS-STACK is developed by utilizing the ML and DL techniques for biomarker identification in multi-omics required for disease prediction. The iMVAN is developed using the phases described in the research methodology by employing hierarchical biomarker selection and a stacked ensemble model. The performance of HBS-STACK is validated using multiple cancers (breast, kidney) and diseases (Alzheimer's). The ability of the HBS-STACK is to identify the most important biomarkers for disease prediction that will help clinicians guide treatment therapies by focusing only on the identified markers.

The summarized results of the proposed approaches, including, BioSurv, iMVAN, and HBS-STACK, are given in Table 7.1.

7.2 Future Scope

This section provides some possible future directions related to this research work.

- The work is designed for biomarker identification in multi-omics for survival prediction, subtype classification, and disease prediction. With the availability of datasets containing information about disease recurrence and treatment/ response, it could be used to identify the biomarkers for a disease recurring after treatment.

- The present work primarily emphasizes biomarker identification for survival prediction in the context of cancer. However, the designed BioSurv for biomarker identification in multi-omics required for survival prediction holds the potential for extension to other diseases like Alzheimer's. Exploring its applicability and performance in identifying the biomarkers for predicting survival outcomes for different diseases could contribute to personalized treatment approaches across various diseases.

- The proposed BioSurv, iMVAN, and HBS-STACK framework can be used for drug discovery and development. The predictive models can be trained to find molecules with therapeutic potential, improve drug design, and predict treatment effectiveness by utilizing multi-omics data.

Table 7.1: Summarized Result of Proposed Approaches

Approach	**Disease**	**Accuracy**	**Identified Markers**
BioSurv	BRCA	91%	FGFR3, YWHAG, NFKB2, RAB2A, CHEK1, ATG5, miR-106b, miR-132, miR-222, miR-143, miR-98, STK11, ROCK1, IL13, SMC3, TSC2, ARNT2, AXIN1, DLL1, LAMA5, PLCG2
	LUAD	90%	FN1, ITGA3, PRKAA2, SGK2, CASP8, FZD3, RHOA, TGFB1, RRAS, miR-132, miR-155, miR-221, miR-222, BPIFB1, MAP2K4, NLRP1, CYCS, TYK2, AP2A2, NEU1, SYNJ2, GIT1, TBCD
iMVAN	BRCA	87%	GSTM1, AGT, CDH1, RET, CALML5, ERBB2, PTEN, ESR1, CCND1, FZD6, LRP5, FZFR1, BRAF, BCL2, AR, ETS1, MSH6
	KIPAN	92%	FGB, C6, FGG, PLG, EFNA5, EFNA5, ROBO1, SEMA3E, BRAF, MAPK9, CDKN1A, BAK1, EGFR
	CESC	86%	MMP13, CXCL6, CXCL5, MMP1 BIRC2, BIRC3, MAPK10, RAD51, MYC, DVL3, ETS1, VHL
HBS-STACK	BRCA	99.60%	IQSEC1, LHX5-AS1, ADAMTS5, VEGFD, LIMA1
	KIRC	99.03%	ZFHX3, SLC9A3R2, CTBP2, AQP2, HOXA6
	Alzheimer	92.05%	cg02008416, CR1, cg21427464, MIRN451, MIRN-132

- The current work focused on biomarker identification in multi-omics data including genomic, transcriptomic, and proteomics. The multi-omics data can be integrated with imaging data, electronic health records, and environmental factors to obtain an in-depth understanding of health conditions. This can lead to better biomarker identification for disease diagnosis, prognosis, and treatment suggestions.
- The transfer learning technique can be used to enhance the identification of biomarkers in multi-omics dataset with less samples using the knowledge learned from multi-omics dataset in this work.

References

[1] Arwinder Dhillon, Ashima Singh, and Vinod Kumar Bhalla. A systematic review on biomarker identification for cancer diagnosis and prognosis in multi-omics: from computational needs to machine learning and deep learning. *Archives of Computational Methods in Engineering*, 30(2):917–949, 2023.

[2] Arwinder Dhillon, Ashima Singh, and Vinod Kumar Bhalla. Biomarker identification and cancer survival prediction using random spatial local best cat swarm and bayesian optimized dnn. *Applied Soft Computing*, 146:110649, 2023.

[3] Arwinder Dhillon, Ashima Singh, and Vinod Kumar Bhalla. imvan: integrative multimodal variational autoencoder and network fusion for biomarker identification and cancer subtype classification. *Applied Intelligence*, 53(22):26672–26689, 2023.

[4] Arwinder Dhillon, Ashima Singh, and Vinod Kumar Bhalla. Hbs–stack: hierarchical biomarker selection and stacked ensemble model for biomarker identification and cancer prediction on multi-omics. *Neural Computing and Applications*, pages 1–19, 2024.

[5] Caris Life Sciences. What are biomarkers? `https://www.mycancer.com/resources/what-are-biomarkers/`, Jan 2021.

[6] Francis S Collins and Harold Varmus. A new initiative on precision medicine. *New England journal of medicine*, 372(9):793–795, 2015.

[7] Ruth Clifford, Tania Louis, Pauline Robbe, Sam Ackroyd, Adam Burns, Adele T Timbs, Glen Wright Colopy, Helene Dreau, Francois Sigaux, Jean Gabriel Judde, et al. Samhd1 is mutated recurrently in chronic lymphocytic leukemia and is involved in response to dna damage. *Blood, The Journal of the American Society of Hematology*, 123(7):1021–1031, 2014.

[8] Constanze Schneider, Thomas Oellerich, Hanna-Mari Baldauf, Sarah-Marie Schwarz, Dominique Thomas, Robert Flick, Hanibal Bohnenberger, Lars Kaderali, Lena Stegmann, Anjali Cremer, et al. Samhd1 is a biomarker for cytarabine response and a therapeutic target in acute myeloid leukemia. *Nature medicine*, 23(2):250–255, 2017.

[9] Shuangyan Tan, Qiheng Gou, Wenchen Pu, Chenglin Guo, Yun Yang, Ke Wu, Yaxin Liu, Lunxu Liu, Yu-Quan Wei, and Yong Peng. Circular

rna f-circea produced from eml4-alk fusion gene as a novel liquid biopsy biomarker for non-small cell lung cancer. *Cell research*, 28(6):693–695, 2018.

[10] Biomarkers Definitions Working Group, Arthur J Atkinson Jr, Wayne A Colburn, Victor G DeGruttola, David L DeMets, Gregory J Downing, Daniel F Hoth, John A Oates, Carl C Peck, Robert T Schooley, et al. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clinical pharmacology & therapeutics*, 69(3):89–95, 2001.

[11] Daniel N Cagney, Joohee Sul, Raymond Y Huang, Keith L Ligon, Patrick Y Wen, and Brian M Alexander. The fda nih biomarkers, endpoints, and other tools (best) resource in neuro-oncology. *Neuro-oncology*, 20(9):1162–1172, 2018.

[12] TK Khan. Introduction to alzheimer's disease biomarkers. In *Biomarkers in Alzheimer's Disease*, page 13. Academic New York, NY, USA, 2016.

[13] Konstantinos Sechidis, Konstantinos Papangelou, Paul D Metcalfe, David Svensson, James Weatherall, and Gavin Brown. Distinguishing prognostic and predictive biomarkers: an information theoretic approach. *Bioinformatics*, 34(19):3365–3376, 2018.

[14] Indhupriya Subramanian, Srikant Verma, Shiva Kumar, Abhay Jere, and Krishanpal Anamika. Multi-omics data integration, interpretation, and its application. *Bioinformatics and biology insights*, 14:1177932219899051, 2020.

[15] Laura Bravo-Merodio, John A Williams, Georgios V Gkoutos, and Animesh Acharjee. -omics biomarker identification pipeline for translational medicine. *Journal of translational medicine*, 17(1):1–10, 2019.

[16] Mestrovic T. Proteomics uses. `https://www.news-medical.net/life-sciences/Proteomics-Uses.aspx`, Jan 2020.

[17] Minseung Kim and Ilias Tagkopoulos. Data integration and predictive modeling methods for multi-omics datasets. *Molecular omics*, 14(1):8–25, 2018.

[18] Yehudit Hasin, Marcus Seldin, and Aldons Lusis. Multi-omics approaches to disease. *Genome biology*, 18(1):1–15, 2017.

[19] Miriam M Cortese-Krott, Jerome Santolini, Steve A Wootton, Alan A Jackson, and Martin Feelisch. The reactive species interactome. In *Oxidative Stress*, pages 51–64. Elsevier, 2020.

[20] Ana Conesa and Ali Mortazavi. The common ground of genomics and systems biology. *BMC systems biology*, 8(2):1–10, 2014.

[21] Vessela N Kristensen, Ole Christian Lingjærde, Hege G Russnes, Hans Kristian M Vollan, Arnoldo Frigessi, and Anne-Lise Børresen-Dale. Principles and methods of integrative genomic analyses in cancer. *Nature Reviews*

Cancer, 14(5):299–313, 2014.

[22] U.S. Department of Health and Human Services. Genomic data commons data portal. `https://portal.gdc.cancer.gov/`, Feb 2021.

[23] cBioPortal. Breast cancer (metabric, nature 2012 nat commun 2016). `https://www.cbioportal.org/study/summary?id=brca_metabric`, December 2022.

[24] Richard J Hodes and Neil Buckholtz. Accelerating medicines partnership: Alzheimer's disease (amp-ad) knowledge portal aids alzheimer's drug discovery through open data sharing. *Expert opinion on therapeutic targets*, 20(4):389–391, 2016.

[25] Rui Ding, Shiqiao Zhang, Yawen Chen, Zhiyan Rui, Kang Hua, Yongkang Wu, Xiaoke Li, Xiao Duan, Xuebin Wang, Jia Li, et al. Application of machine learning in optimizing proton exchange membrane fuel cells: a review. *Energy and AI*, 9:100170, 2022.

[26] Prabhaker Mishra, Chandra Mani Pandey, Uttam Singh, Amit Keshri, and Mayilvaganan Sabaretnam. Selection of appropriate statistical methods for data analysis. *Annals of cardiac anaesthesia*, 22(3):297, 2019.

[27] Arwinder Dhillon and Ashima Singh. Machine learning in healthcare data analysis: a survey. *Journal of Biology and Today's World*, 8(6):1–10, 2019.

[28] Poulmanogo Illy, Georges Kaddoum, Paulo Freitas de Araujo-Filho, Kuljeet Kaur, and Sahil Garg. A hybrid multistage dnn-based collaborative idps for high-risk smart factory networks. *IEEE Transactions on Network and Service Management*, 19(4):4273–4283, 2022.

[29] Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T Dudley. Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics*, 19(6):1236–1246, 2018.

[30] Yolande FM Ramos, Sarah J Rice, Shabana Amanda Ali, Chiara Pastrello, Igor Jurisica, Muhammad Farooq Rai, Kelsey H Collins, Annemarie Lang, Tristan Maerz, Jeroen Geurts, et al. Evolution and advancements in genomics and epigenomics in oa research: How far we have come. *Osteoarthritis and Cartilage*, 2024.

[31] Jordi Martorell-Marugán, Siham Tabik, Yassir Benhammou, Coral del Val, Igor Zwir, Francisco Herrera, and Pedro Carmona-Sáez. Deep learning in omics data analysis and precision medicine. *Exon Publications*, pages 37–53, 2019.

[32] Hanieh Azari, Elham Nazari, Reza Mohit, Alireza Asadnia, Mina Maftooh, Mohammadreza Nassiri, Seyed Mahdi Hassanian, Majid Ghayour-

Mobarhan, Soodabeh Shahidsales, Majid Khazaei, et al. Machine learning algorithms reveal potential mirnas biomarkers in gastric cancer. *Scientific Reports*, 13(1):6147, 2023.

[33] Baoshan Ma, Bingjie Chai, Heng Dong, Jishuang Qi, Pengcheng Wang, Tong Xiong, Yi Gong, Di Li, Shuxin Liu, and Fengju Song. Diagnostic classification of cancers using dna methylation of paracancerous tissues. *Scientific Reports*, 12(1):10646, 2022.

[34] Ting Jin, Nam D Nguyen, Flaminia Talos, and Daifeng Wang. Ecmarker: interpretable machine learning model identifies gene expression biomarkers predicting clinical outcomes and reveals molecular mechanisms of human disease in early stages. *Bioinformatics*, 37(8):1115–1124, 2021.

[35] Ying Xie, Wei-Yu Meng, Run-Ze Li, Yu-Wei Wang, Xin Qian, Chang Chan, Zhi-Fang Yu, Xing-Xing Fan, Hu-Dan Pan, Chun Xie, et al. Early lung cancer diagnostic biomarker discovery by machine learning methods. *Translational oncology*, 14(1):100907, 2021.

[36] Fariha Muazzam. Multi-class cancer classification and biomarker identification using deep learning. *bioRxiv*, pages 2020–12, 2020.

[37] Indu Khatri and Manoj K Bhasin. A transcriptomics-based meta-analysis combined with machine learning approach identifies a secretory biomarker panel for diagnosis of pancreatic adenocarcinoma. *medRxiv*, pages 2020–04, 2020.

[38] Xin Zhao, Jian Dou, Jinglin Cao, Yang Wang, Qingjun Gao, Qiang Zeng, Wenpeng Liu, Baowang Liu, Ziqiang Cui, Liang Teng, et al. Uncovering the potential differentially expressed mirnas as diagnostic biomarkers for hepatocellular carcinoma based on machine learning in the cancer genome atlas database. *Oncology reports*, 43(6):1771–1784, 2020.

[39] Oneeb Rehman, Hanqi Zhuang, Ali Muhamed Ali, Ali Ibrahim, and Zhongwei Li. Validation of mirnas as breast cancer biomarkers with a machine learning approach. *Cancers*, 11(3):431, 2019.

[40] Abedalrhman Alkhateeb, Iman Rezaeian, Siva Singireddy, Dora Cavallo-Medved, Lisa A Porter, and Luis Rueda. Transcriptomics signature from next-generation sequencing data reveals new transcriptomic biomarkers related to prostate cancer. *Cancer informatics*, 18:1176935119835522, 2019.

[41] Biao Liu, Yulu Liu, Xingxin Pan, Mengyao Li, Shuang Yang, and Shuai Cheng Li. Dna methylation markers for pan-cancer prediction by deep learning. *Genes*, 10(10):778, 2019.

[42] Reka Toth, Heiko Schiffmann, Claudia Hube-Magg, Franziska Büscheck,

Doris Höflmayer, Sören Weidemann, Patrick Lebok, Christoph Fraune, Sarah Minner, Thorsten Schlomm, et al. Random forest-based modelling to detect biomarkers for prostate cancer progression. *Clinical epigenetics*, 11:1–15, 2019.

[43] Nivedhitha Mahendran and Durai Raj Vincent PM. Deep belief network-based approach for detecting alzheimer's disease using the multi-omics data. *Computational and Structural Biotechnology Journal*, 21:1651–1660, 2023.

[44] Ping Gong, Lei Cheng, Zhiyuan Zhang, Ao Meng, Enshuo Li, Jie Chen, and Longzhen Zhang. Multi-omics integration method based on attention deep learning network for biomedical data classification. *Computer Methods and Programs in Biomedicine*, 231:107377, 2023.

[45] Yanyu Hu, Long Zhao, Zhao Li, Xiangjun Dong, Tiantian Xu, and Yuhai Zhao. Classifying the multi-omics data of gastric cancer using a deep feature selection method. *Expert Systems with Applications*, 200:116813, 2022.

[46] Min-Koo Park, Jin-Muk Lim, Jinwoo Jeong, Yeongjae Jang, Ji-Won Lee, Jeong-Chan Lee, Hyungyu Kim, Euiyul Koh, Sung-Joo Hwang, Hong-Gee Kim, et al. Deep-learning algorithm and concomitant biomarker identification for nsclc prediction using multi-omics data integration. *Biomolecules*, 12(12):1839, 2022.

[47] Jie Feng, Limin Jiang, Shuhao Li, Jijun Tang, and Lan Wen. Multi-omics data fusion via a joint kernel learning model for cancer subtype discovery and essential gene identification. *Frontiers in genetics*, 12:647141, 2021.

[48] Junhao Liu, Zexuan Liu, Yangying Zhou, Manting Zeng, Sanshui Pan, Huan Liu, Qiong Liu, and Hong Zhu. Identification of a novel transcription factor prognostic index for breast cancer. *Frontiers in Oncology*, 11:666505, 2021.

[49] Ge Zhang, Zijing Xue, Chaokun Yan, Jianlin Wang, and Huimin Luo. A novel biomarker identification approach for gastric cancer using gene expression and dna methylation dataset. *Frontiers in Genetics*, 12:644378, 2021.

[50] Ming Zhang, Yilin Wang, Yan Wang, Longyang Jiang, Xueping Li, Hua Gao, Minjie Wei, and Lin Zhao. Integrative analysis of dna methylation and gene expression to determine specific diagnostic biomarkers and prognostic biomarkers of breast cancer. *Frontiers in Cell and Developmental Biology*, 8:529386, 2020.

[51] Meijie Zhang, Luyang Cheng, and Yina Zhang. Characterization of dysregulated lncrna-associated cerna network reveals novel lncrnas with cerna activity as epigenetic diagnostic biomarkers for osteoporosis risk. *Frontiers in Cell and Developmental Biology*, 8:184, 2020.

[52] Pengfei Liu and Weidong Tian. Identification of dna methylation patterns and biomarkers for clear-cell renal cell carcinoma by multi-omics data analysis. *PeerJ*, 8:e9654, 2020.

[53] Prasoon Joshi, Seokho Jeong, and Taesung Park. Sparse superlayered neural network-based multi-omics cancer subtype classification. *International Journal of Data Mining and Bioinformatics*, 24(1):58–73, 2020.

[54] Xiao Ouyang, Qingju Fan, Guang Ling, Yu Shi, and Fuyan Hu. Identification of diagnostic biomarkers and subtypes of liver hepatocellular carcinoma by multi-omics data analysis. *Genes*, 11(9):1051, 2020.

[55] Musalula Sinkala, Nicola Mulder, and Darren Martin. Machine learning and network analyses reveal disease subtypes of pancreatic cancer and their molecular characteristics. *Scientific reports*, 10(1):1212, 2020.

[56] Long-Yi Guo, Ai-Hua Wu, Yong-xia Wang, Li-ping Zhang, Hua Chai, and Xue-Fang Liang. Deep learning-based ovarian cancer subtypes identification using multi-omics data. *BioData Mining*, 13(1):1–12, 2020.

[57] Osama Hamzeh, Abedalrhman Alkhateeb, Julia Zhuoran Zheng, Srinath Kandalam, Crystal Leung, Govindaraja Atikukke, Dora Cavallo-Medved, Nallasivam Palanisamy, and Luis Rueda. A hierarchical machine learning model to discover gleason grade-specific biomarkers in prostate cancer. *Diagnostics*, 9(4):219, 2019.

[58] Wanxue Xu, Mengyao Xu, Longlong Wang, Wei Zhou, Rong Xiang, Yi Shi, Yunshan Zhang, and Yongjun Piao. Integrative analysis of dna methylation and gene expression identified cervical cancer-specific diagnostic biomarkers. *Signal transduction and targeted therapy*, 4(1):55, 2019.

[59] Nguyen Phuoc Long, Kyung Hee Jung, Nguyen Hoang Anh, Hong Hua Yan, Tran Diem Nghi, Seongoh Park, Sang Jun Yoon, Jung Eun Min, Hyung Min Kim, Joo Han Lim, et al. An integrative data mining and omics-based translational model for the identification and validation of oncogenic biomarkers of pancreatic cancer. *Cancers*, 11(2):155, 2019.

[60] Jianfeng Shu, Jinni Jiang, and Guofang Zhao. Identification of novel gene signature for lung adenocarcinoma by machine learning to predict immunotherapy and prognosis. *Frontiers in Immunology*, 14, 2023.

[61] Kountay Dwivedi, Ankit Rajpal, Sheetal Rajpal, Manoj Agarwal, Virendra Kumar, and Naveen Kumar. An explainable ai-driven biomarker discovery framework for non-small cell lung cancer classification. *Computers in Biology and Medicine*, 153:106544, 2023.

[62] Eskezeia Yihunie Dessie, Jan-Gowth Chang, and Ya-Sian Chang. A nine-

gene signature identification and prognostic risk prediction for patients with lung adenocarcinoma using novel machine learning approach. *Computers in Biology and Medicine*, 145:105493, 2022.

[63] Jnanendra Prasad Sarkar, Indrajit Saha, Anasua Sarkar, and Ujjwal Maulik. Machine learning integrated ensemble of feature selection methods followed by survival analysis for predicting breast cancer subtype specific mirna biomarkers. *Computers in Biology and Medicine*, 131:104244, 2021.

[64] Shuai Liu, Han Li, Qichen Zheng, Lu Yang, Meiyu Duan, Xin Feng, Fei Li, Lan Huang, and Fengfeng Zhou. Survival time prediction of breast cancer patients using feature selection algorithm crystall. *IEEE Access*, 9:24433–24445, 2021.

[65] Baoshan Ma, Yao Geng, Fanyu Meng, Ge Yan, and Fengju Song. Identification of a sixteen-gene prognostic biomarker for lung adenocarcinoma using a machine learning method. *Journal of Cancer*, 11(5):1288, 2020.

[66] Suman Ghosal, Shaoli Das, Ying Pang, Melissa K Gonzales, Thanh-Truc Huynh, Yanqin Yang, David Taieb, Joakim Crona, Uma T Shankavaram, and Karel Pacak. Long intergenic noncoding rna profiles of pheochromocytoma and paraganglioma: a novel prognostic biomarker. *International journal of cancer*, 146(8):2326–2335, 2020.

[67] Md Ali Hossain, Sheikh Muhammad Saiful Islam, Julian MW Quinn, Fazlul Huq, and Mohammad Ali Moni. Machine learning and bioinformatics models to identify gene expression patterns of ovarian cancer associated with disease progression and mortality. *Journal of biomedical informatics*, 100:103313, 2019.

[68] Feng Liu, Lu Xing, Xiaoqian Zhang, and Xiaoqi Zhang. A four-pseudogene classifier identified by machine learning serves as a novel prognostic marker for survival of osteosarcoma. *Genes*, 10(6):414, 2019.

[69] Jun Yu, Ming Zhu, Min Lv, Xiaoliu Wu, Xiaomei Zhang, Yuanying Zhang, Jintian Li, and Qin Zhang. Characterization of a five-microrna signature as a prognostic biomarker for esophageal squamous cell carcinoma. *Scientific Reports*, 9(1):19847, 2019.

[70] Adrian Pino Angulo, Kilho Shin, and Camilo Velázquez-Rodríguez. Improving the genetic bee colony optimization algorithm for efficient gene selection in microarray data. *Progress in Artificial Intelligence*, 7:399–410, 2018.

[71] Yongqing Zhang, Shuwen Xiong, Zixuan Wang, Yuhang Liu, Hong Luo, Beichen Li, and Quan Zou. Local augmented graph neural network for multi-omics cancer prognosis prediction and analysis. *Methods*, 213:1–9,

2023.

[72] Siamak Salimy, Hossein Lanjanian, Karim Abbasi, Mahdieh Salimi, Ali Najafi, Leili Tapak, and Ali Masoudi-Nejad. A deep learning-based framework for predicting survival-associated groups in colon cancer by integrating multi-omics and clinical data. *Heliyon*, 9(7), 2023.

[73] Xu Chen, Jing Yang, Zhengshu Lu, and Yanrui Ding. A 70-rna model based on svr and rfe for predicting the pancreatic cancer clinical prognosis. *Methods*, 204:278–285, 2022.

[74] Chengming Zhang, Yabin Chen, Tao Zeng, Chuanchao Zhang, and Luonan Chen. Deep latent space fusion for adaptive representation of heterogeneous multi-omics data. *Briefings in Bioinformatics*, 23(2):bbab600, 2022.

[75] Yeye Fan, Chunyu Kao, Fu Yang, Fei Wang, Gengshen Yin, He Yong Wang, Yongjiu, Liu Ji, Jiadong, and Liyuan. Integrated multi-omics analysis model to identify biomarkers associated with prognosis of breast cancer. *Frontiers in Oncology*, 12, 2022.

[76] Tongxin Wang, Wei Shao, Zhi Huang, Haixu Tang, Jie Zhang, Zhengming Ding, and Kun Huang. Mogonet integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. *Nature Communications*, 12(1):3445, 2021.

[77] Hua Chai, Xiang Zhou, Zhongyue Zhang, Jiahua Rao, Huiying Zhao, and Yuedong Yang. Integrating multi-omics data through deep learning for accurate cancer prognosis prediction. *Computers in biology and medicine*, 134:104481, 2021.

[78] Dafeng Xu, Yu Wang, Xiangmei Liu, Kailun Zhou, Jincai Wu, Jiacheng Chen, Cheng Chen, Liang Chen, and Jinfang Zheng. Development and clinical validation of a novel 9-gene prognostic model based on multi-omics in pancreatic adenocarcinoma. *Pharmacological Research*, 164:105370, 2021.

[79] Ning Zhao, Maozu Guo, Kuanquan Wang, Chunlong Zhang, and Xiaoyan Liu. Identification of pan-cancer prognostic biomarkers through integration of multi-omics data. *Frontiers in Bioengineering and Biotechnology*, 8:268, 2020.

[80] Yu-Heng Lai, Wei-Ning Chen, Te-Cheng Hsu, Che Lin, Yu Tsao, and Semon Wu. Overall survival prediction of non-small cell lung cancer by integrating microarray and clinical data with deep learning. *Scientific reports*, 10(1):4679, 2020.

[81] Lei Cui, Hansheng Li, Wenli Hui, Sitong Chen, Lin Yang, Yuxin Kang, Qirong Bo, and Jun Feng. A deep learning-based framework for lung cancer

survival analysis with biomarker interpretation. *BMC bioinformatics*, 21:1–14, 2020.

[82] Wenju Mo, Yuqin Ding, Shuai Zhao, Dehong Zou, and Xiaowen Ding. Identification of a 6-gene signature for the survival prediction of breast cancer patients based on integrated multi-omics data analysis. *PloS one*, 15(11):e0241924, 2020.

[83] Qianxing Mo, Roger Li, Dennis O Adeegbe, Guang Peng, and Keith Syson Chan. Integrative multi-omics analysis of muscle-invasive bladder cancer identifies prognostic biomarkers for frontline chemotherapy and immunotherapy. *Communications Biology*, 3(1):784, 2020.

[84] Zhiqiang Chang, Xiuxiu Miao, and Wenyuan Zhao. Identification of prognostic dosage-sensitive genes in colorectal cancer based on multi-omics. *Frontiers in Genetics*, 10:1310, 2020.

[85] Yang Yuan, Pan Qi, Wang Xiang, Liu Yanhui, Li Yu, and Mao Qing. Multi-omics analysis reveals novel subtypes and driver genes in glioblastoma. *Frontiers in Genetics*, 11:565341, 2020.

[86] Yanhui Jia, Meiyan Shen, Yan Zhou, and Huaiping Liu. Development of a 12-biomarkers-based prognostic model for pancreatic cancer using multi-omics integrated analysis. *Acta Biochimica Polonica*, 67(4):501–508, 2020.

[87] Tzong-Yi Lee, Kai-Yao Huang, Cheng-Hsiang Chuang, Cheng-Yang Lee, and Tzu-Hao Chang. Incorporating deep learning and multi-omics autoencoding for analysis of lung adenocarcinoma prognostication. *Computational Biology and Chemistry*, 87:107277, 2020.

[88] Nitish Kumar Mishra, Siddesh Southekal, and Chittibabu Guda. Survival analysis of multi-omics data identifies potential prognostic markers of pancreatic ductal adenocarcinoma. *Frontiers in genetics*, 10:624, 2019.

[89] Xuesi Dong, Ruyang Zhang, Jieyu He, Linjing Lai, Raphael N Alolga, Sipeng Shen, Ying Zhu, Dongfang You, Lijuan Lin, Chao Chen, et al. Trans-omics biomarker model improves prognostic prediction accuracy for early-stage lung adenocarcinoma. *Aging (Albany NY)*, 11(16):6312, 2019.

[90] Chen Peng, Yang Zheng, and De-Shuang Huang. Capsule network based modeling of multi-omics data for discovery of breast cancer-related genes. *IEEE/ACM transactions on computational biology and bioinformatics*, 17(5):1605–1612, 2019.

[91] Jayeon Lim, SoYoun Bang, Jiyeon Kim, Cheolyong Park, JunSang Cho, SungHwan Kim, et al. Integrative deep learning for identifying differentially expressed (de) biomarkers. *Computational and mathematical methods in*

medicine, 2019, 2019.

[92] Siquan Zhou, Guochen Ma, Hang Luo, Shufang Shan, Jingyuan Xiong, and Guo Cheng. Identification of 5 potential predictive biomarkers for alzheimer's disease by integrating the unified test for molecular signatures and weighted gene coexpression network analysis. *The Journals of Gerontology: Series A*, 78(4):653–658, 2023.

[93] Tianyi Zhao, Yang Hu, Jiajie Peng, and Liang Cheng. Deeplgp: a novel deep learning method for prioritizing lncrna target genes. *Bioinformatics*, 36(16):4466–4472, 2020.

[94] Broad Institute of MIT and Harward. FireBrowse. `https://xenabrowser.net/datapages/`, 2019. Online; accessed 31 January 2020.

[95] Yu Zhang, Yuanzhu Chen, and Ting Hu. Panda: Prioritization of autism-genes using network-based deep-learning approach. *Genetic epidemiology*, 44(4):382–394, 2020.

[96] Xue Jiang, Jingjing Zhao, Wei Qian, Weichen Song, and Guan Ning Lin. A generative adversarial network model for disease gene prediction with rna-seq data. *IEEE Access*, 8:37352–37360, 2020.

[97] Yonghyun Nam, Jong Ho Jhee, Junhee Cho, Ji-Hyun Lee, and Hyunjung Shin. Disease gene identification based on generic and disease-specific genome networks. *Bioinformatics*, 35(11):1923–1930, 2019.

[98] CHDI Foundation. Huntington's Disease in High Definition. `https://www.hdinhd.org/`, 2019. Online; accessed 10 February 2020.

[99] Runzhi Zhang and Susmita Datta. Adaptive sparse multi-block pls discriminant analysis: An integrative method for identifying key biomarkers from multi-omics data. *Genes*, 14(5):961, 2023.

[100] Jae-Ho Cheong, Sam C Wang, Sunho Park, Matthew R Porembka, Alana L Christie, Hyunki Kim, Hyo Song Kim, Hong Zhu, Woo Jin Hyung, Sung Hoon Noh, et al. Development and validation of a prognostic and predictive 32-gene signature for gastric cancer. *Nature communications*, 13(1):774, 2022.

[101] Pouria Samadi, Meysam Soleimani, Fatemeh Nouri, Fatemeh Rahbarizadeh, Rezvan Najafi, and Akram Jalali. An integrative transcriptome analysis reveals potential predictive, prognostic biomarkers and therapeutic targets in colorectal cancer. *BMC cancer*, 22(1):1–22, 2022.

[102] Kuo Yang, Kezhi Lu, Yang Wu, Jian Yu, Baoyan Liu, Yi Zhao, Jianxin Chen, and Xuezhong Zhou. A network-based machine-learning framework to identify both functional modules and disease genes. *Human Genetics*,

140:897–913, 2021.

[103] Haixia Shang and Zhi-Ping Liu. Network-based prioritization of cancer genes by integrative ranks from multi-omics data. *Computers in biology and medicine*, 119:103692, 2020.

[104] Pi-Jing Wei, Fang-Xiang Wu, Junfeng Xia, Yansen Su, Jing Wang, and Chun-Hou Zheng. Prioritizing cancer genes based on an improved random walk method. *Frontiers in genetics*, 11:377, 2020.

[105] Alberto Valdeolivas, Laurent Tichit, Claire Navarro, Sophie Perrin, Gaelle Odelin, Nicolas Levy, Pierre Cau, Elisabeth Remy, and Anaïs Baudot. Random walk with restart on multiplex and heterogeneous biological networks. *Bioinformatics*, 35(3):497–505, 2019.

[106] Zhen Zeng, Yuefeng Lu, Judong Shen, Wei Zheng, Peter Shaw, and Mary Beth Dorr. A random interaction forest for prioritizing predictive biomarkers. *arXiv preprint arXiv:1910.01786*, 2019.

[107] Naoya Fujita, Shinji Mizuarai, Katsuhiko Murakami, and Kenta Nakai. Biomarker discovery by integrated joint non-negative matrix factorization and pathway signature analyses. *Scientific reports*, 8(1):9743, 2018.

[108] Christos Dimitrakopoulos, Sravanth Kumar Hindupur, Luca Häfliger, Jonas Behr, Hesam Montazeri, Michael N Hall, and Niko Beerenwinkel. Network-based integration of multi-omics data for prioritizing cancer genes. *Bioinformatics*, 34(14):2441–2448, 2018.

[109] Yuanfang Guan, Tingyang Li, Hongjiu Zhang, Fan Zhu, and Gilbert S Omenn. Prioritizing predictive biomarkers for gene essentiality in cancer cells with mrna expression data and dna copy number profile. *Bioinformatics*, 34(23):3975–3982, 2018.

[110] Tiejun Zhang and Di Zhang. Integrating omics data and protein interaction networks to prioritize driver genes in cancer. *Oncotarget*, 8(35):58050, 2017.

[111] Huihui Fan, Hongying Zhao, Lin Pang, Ling Liu, Guanxiong Zhang, Fulong Yu, Tingting Liu, Chaohan Xu, Yun Xiao, and Xia Li. Systematically prioritizing functional differentially methylated regions (fdmrs) by integrating multi-omics data in colorectal cancer. *Scientific reports*, 5(1):12789, 2015.

[112] Qianlan Yao, Yanjun Xu, Haixiu Yang, Desi Shang, Chunlong Zhang, Yunpeng Zhang, Zeguo Sun, Xinrui Shi, Li Feng, Junwei Han, et al. Global prioritization of disease candidate metabolites based on a multi-omics composite network. *Scientific reports*, 5(1):17201, 2015.

[113] Vittorio Fortino, Pia Kinaret, Nanna Fyhrquist, Harri Alenius, and Dario Greco. A robust and accurate method for feature selection and prioritization

from multi-class omics data. *PloS one*, 9(9):e107801, 2014.

[114] Andrew Ke-Ming Lu, Shulan Hsieh, Cheng-Ta Yang, Xin-Yu Wang, and Sheng-Hsiang Lin. Dna methylation signature of psychological resilience in young adults: Constructing a methylation risk score using a machine learning method. *Frontiers in Genetics*, 13:1046700, 2023.

[115] NIH Policy for Data Management and Sharing. Broad-DREAM Gene Essentiality Prediction Challenge. `https://www.synapse.org/#!Synapse:syn2384331/wiki/62826`, 2019. Online; accessed 4 July 2020.

[116] St. Jude Research. Setting the Standard for Research and Discovery. `https://www.stjude.org/research`, 2019. Online; accessed 15 July 2020.

[117] JungHo Kong, Doyeon Ha, Juhun Lee, Inhae Kim, Minhyuk Park, Sin-Hyeog Im, Kunyoo Shin, and Sanguk Kim. Network-based machine learning approach to predict immunotherapy response in cancer patients. *Nature communications*, 13(1):3703, 2022.

[118] Danqing Luo, Jing Yang, Junji Liu, Xia Yong, and Zhimin Wang. Identification of four novel hub genes as monitoring biomarkers for colorectal cancer. *Hereditas*, 159(1):11, 2022.

[119] JungHo Kong, Heetak Lee, Donghyo Kim, Seong Kyu Han, Doyeon Ha, Kunyoo Shin, and Sanguk Kim. Network-based machine learning in colorectal and bladder organoid models predicts anti-cancer drug efficacy in patients. *Nature communications*, 11(1):5485, 2020.

[120] Yun Cai, Jie Mei, Zhuang Xiao, Bujie Xu, Xiaozheng Jiang, Yongjie Zhang, and Yichao Zhu. Identification of five hub genes as monitoring biomarkers for breast cancer metastasis in silico. *Hereditas*, 156(1):1–12, 2019.

[121] Anna Papiez, Christophe Badie, and Joanna Polanska. Machine learning techniques combined with dose profiles indicate radiation response biomarkers. *International Journal of Applied Mathematics and Computer Science*, 29(1), 2019.

[122] Ganxun Li, Dongyi Wan, Junnan Liang, Peng Zhu, Zeyang Ding, and Bixiang Zhang. Imopac: A web server for interactive multiomics and pharmacological analyses of patient-derived cancer cell lines. *Computational and Structural Biotechnology Journal*, 21:3705–3714, 2023.

[123] Anqi Lin, Chang Qi, Ting Wei, Mengyao Li, Quan Cheng, Zaoqu Liu, Peng Luo, and Jian Zhang. Camoip: a web server for comprehensive analysis on multi-omics of immunotherapy in pan-cancer. *Briefings in bioinformatics*, 23(3):bbac129, 2022.

[124] Furkan M Torun, Sebastian Virreira Winter, Sophia Doll, Felix M Riese,

Artem Vorobyev, Johannes B Mueller-Reif, Philipp E Geyer, and Maximilian T Strauss. Transparent exploration of machine learning for biomarker discovery from proteomics and omics data. *Journal of Proteome Research*, 22(2):359–367, 2022.

[125] Salim Ghannoum, Waldir Leoncio Netto, Damiano Fantini, Benjamin Ragan-Kelley, Amirabbas Parizadeh, Emma Jonasson, Anders Ståhlberg, Hesso Farhan, and Alvaro Köhn-Luque. Discbio: a user-friendly pipeline for biomarker discovery in single-cell transcriptomics. *International journal of molecular sciences*, 22(3):1399, 2021.

[126] Dongqiang Zeng, Zilan Ye, Rongfang Shen, Guangchuang Yu, Jiani Wu, Yi Xiong, Rui Zhou, Wenjun Qiu, Na Huang, Li Sun, et al. Iobr: multi-omics immuno-oncology biological research to decode tumor microenvironment and signatures. *Frontiers in immunology*, 12:687975, 2021.

[127] Huan Dong, Qiang Wang, Guosen Zhang, Ning Li, Mengsi Yang, Yang An, Longxiang Xie, Huimin Li, Lu Zhang, Wan Zhu, et al. Osdlbcl: An online consensus survival analysis web server based on gene expression profiles of diffuse large b-cell lymphoma. *Cancer Medicine*, 9(5):1790–1797, 2020.

[128] Harpreet Kaur, Anjali Dhall, Rajesh Kumar, and Gajendra PS Raghava. Identification of platform-independent diagnostic biomarker panel for hepatocellular carcinoma using large-scale transcriptomics data. *Frontiers in genetics*, 10:1306, 2020.

[129] Amrit Singh, Casey P Shannon, Benoît Gautier, Florian Rohart, Michaël Vacher, Scott J Tebbutt, and Kim-Anh Lê Cao. Diablo: an integrative approach for identifying key molecular drivers from multi-omics assays. *Bioinformatics*, 35(17):3055–3062, 2019.

[130] Dvir Netanely, Neta Stern, Itay Laufer, and Ron Shamir. Promo: an interactive tool for analyzing clinically-labeled multi-omic cancer datasets. *Bmc Bioinformatics*, 20:1–10, 2019.

[131] Harpreet Kaur, Sherry Bhalla, and Gajendra PS Raghava. Classification of early and late stage liver hepatocellular carcinoma patients from their genomics and epigenomics profiles. *PloS one*, 14(9):e0221476, 2019.

[132] Ajanthah Sangaralingam, Abu Z Dayem Ullah, Jacek Marzec, Emanuela Gadaleta, Ai Nagano, Helen Ross-Adams, Jun Wang, Nicholas R Lemoine, and Claude Chelala. 'multi-omic'data analysis using o-miner. *Briefings in bioinformatics*, 20(1):130–143, 2019.

[133] Mickael Leclercq, Benjamin Vittrant, Marie Laure Martin-Magniette, Marie Pier Scott Boyer, Olivier Perin, Alain Bergeron, Yves Fradet, and

Arnaud Droit. Large-scale automatic feature selection for biomarker discovery in high-dimensional omics data. *Frontiers in genetics*, 10:452, 2019.

[134] Xiaoyu Song, Jiayi Ji, Kevin J Gleason, Fan Yang, John A Martignetti, Lin S Chen, and Pei Wang. Insights into impact of dna copy number alteration and methylation on the proteogenomic landscape of human ovarian cancer via a multi-omics integrative analysis. *Molecular & Cellular Proteomics*, 18(8):S52–S65, 2019.

[135] Zefang Tang, Boxi Kang, Chenwei Li, Tianxiang Chen, and Zemin Zhang. Gepia2: an enhanced web server for large-scale expression profiling and interactive analysis. *Nucleic acids research*, 47(W1):W556–W560, 2019.

[136] Qiang Wang, Lu Zhang, Zhongyi Yan, Longxiang Xie, Yang An, Huimin Li, Yali Han, Guosen Zhang, Huan Dong, Hong Zheng, et al. Oscc: an online survival analysis web server to evaluate the prognostic value of biomarkers in cervical cancer. *Future Oncology*, 15(32):3693–3699, 2019.

[137] Yeongjun Jang, Jihae Seo, Insu Jang, Byungwook Lee, Sun Kim, and Sanghyuk Lee. Capssa: visual evaluation of cancer biomarker genes for patient stratification and survival analysis using mutation and expression data. *Bioinformatics*, 35(24):5341–5343, 2019.

[138] Magali Champion, Kevin Brennan, Tom Croonenborghs, Andrew J Gentles, Nathalie Pochet, and Olivier Gevaert. Module analysis captures pancancer genetically and epigenetically deregulated cancer driver genes for smoking and antiviral response. *EBioMedicine*, 27:156–166, 2018.

[139] Bingbing Xie, Zifeng Yuan, Yadong Yang, Zhidan Sun, Shuigeng Zhou, and Xiangdong Fang. Mobcdb: a comprehensive database integrating multi-omics data on breast cancer for precision medicine. *Breast cancer research and treatment*, 169:625–632, 2018.

[140] Akram Mohammed, Greyson Biegert, Jiri Adamec, and Tomáš Helikar. Cancerdiscover: an integrative pipeline for cancer biomarker and cancer class prediction from high-throughput sequencing data. *Oncotarget*, 9(2):2565, 2018.

[141] Jasmine Chong, Othman Soufan, Carin Li, Iurie Caraus, Shuzhao Li, Guillaume Bourque, David S Wishart, and Jianguo Xia. Metaboanalyst 4.0: towards more transparent and integrative metabolomics analysis. *Nucleic acids research*, 46(W1):W486–W494, 2018.

[142] Chun-Jie Liu, Fei-Fei Hu, Meng-Xuan Xia, Leng Han, Qiong Zhang, and An-Yuan Guo. Gscalite: a web server for gene set cancer analysis. *Bioinformatics*, 34(21):3771–3772, 2018.

[143] Javier E Flores, Daniel M Claborne, Zachary D Weller, Bobbie-Jo M Webb-Robertson, Katrina M Waters, and Lisa M Bramer. Missing data in multi-omics integration: Recent advances through artificial intelligence. *Frontiers in Artificial Intelligence*, 6:1098308, 2023.

[144] Akhil Kadiyala and Ashok Kumar. Applications of python to evaluate environmental data science problems. *Environmental Progress & Sustainable Energy*, 36(6):1580–1586, 2017.

[145] Marley W Watkins. *A step-by-step guide to exploratory factor analysis with R and RStudio*. Routledge, 2020.

[146] Pawan Kumar Mall, Pradeep Kumar Singh, Swapnita Srivastav, Vipul Narayan, Marcin Paprzycki, Tatiana Jaworska, and Maria Ganzha. A comprehensive review of deep neural networks for medical image processing: Recent developments and future opportunities. *Healthcare Analytics*, page 100216, 2023.

[147] Arnab Mukherjee, Suzanna Abraham, Akshita Singh, S Balaji, and KS Mukunthan. From data to cure: A comprehensive exploration of multi-omics data analysis for targeted therapies. *Molecular Biotechnology*, pages 1–21, 2024.

[148] Kaushik Roy Chaudhary. Knnimputer — way to impute missing values. `https://shorturl.at/cgxHY`, July 2022.

[149] Hao Ding. *Visualization and integrative analysis of cancer multi-omics data*. PhD thesis, The Ohio State University, 2016.

[150] Andrea Bommert, Xudong Sun, Bernd Bischl, Jörg Rahnenführer, and Michel Lang. Benchmark for filter methods for feature selection in high-dimensional classification data. *Computational Statistics & Data Analysis*, 143:106839, 2020.

[151] Arwinder Dhillon and Ashima Singh. ebrecap: extreme learning-based model for breast cancer survival prediction. *IET Systems Biology*, 14(3):160–169, 2020.

[152] Priti Bansal, Sachin Kumar, Sagar Pasrija, and Sachin Singh. A hybrid grasshopper and new cat swarm optimization algorithm for feature selection and optimization of multi-layer perceptron. *Soft computing*, 24:15463–15489, 2020.

[153] Bach Hoai Nguyen, Bing Xue, and Mengjie Zhang. A constrained competitive swarm optimiser with an svm-based surrogate model for feature selection. *IEEE Transactions on Evolutionary Computation*, 2022.

[154] Zitao Shen. Classification model - lightgbm with neural

net. https://zitaoshen.rbind.io/project/machine_learning/a-novel-hybrid-classification-model-lightgbm-with-neural-net/, Jan 2022.

[155] Kyoungmi Hwang, Dohyun Kim, Kyungsik Lee, Chungmok Lee, and Sungsoo Park. Embedded variable selection method using signomial classification. *Annals of Operations Research*, 254:89–109, 2017.

[156] G Thippa Reddy, M Praveen Kumar Reddy, Kuruva Lakshmanna, Rajesh Kaluri, Dharmendra Singh Rajput, Gautam Srivastava, and Thar Baker. Analysis of dimensionality reduction techniques on big data. *Ieee Access*, 8:54776–54788, 2020.

[157] Muta Tah Hira, M. A. Razzaque, Claudio Angione, James Scrivens, Saladin Sawan, and Mosharraf Sarkar. Integrated multi-omics analysis of ovarian cancer using variational autoencoders. *Scientific Reports*, 11(1), 2021.

[158] Yajie Meng and Min Jin. Hfs-slpee: A novel hierarchical feature selection and second learning probability error ensemble model for precision cancer diagnosis. *Frontiers in Cell and Developmental Biology*, 9:696359, 2021.

[159] Ke Wang, Ruo Chen, Zhuan Feng, Yu-Meng Zhu, Xiu-Xuan Sun, Wan Huang, and Zhi-Nan Chen. Identification of differentially expressed genes in non-small cell lung cancer. *Aging (Albany NY)*, 11(23):11170, 2019.

[160] Tomozumi Imamichi. DAVID Bioinformatics Resources. https://david.ncifcrf.gov/, July 2022.

[161] Changhee Lee, William Zame, Jinsung Yoon, and Mihaela Van Der Schaar. Deephit: A deep learning approach to survival analysis with competing risks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[162] Viv Bewick, Liz Cheek, and Jonathan Ball. Statistics review 12: survival analysis. *Critical care*, 8:1–6, 2004.

[163] András Lánczky and Balázs Győrffy. Web-based survival analysis tool tailored for medical research (KMplot): Development and implementation. *Journal of Medical Internet Research*, 23(7):1–7, 2021.

[164] Terry Therneau et al. A package for survival analysis in s. *R package version*, 2(7), 2015.

[165] Markus S Schröder, Aedín C Culhane, John Quackenbush, and Benjamin Haibe-Kains. survcomp: an r/bioconductor package for performance assessment and comparison of survival models. *Bioinformatics*, 27(22):3206–3208, 2011.

[166] Bo Wang, Aziz M. Mezlini, Feyyaz Demir, Marc Fiume, Zhuowen Tu,

Michael Brudno, Benjamin Haibe-Kains, and Anna Goldenberg. Similarity network fusion for aggregating data types on a genomic scale. *Nature Methods*, 11(3):333–337, 2014.

[167] Devesh Kumar, Rishabh Abhinav, and Naran Pindoriya. An ensemble model for short-term wind power forecasting using deep learning and gradient boosting algorithms. In *2020 21st National Power Systems Conference (NPSC)*, pages 1–6. IEEE, 2020.

[168] Jason Brownlee. Naive bayes for machine learning. `https://machinelearningmastery.com/naive-bayes-for-machine-learning/`, Feb 2022.

[169] Andy Liaw, Matthew Wiener, et al. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

[170] Alexey Natekin and Alois Knoll. Gradient boosting machines, a tutorial. *Frontiers in neurorobotics*, 7:21, 2013.

[171] Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Methods for interpreting and understanding deep neural networks. *Digital signal processing*, 73:1–15, 2018.

[172] Felix Wu, Tianyi Zhang, Amauri Holanda de Souza, Christopher Fifty, Tao Yu, and Kilian Q. Weinberger. Simplifying graph convolutional networks. *36th International Conference on Machine Learning, ICML 2019*, 2019-June:11884–11894, 2019.

[173] Daniel Mesafint Belete and Manjaiah D Huchaiah. Grid search in hyperparameter optimization of machine learning models for prediction of hiv/aids test results. *International Journal of Computers and Applications*, pages 1–12, 2021.

[174] Annibale Panichella. A systematic comparison of search-based approaches for lda hyperparameter tuning. *Information and Software Technology*, 130:106411, 2021.

[175] Mohit Malu, Gautam Dasarathy, and Andreas Spanias. Bayesian optimization in high-dimensional spaces: A brief survey. In *2021 12th International Conference on Information, Intelligence, Systems & Applications (IISA)*, pages 1–8. IEEE, 2021.

[176] LuHuai Jiao, Xin Ma, YuanNong Zhang, TaiFeng Jin, Song Fu, and BinBin Ni. A statistical analysis of the kappa-type energy spectrum distribution of radiation belt electrons observed by van allen probes. *Earth and Planetary Physics*, 2023.

[177] Ashima Singh, Arwinder Dhillon, Neeraj Kumar, M Shamim Hossain, Ghu-

lam Muhammad, and Manoj Kumar. ediapredict: An ensemble-based framework for diabetes prediction. *ACM Transactions on Multimidia Computing Communications and Applications*, 17(2s):1–26, 2021.

[178] Kevin Bi, Meng Xiao He, Ziad Bakouny, Abhay Kanodia, Sara Napolitano, Jingyi Wu, Grace Grimaldi, David A Braun, Michael S Cuoco, Angie Mayorga, et al. Tumor and immune reprogramming during immunotherapy in advanced renal cell carcinoma. *Cancer Cell*, 39(5):649–661, 2021.

[179] Mayo Foundation for Medical Education and Research (MFMER). Cancer survival rate: What it means for your prognosis print. `http://surl.li/kxpfi`, March 2023.

[180] Suhas V Vasaikar, Peter Straub, Jing Wang, and Bing Zhang. Linkedomics: analyzing multi-omics data within and across 32 cancer types. *Nucleic acids research*, 46(D1):D956–D963, 2018.

[181] Jie Tan, John H Hammond, Deborah A Hogan, and Casey S Greene. Adage analysis of publicly available gene expression data collections illuminates pseudomonas aeruginosa-host interactions. *BioRxiv*, page 030650, 2015.

[182] Baljinder Singh Heera, Yatindra Nath Singh, and Anjali Sharma. Congestion-aware dynamic rmcsa algorithm for spatially multiplexed elastic optical networks. In *2023 International Conference on Optical Network Design and Modeling (ONDM)*, pages 1–6. IEEE, 2023.

[183] Illumina. More than data: we empower understanding. `https://shorturl.at/aJKZ2`, March 2022.

[184] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[185] Apeksha Mittal, Amit Prakash Singh, and Pravin Chandra. Weight and bias initialization routines for sigmoidal feedforward network. *Applied Intelligence*, 51:2651–2671, 2021.

[186] Wei Wang. Bayesian optimization concept explained in layman terms. `https://shorturl.at/xEFP1`, March 2020.

[187] Kanehisa. KEGG PATHWAY Database. `https://www.genome.jp/kegg/pathway.html`, July 2022.

[188] Nicole J Chew, Elizabeth V Nguyen, Shih-Ping Su, Karel Novy, Howard C Chan, Lan K Nguyen, Jennii Luu, Kaylene J Simpson, Rachel S Lee, and Roger J Daly. Fgfr3 signaling and function in triple negative breast cancer. *Cell Communication and Signaling*, 18:1–17, 2020.

[189] Jie Mei, Yan Liu, Xinqian Yu, Leiyu Hao, Tao Ma, Qiang Zhan, Yan Zhang, and Yichao Zhu. Ywhaz interacts with daam1 to promote cell migration in

breast cancer. *Cell death discovery*, 7(1):221, 2021.

[190] Jian Li, Chunling Qi, Qing Li, and Fei Liu. Construction and validation of an aging-related gene signature for prognosis prediction of patients with breast cancer. *Cancer Reports*, 6(3):e1741, 2023.

[191] Zhixing Wang and Fan Wang. Identification of ten-gene related to lipid metabolism for predicting overall survival of breast invasive carcinoma. *Contrast Media & Molecular Imaging*, 2022, 2022.

[192] Mei Wu, Jin-Shu Pang, Qi Sun, Yu Huang, Jia-Yin Hou, Gang Chen, Jing-Jing Zeng, and Zhen-Bo Feng. The clinical significance of chek1 in breast cancer: a high-throughput data analysis and immunohistochemical study. *International journal of clinical and experimental pathology*, 12(1):1, 2019.

[193] Céline Grandvallet, Jean Paul Feugeas, Franck Monnien, Gilles Despouy, Perez Valérie, Guittaut Michaël, Eric Hervouet, and Paul Peixoto. Autophagy is associated with a robust specific transcriptional signature in breast cancer subtypes. *Genes & cancer*, 11(3-4):154, 2020.

[194] Nana Li, Yuan Miao, Yujia Shan, Bing Liu, Yang Li, Lifen Zhao, and Li Jia. Mir-106b and mir-93 regulate cell progression by suppression of pten via pi3k/akt pathway in breast cancer. *Cell death & disease*, 8(5):e2796–e2796, 2017.

[195] Dan Wang, Jin Ren, Hui Ren, Jin-ling Fu, and Dan Yu. Microrna-132 suppresses cell proliferation in human breast cancer by directly targeting foxa1. *Acta Pharmacologica Sinica*, 39(1):124–131, 2018.

[196] Jungho Kim, Sehee Oh, Sunyoung Park, Sungwoo Ahn, Yeonim Choi, Geehyuk Kim, Seung Il Kim, and Hyeyoung Lee. Circulating mir-221 and mir-222 as potential biomarkers for screening of breast cancer. 2019.

[197] Alexandra Triantafyllou, Nikolaos Dovrolis, Eleni Zografos, Charalampos Theodoropoulos, George C Zografos, Nikolaos V Michalopoulos, and Maria Gazouli. Circulating mirna expression profiling in breast cancer molecular subtypes: Applying machine learning analysis in bioinformatics. *Cancer Diagnosis & Prognosis*, 2(6):739, 2022.

[198] Forough Firoozbakht, Iman Rezaeian, Michele D'agnillo, Lisa Porter, Luis Rueda, and Alioune Ngom. An integrative approach for identifying network biomarkers of breast cancer subtypes using genomic, interactomic, and transcriptomic data. *Journal of Computational Biology*, 24(8):756–766, 2017.

[199] Fangfang Xu, Hui Li, and Chengjiu Hu. Mir-202 inhibits cell proliferation, invasion, and migration in breast cancer by targeting rock1 gene. *Journal of cellular biochemistry*, 120(9):16008–16018, 2019.

[200] Shen Li, Yan Xu, Yao Zhang, Lili Nie, Zhihua Ma, Ling Ma, Xiaoyu Fang, and Xiangyu Ma. Mendelian randomization analyses of genetically predicted circulating levels of cytokines with risk of breast cancer. *NPJ precision oncology*, 4(1):25, 2020.

[201] Indu Sinha, Rachel L Fogle, Gizem Gulfidan, Anne E Stanley, Vonn Walter, Christopher S Hollenbeak, Kazim Y Arga, and Raghu Sinha. Potential early markers for breast cancer: A proteomic approach comparing saliva and serum samples in a pilot study. *International journal of molecular sciences*, 24(4):4164, 2023.

[202] Siji Zhu, Haoyu Wang, Lin Lin, Xiaochun Fei, and Jiayi Wu. Primary breast osteosarcoma in a patient treated previously for ipsilateral invasive ductal carcinoma: An unusual case report with clinical and genomic features. *Frontiers in Oncology*, 12:7305, 2022.

[203] Wenhua Yang, Guozhong Cui, Mingjian Ding, Meng Yang, and Dianlu Dai. Microrna-124-3p. 1 promotes cell proliferation through axin1-dependent wnt signaling pathway and predicts a poor prognosis of triple-negative breast cancer. *Journal of Clinical Laboratory Analysis*, 34(7):e23266, 2020.

[204] Sushil Kumar, Ratnesh Kumar Srivastav, David W Wilkes, Taylor Ross, Sabrina Kim, Jules Kowalski, Srinivas Chatla, Qing Zhang, Anupma Nayak, Manti Guha, et al. Estrogen-dependent dll1-mediated notch signaling promotes luminal breast cancer. *Oncogene*, 38(12):2092–2107, 2019.

[205] Feng Qi, Wen-Xing Qin, and Yuan-Sheng Zang. Molecular mechanism of triple-negative breast cancer-associated brca1 and the identification of signaling pathways. *Oncology letters*, 17(3):2905–2914, 2019.

[206] Tian Hua, Bei-bei Zhao, Shao-bei Fan, Cai-fen Zhao, Yun-hong Kong, Rui-qing Tian, and Bao-ying Zhang. Prognostic implications of ppl expression in ovarian cancer. *Discover Oncology*, 13(1):35, 2022.

[207] Yan Kong, Zhi Qiao, Yongyong Ren, Georgi Z Genchev, Maolin Ge, Hua Xiao, Hongyu Zhao, and Hui Lu. Integrative analysis of membrane proteome and microrna reveals novel lung cancer metastasis biomarkers. *Frontiers in Genetics*, 11:1023, 2020.

[208] Jiaxi Yao, Yuchong Zhang, Mengling Li, Zuyu Sun, Tao Liu, Mingfang Zhao, and Zhi Li. Single-cell rna-seq reveals the promoting role of ferroptosis tendency during lung adenocarcinoma emt progression. *Frontiers in Cell and Developmental Biology*, 9:3951, 2022.

[209] Talip Zengin and Tuğba Önal-Süzek. Analysis of genomic and transcriptomic variations as prognostic signature for lung adenocarcinoma. *BMC*

bioinformatics, 21(14):1–28, 2020.

[210] Tao Li, Ning Liu, Guangyuan Zhang, and Ming Chen. Casp4 and casp8 as newly defined autophagy-pyroptosis-related genes associated with clinical and prognostic features of renal cell carcinoma. *Journal of Cancer Research and Therapeutics*, 18(7):1952–1960, 2022.

[211] Maryam Kohansal, Ali Ghanbarisad, Reza Tabrizi, Abdolreza Daraei, Mojtaba Kashfi, Hailin Tang, Cailu Song, and Yongming Chen. trna-derived fragments in gastric cancer: Biomarkers and functions. *Journal of Cellular and Molecular Medicine*, 26(18):4768–4780, 2022.

[212] Anqi Lin, Lingxuan Zhu, Aimin Jiang, Weiming Mou, Jian Zhang, Peng Luo, et al. Activation of the tgf-β pathway enhances the efficacy of platinum-based chemotherapy in small cell lung cancer patients. *Disease Markers*, 2022, 2022.

[213] Robert Fred Henry Walter, Robert Werner, Claudia Vollbrecht, Thomas Hager, Elena Flom, Daniel Christian Christoph, Jan Schmeller, Kurt Werner Schmid, Jeremias Wohlschlaeger, and Fabian Dominik Mairinger. Actb, cdkn1b, gapdh, grb2, rhoa and sdcbp were identified as reference genes in neuroendocrine lung cancer via the ncounter technology. *PLoS One*, 11(11):e0165181, 2016.

[214] Huihui Guo, Xilin Zhang, Qiuqiang Chen, Ying Bao, Chaohui Dong, and Xiang Wang. mir-132 suppresses the migration and invasion of lung cancer cells by blocking usp9x-induced epithelial-mesenchymal transition. *American journal of translational research*, 10(1):224, 2018.

[215] Chuchu Shao, Fengming Yang, Zhiqiang Qin, Xinming Jing, Yongqian Shu, and Hua Shen. The value of mir-155 as a biomarker for the diagnosis and prognosis of lung cancer: a systematic review with meta-analysis. *Bmc Cancer*, 19:1–10, 2019.

[216] Yuefan Guo, Guangxue Wang, Zhongrui Wang, Xin Ding, Lu Qian, Ya Li, Zhen Ren, Pengfei Liu, Wenjing Ma, Danni Li, et al. Reck-notch1 signaling mediates mir-221/222 regulation of lung cancer stem cells in nsclc. *Frontiers in Cell and Developmental Biology*, 9:663279, 2021.

[217] Li Wan, Lin Zhang, Kai Fan, Zai-Xing Cheng, Quan-Chao Sun, and Jian-Jun Wang. Circular rna-itch suppresses lung cancer proliferation via inhibiting the wnt/β-catenin pathway. *BioMed research international*, 2016, 2016.

[218] Bing Wang, Shengrong Yang, Yang Jia, Jianru Yang, Kun Du, Yujie Luo, Yunhe Li, Zhenghong Wang, Yi Liu, and Bing Zhu. Pcat19 regulates the proliferation and apoptosis of lung cancer cells by inhibiting mir-25-3p via

targeting the map2k4 signal axis. *Disease Markers*, 2022, 2022.

[219] Yuan Zhang, Jianbo Chen, Yunan Zhao, Lihong Weng, and Yiquan Xu. Ceramide pathway regulators predict clinical prognostic risk and affect the tumor immune microenvironment in lung adenocarcinoma. *Frontiers in Oncology*, 10:562574, 2020.

[220] Edward Shen, Ying Han, Changjing Cai, Ping Liu, Yihong Chen, Le Gao, Qiaoqiao Huang, Hong Shen, Shan Zeng, and Min He. Low expression of nlrp1 is associated with a poor prognosis and immune infiltration in lung adenocarcinoma patients. *Aging (Albany NY)*, 13(5):7570, 2021.

[221] Hend Baghoum, Hend Alahmed, Mahmood Hachim, Abiola Senok, Nour Jalaleddine, and Saba Al Heialy. Simulated microgravity influences immunity-related biomarkers in lung cancer. *International Journal of Molecular Sciences*, 24(1):155, 2022.

[222] Aoxiao He, Rongguiyi Zhang, Jiakun Wang, Zhihao Huang, Wenjun Liao, Yong Li, Cong Wang, Jun Yang, Qian Feng, and Linquan Wu. Tyk2 is a prognostic biomarker and associated with immune infiltration in the lung adenocarcinoma microenvironment. *Asia-Pacific Journal of Clinical Oncology*, 18(2):e129–e140, 2022.

[223] Mingyuan Luan, Fucheng Song, Shuyuan Qu, XI Meng, Junjie Ji, Yunbo Duan, Changgang Sun, Hongzong Si, and Honglin Zhai. Multi-omics integrative analysis and survival risk model construction of non-small cell lung cancer based on the cancer genome atlas datasets. *Oncology letters*, 20(4):1–1, 2020.

[224] Zhenyu Zhao, Boxue He, Qidong Cai, Pengfei Zhang, Xiong Peng, Yuqian Zhang, Hui Xie, and Xiang Wang. A model of twenty-three metabolic-related genes predicting overall survival for lung adenocarcinoma. *PeerJ*, 8:e10008, 2020.

[225] Wei Hou, Guo-Sheng Li, Li Gao, Hui-Ping Lu, Hua-Fu Zhou, Jin-Liang Kong, Gang Chen, Shuang Xia, and Hong-Yu Wei. Synj2 is a novel and potential biomarker for the prediction and treatment of cancers: from lung squamous cell carcinoma to pan-cancer. *BMC Medical Genomics*, 15(1):1–17, 2022.

[226] Tao Wang, Kun Su, Lianming Wang, Yanmei Shi, Yichun Niu, Yahao Zhou, Ayong Wang, and Tao Wu. Pan-cancer analysis of the oncogenic effects of g-protein-coupled receptor kinase-interacting protein-1 and validation on liver hepatocellular carcinoma. *Advances in clinical and experimental medicine: official organ Wroclaw Medical University*.

[227] Rongjiong Zheng, Haiqi Xu, Wenjie Mao, Zhennan Du, Mingming Wang, Meiling Hu, and Xiaolong Gu. A novel cpg-based signature for survival prediction of lung adenocarcinoma patients. *Experimental and Therapeutic Medicine*, 19(1):280–286, 2020.

[228] Yayun Gu, Huanyao Gao, Huan Zhang, August John, Xiujuan Zhu, Suganti Shivaram, Jia Yu, Richard M Weinshilboum, and Liewei Wang. Traf4 hyperactivates her2 signaling and contributes to trastuzumab resistance in her2-positive breast cancer. *Oncogene*, 41(35):4119–4129, 2022.

[229] Ming Niu, Ming Shan, Yang Liu, Yanni Song, Ji-Guang Han, Shanshan Sun, Xiao-Shuan Liang, and Guo-qiang Zhang. Dctpp1, an oncogene regulated by mir-378a-3p, promotes proliferation of breast cancer via dna repair signaling pathway. *Frontiers in oncology*, page 723, 2021.

[230] Manar Ahmed Abdel-Rahman, Mena Mahfouz, and Hany Onsy Habashy. Rrm2 expression in different molecular subtypes of breast cancer and its prognostic significance. *Diagnostic Pathology*, 17(1):1–8, 2022.

[231] So-Jeong Moon, Hyung-Jun Choi, Young-Hyeon Kye, Ga-Young Jeong, Hyung-Yong Kim, Jae-Kyung Myung, and Gu Kong. Cttn overexpression confers cancer stem cell-like properties and trastuzumab resistance via dkk-1/wnt signaling in her2 positive breast cancer. *Cancers*, 15(4):1168, 2023.

[232] Zelin Tian, Jianing Tang, Xing Liao, Qian Yang, Yumin Wu, and Gaosong Wu. Identification of a 9-gene prognostic signature for breast cancer. *Cancer Medicine*, 9(24):9471–9484, 2020.

[233] Hao Tian, Tingting Zhao, Yanling Li, Na Sun, Dandan Ma, Qiyun Shi, Guozhi Zhang, Qingqiu Chen, Kongyong Zhang, Ceshi Chen, et al. Chromobox family proteins as putative biomarkers for breast cancer management: A preliminary study based on bioinformatics analysis and qrt-pcr validation. *Breast Cancer: Targets and Therapy*, pages 515–535, 2022.

[234] Qian Zhou, Xiaofeng Liu, Mingming Lv, Erhu Sun, Xun Lu, and Cheng Lu. Genes that predict poor prognosis in breast cancer via bioinformatical analysis. *BioMed Research International*, 2021:1–8, 2021.

[235] Quang-Huy Nguyen, Hung Nguyen, Tin Nguyen, and Duc-Hau Le. Multi-omics analysis detects novel prognostic subgroups of breast cancer. *Frontiers in genetics*, 11:574661, 2020.

[236] Eriko Katsuta, Malgorzata Gil-moore, Justine Moore, Mohamed Yousif, Alex A Adjei, Yi Ding, Justin Caserta, Carmen M Baldino, Kelvin P Lee, Irwin H Gelman, et al. Targeting pim2 by jp11646 results in significant anti-tumor effects in solid tumors. *International journal of oncology*, 61(4):1–10,

2022.

[237] He JingSong, Guan Hong, Jianbo Yang, Zheng Duo, Fu Li, Chen WeiCai, Luo XueYing, Mao YouSheng, OuYang YiWen, Pan Yue, et al. sirna-mediated suppression of collagen type iv alpha 2 (col4a2) mrna inhibits triple-negative breast cancer cell proliferation and migration. *Oncotarget*, 8(2):2585, 2017.

[238] Roohollah Etemadi, Abedalrhman Alkhateeb, Iman Rezaeian, and Luis Rueda. Identification of discriminative genes for predicting breast cancer subtypes. In *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1184–1188. IEEE, 2016.

[239] Jia-Feng Huang, Chun-Jie Wen, Guo-Zhi Zhao, Yi Dai, Ying Li, Lan-Xiang Wu, and Hong-Hao Zhou. Overexpression of abcb4 contributes to acquired doxorubicin resistance in breast cancer cells in vitro. *Cancer chemotherapy and pharmacology*, 82:199–210, 2018.

[240] Mateusz Bujko, Paulina Kober, Michal Mikula, Marcin Ligaj, Jerzy Ostrowski, and Janusz Aleksander Siedlecki. Expression changes of cell-cell adhesion-related genes in colorectal tumors. *Oncology letters*, 9(6):2463–2470, 2015.

[241] Mingfei Xu, Chaoyue Liu, Lulan Pu, Jinrong Lai, Jingjia Li, Qianwen Ning, Xin Liu, and Shishan Deng. Systemic analysis of the expression levels and prognosis of breast cancer-related cadherins. *Experimental Biology and Medicine*, 246(15):1706–1720, 2021.

[242] Beste Turanli, Kubra Karagoz, Gholamreza Bidkhori, Raghu Sinha, Michael L Gatza, Mathias Uhlen, Adil Mardinoglu, and Kazim Yalcin Arga. Multi-omic data interpretation to repurpose subtype specific drug candidates for breast cancer. *Frontiers in genetics*, 10:420, 2019.

[243] Joaquín Derrac, Salvador García, Daniel Molina, and Francisco Herrera. A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms. *Swarm and Evolutionary Computation*, 1(1):3–18, 2011.

[244] Zichen Zhang and Wei-Chiang Hong. Application of variational mode decomposition and chaotic grey wolf optimizer with support vector regression for forecasting electric loads. *Knowledge-Based Systems*, 228:107297, 2021.

[245] Ming-Wei Li, Dong-Yang Xu, Jing Geng, and Wei-Chiang Hong. A hybrid approach for forecasting ship motion using cnn–gru–am and gcwoa. *Applied Soft Computing*, 114:108084, 2022.

[246] Suhas V. Vasaikar, Peter Straub, Jing Wang, and Bing Zhang. LinkedOmics:

Analyzing multi-omics data within and across 32 cancer types. *Nucleic Acids Research*, 46(D1):D956–D963, 2018.

[247] Andrea Asperti and Matteo Trentin. Balancing reconstruction error and kullback-leibler divergence in variational autoencoders. *IEEE Access*, 8:199440–199448, 2020.

[248] Shelly Sachdeva and Subhash Bhalla. Using knowledge graph structures for semantic interoperability in electronic health records data exchanges. *Information*, 13(2):52, 2022.

[249] Pakize Taylan, Fatma Yerlikaya-Özkurt, Burcu Bilgic Ucak, and Gerhard-Wilhelm Weber. A new outlier detection method based on convex optimization: application to diagnosis of parkinson's disease. *Journal of Applied Statistics*, 48(13-15):2421–2440, 2021.

[250] Xun Liu, Fangyuan Lei, Guoqing Xia, Yikuan Zhang, and Wenguo Wei. Adjmix: simplifying and attending graph convolutional networks. *Complex & Intelligent Systems*, pages 1–10, 2022.

[251] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30, 2017.

[252] Parampreet Kaur, Ashima Singh, and Inderveer Chana. Bsense: a parallel bayesian hyperparameter optimized stacked ensemble model for breast cancer survival prediction. *Journal of Computational Science*, 60:101570, 2022.

[253] Dongdong Sun, Ao Li, Bo Tang, and Minghui Wang. Integrating genomic data and pathological images to effectively predict breast cancer clinical outcome. *Computer methods and programs in biomedicine*, 161:45–53, 2018.

[254] Wei Chen, Yihuan Chen, Kai Zhang, Wanjing Yang, Xiang Li, Jun Zhao, Kangdong Liu, Ziming Dong, and Jing Lu. Agt serves as a potential biomarker and drives tumor progression in colorectal carcinoma. *International Immunopharmacology*, 101:108225, 2021.

[255] Peng Liu, Fan Li, Jian Lin, Ling Li, and Li Wang. Cdh1 as a therapeutic target for breast cancer treatment. *Scientific reports*, 9(1):1–13, 2019.

[256] Ana Carolina Pavanelli, Flavia Rotea Mangone, Piriya Yoganathan, Simone Aparecida Bessa, Suely Nonogaki, Cynthia AB de Toledo Osório, Victor Piana de Andrade, Iberê Cauduro Soares, Evandro Sobrosa de Mello, Lois M Mulligan, et al. Comprehensive immunohistochemical analysis of ret, bcar1, and bcar3 expression in patients with luminal a and b breast cancer subtypes. *Breast Cancer Research and Treatment*, 192(1):43–52, 2022.

[257] Maria Esperanza Rodriguez-Ruiz, Aitziber Buqué, Michal Hensler, Jonathan Chen, Norma Bloy, Giulia Petroni, Ai Sato, Takahiro Yamazaki, Jitka Fucikova, and Lorenzo Galluzzi. Apoptotic caspases inhibit abscopal responses to radiation and identify a new prognostic biomarker for breast cancer patients. *Oncoimmunology*, 8(11):e1655964, 2019.

[258] Sasagu Kurozumi, Mansour Alsaleem, Cíntia J Monteiro, Kartikeya Bhardwaj, Stacey EP Joosten, Takaaki Fujii, Ken Shirabe, Andrew R Green, Ian O Ellis, Emad A Rakha, et al. Targetable erbb2 mutation status is an independent marker of adverse prognosis in estrogen receptor positive, erbb2 non-amplified primary lobular breast carcinoma: a retrospective in silico analysis of public datasets. *Breast Cancer Research*, 22:1–11, 2020.

[259] Marit Valla, Elise Klæstad, Borgny Ytterhus, and Anna M Bofin. Ccnd1 amplification in breast cancer-associations with proliferation, histopathological grade, molecular subtype and prognosis. *Journal of Mammary Gland Biology and Neoplasia*, 27(1):67–77, 2022.

[260] Amal Ramadan, Maha Hashim, Amr Abouzid, and Menha Swellam. Clinical impact of pten methylation status as a prognostic marker for breast cancer. *Journal of Genetic Engineering and Biotechnology*, 19(1):1–11, 2021.

[261] Irene De Santo, Amelia McCartney, Ilenia Migliaccio, Angelo Di Leo, and Luca Malorni. The emerging role of esr1 mutations in luminal breast cancer as a prognostic and predictive biomarker of response to endocrine therapy. *Cancers*, 11(12):1894, 2019.

[262] Belhadj Amina, Addou Klouche Lynda, Seddiki Sonia, Belhadj Adel, Benammar H Jelloul, Medjamia Miloud, Sahraoui Tewfik, et al. Fibroblast growth factor receptor 1 protein (fgfr1) as potential prognostic and predictive marker in patients with luminal b breast cancers overexpressing human epidermal receptor 2 protein (her2). *Indian Journal of Pathology and Microbiology*, 64(2):254, 2021.

[263] Gabriele Corda, Gianluca Sala, Rossano Lattanzio, Manuela Iezzi, Michele Sallese, Giorgia Fragassi, Alessia Lamolinara, Hasan Mirza, Daniela Barcaroli, Sibylle Ermler, et al. Functional and prognostic significance of the genomic amplification of frizzled 6 (fzd6) in breast cancer. *The Journal of pathology*, 241(3):350–361, 2017.

[264] Dan-ni Ren, Jinxiao Chen, Zhi Li, Hongwei Yan, Yan Yin, Da Wo, Jiankang Zhang, Luoquan Ao, Bo Chen, Takashi K Ito, et al. Lrp5/6 directly bind to frizzled and prevent frizzled-regulated tumour metastasis. *Nature communications*, 6(1):1–13, 2015.

[265] Li Zhang, Cheng Fang, Xianqun Xu, Anling Li, Qing Cai, and Xinghua Long. Androgen receptor, egfr, and brca1 as biomarkers in triple-negative breast cancer: a meta-analysis. *BioMed research international*, 2015, 2015.

[266] Katerina Bouchalova, Gvantsa Kharaishvili, Jan Bouchal, Jana Vrbkova, Magdalena Megova, and Alice Hlobilkova. Triple negative breast cancer-bcl2 in prognosis and prediction. review. *Current drug targets*, 15(12):1166–1175, 2014.

[267] Maegan E Roberts, Sarah A Jackson, Lisa R Susswein, Nur Zeinomar, Xinran Ma, Megan L Marshall, Amy R Stettner, Becky Milewski, Zhixiong Xu, Benjamin D Solomon, et al. Msh6 and pms2 germ-line pathogenic variants implicated in lynch syndrome are associated with breast cancer. *Genetics in Medicine*, 20(10):1167–1174, 2018.

[268] Karina J Matissek, Maristela L Onozato, Sheng Sun, Zongli Zheng, Andrew Schultz, Jesse Lee, Kristofer Patel, Piiha-Lotta Jerevall, Srinivas Vinod Saladi, Allison Macleay, et al. Expressed gene fusions as frequent drivers of poor outcomes in hormone receptor–positive breast cancerfrequent expressed gene fusions in hr+ breast cancer. *Cancer discovery*, 8(3):336–353, 2018.

[269] Yanlin Li, Tiantian Wu, Ziluo Peng, Xianyan Tian, Qian Dai, Miao Chen, Jun Zhu, Song Xia, Aiqin Sun, Wannian Yang, et al. Ets1 is a prognostic biomarker of triple-negative breast cancer and promotes the triple-negative breast cancer progression through the yap signaling. *American Journal of Cancer Research*, 12(11):5074, 2022.

[270] Tongxin Wang, Wei Shao, Zhi Huang, Haixu Tang, Jie Zhang, Zhengming Ding, and Kun Huang. MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. *Nature Communications*, 12(1):1–13, 2021.

[271] Inuk Jung, Minsu Kim, Sungmin Rhee, Sangsoo Lim, and Sun Kim. Monti: A multi-omics non-negative tensor decomposition framework for gene-level integrative analysis. *Frontiers in Genetics*, page 1635, 2021.

[272] Hongjun Guo, Siqiao Wang, Min Ju, Penghui Yan, Wenhuizi Sun, Zhenyu Li, Siyu Wu, Ruoyi Lin, Shuyuan Xian, Daoke Yang, et al. Identification of stemness-related genes for cervical squamous cell carcinoma and endocervical adenocarcinoma by integrated bioinformatics analysis. *Frontiers in Cell and Developmental Biology*, 9:642724, 2021.

[273] Data Camp. impute.knn: A function to impute missing expression data. `https://www.rdocumentation.org/packages/impute/versions/1.46.0/topics/impute.knn`, Jan 2022.

[274] Kgaugelo Moses Dolo and Ernest Mnkandla. Differential evolution-based weighted voting stacking ensemble classifier for highly skewed binary data distribution. In *International Conference on Wireless Intelligent and Distributed Environment for Communication*, pages 13–27. Springer, 2023.

[275] Bohdan Pavlyshenko. Using stacking approaches for machine learning models. In *2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP)*, pages 255–258. IEEE, 2018.

[276] Bradley Boehmke Brandon Greenwell. Stacked models, hands-on machine learning with r. `https://bradleyboehmke.github.io/HOML/stacking.html`, January 2022.

[277] Ahsan Nazir, Jingsha He, Nafei Zhu, Ahsan Wajahat, Xiangjun Ma, Faheem Ullah, Sirajuddin Qureshi, and Muhammad Salman Pathan. Advancing iot security: A systematic review of machine learning approaches for the detection of iot botnets. *Journal of King Saud University-Computer and Information Sciences*, page 101820, 2023.

[278] Ghazala Sultan, Swaleha Zubair, Iftikhar Aslam Tayubi, Hans-Uwe Dahms, and Inamul Hasan Madar. Towards the early detection of ductal carcinoma (a common type of breast cancer) using biomarkers linked to the ppar (γ) signaling pathway. *Bioinformation*, 15(11):799, 2019.

[279] Stephanie Hunter, Braydon Nault, Kingsley Chukwunonso Ugwuagbo, Sujit Maiti, and Mousumi Majumder. Mir526b and mir655 promote tumour associated angiogenesis and lymphangiogenesis in breast cancer. *Cancers*, 11(7):938, 2019.

[280] Emmanuel Martinez-Ledesma, Roeland GW Verhaak, and Victor Treviño. Identification of a multi-cancer gene expression biomarker for cancer clinical outcomes using a network-based algorithm. *Scientific reports*, 5(1):11966, 2015.

[281] Lucas A Salas, Kevin C Johnson, Devin C Koestler, Dylan E O'Sullivan, and Brock C Christensen. Integrative epigenetic and genetic pan-cancer somatic alteration portraits. *Epigenetics*, 12(7):561–574, 2017.

[282] Hehuan Zhu, Jun Lu, Hu Zhao, Zhan Chen, Qiang Cui, Zhiwen Lin, Xuyang Wang, Jie Wang, Huiyue Dong, Shuiliang Wang, et al. Functional long noncoding rnas (lncrnas) in clear cell kidney carcinoma revealed by reconstruction and comprehensive analysis of the lncrna–mirna–mrna regulatory network. *Medical science monitor: international medical journal of experimental and clinical research*, 24:8250, 2018.

[283] Xiaofang Zong, Juexiu Fu, Ziyu Wang, Qianying Wang, et al. The diagnostic

and prognostic values of hoxa gene family in kidney clear cell renal cell carcinoma. *Journal of oncology*, 2022, 2022.

[284] Guangchun Han, Wei Zhao, Xiaofeng Song, Patrick Kwok-Shing Ng, Jose A Karam, Eric Jonasch, Gordon B Mills, Zhongming Zhao, Zhiyong Ding, and Peilin Jia. Unique protein expression signatures of survival time in kidney renal clear cell carcinoma through a pan-cancer screening. *BMC genomics*, 18(6):79–93, 2017.

[285] Xin Zheng, Tao Song, Changwei Dou, Yuli Jia, and Qingguang Liu. Ctbp2 is an independent prognostic marker that promotes gli1 induced epithelial-mesenchymal transition in hepatocellular carcinoma. *Oncotarget*, 6(6):3752, 2015.

[286] Soulaimane Aboulouard, Maxence Wisztorski, Marie Duhamel, Philippe Saudemont, Tristan Cardon, Fabrice Narducci, Anne-Sophie Lemaire, Firas Kobeissy, Eric Leblanc, Isabelle Fournier, et al. In-depth proteomics analysis of sentinel lymph nodes from individuals with endometrial cancer. *Cell Reports Medicine*, 2(6), 2021.

[287] Muhammad Ali, Derek B Archer, Priyanka Gorijala, Daniel Western, Jigyasha Timsina, Maria V Fernández, Ting-Chen Wang, Claudia L Satizabal, Qiong Yang, Alexa S Beiser, et al. Large multi-ethnic genetic analyses of amyloid imaging identify new genes for alzheimer disease. *Acta neuropathologica communications*, 11(1):68, 2023.

[288] Aparna Vasanthakumar, Justin W Davis, Kenneth Idler, Jeffrey F Waring, Elizabeth Asque, Bridget Riley-Gillis, Shaun Grosskurth, Gyan Srivastava, Sungeun Kim, Kwangsik Nho, et al. Harnessing peripheral dna methylation differences in the alzheimer's disease neuroimaging initiative (adni) to reveal novel biomarkers of disease. *Clinical epigenetics*, 12:1–11, 2020.

[289] Gustavo JJ Silva, Anja Bye, Hamid El Azzouzi, and Ulrik Wisløff. Micrornas as important regulators of exercise adaptation. *Progress in cardiovascular diseases*, 60(1):130–151, 2017.

[290] Srinivasulu Yerukala Sathipati and Shinn-Ying Ho. Identifying a mirna signature for predicting the stage of breast cancer. *Scientific reports*, 8(1):16138, 2018.

[291] Abeer A Raweh, Mohammed Nassef, and Amr Badr. A hybridized feature selection and extraction approach for enhancing cancer prediction based on dna methylation. *IEEE Access*, 6:15212–15223, 2018.

[292] Sara Alghunaim and Heyam H Al-Baity. On the scalability of machine-learning algorithms for breast cancer prediction in big data context. *Ieee*

Access, 7:91535–91546, 2019.

[293] Joung Min Choi and Heejoon Chae. mobrca-net: a breast cancer subtype classification framework based on multi-omics attention neural networks. *BMC bioinformatics*, 24(1):1–15, 2023.

[294] Baoshan Ma, Fanyu Meng, Ge Yan, Haowen Yan, Bingjie Chai, and Fengju Song. Diagnostic classification of cancers using extreme gradient boosting algorithm and multi-omics data. *Computers in biology and medicine*, 121:103761, 2020.

[295] Tony Yiu. Understanding random forest. `https://towardsdatascience.com/understanding-random-forest-58381e0602d2`, March 2022.

List of Publications

SCI Journals

- Arwinder Dhillon, Ashima Singh, Vinod Kumar Bhalla, " A Systematic Review on biomarker identification for cancer diagnosis and prognosis in multi-omics: from computational needs to machine learning and deep learning." Archives of Computational Methods in Engineering, vol. 30, no. 2, pp. 917-949, 2023, Springer. [Impact Factor: 9.7]

- Arwinder Dhillon, Ashima Singh, Vinod Kumar Bhalla, "Biomarker identification and cancer survival prediction using random spatial local best cat swarm and Bayesian optimized DNN." Applied Soft Computing, vol. 146, pp. 110649, 2023, Elsevier. [Impact Factor: 8.263]

- Arwinder Dhillon, Ashima Singh, Vinod Kumar Bhalla, "iMVAN: integrative multimodal variational autoencoder and network fusion for biomarker identification and cancer subtype classification." Applied Intelligence, vol. 53, no. 22, pp. 1-18, 2023, Springer. [Impact Factor: 5.019]

- Arwinder Dhillon, Ashima Singh, Vinod Kumar Bhalla, "HBS-STACK: Hierarchical Biomarker Selection and Stacked Ensemble model for Biomarker Identification and Cancer Prediction on Multi-Omics " Neural Computing and Applications, Springer. [Accepted]

Github Link: `https://github.com/Arwin94/PHD_WORK`