Development of Artificial Intelligence Models Based on Genomics Data for Precision Therapy of Glioma

THESIS
Submitted in partial fulfilment of the requirements
for the award of the degree of

Doctor of Philosophy
IN
BIOTECHNOLOGY

BY

SANA MUNQUAD
(Roll No. 718063)

Dr. ASIM BIKAS DAS
RESEARCH SUPERVISOR

DEPARTMENT OF BIOTECHNOLOGY
NATIONAL INSTITUTE OF TECHNOLOGY, WARANGAL
TELANGANA, INDIA
DECEMBER – 2023

With all the heart and soul, I dedicate this piece of my work to my beloved father and the almighty.

NATIONAL INSTITUTE OF TECHNOLOGY

WARANGAL – 506004.

DEPARTMENT OF BIOTECHNOLOGY

THESIS APPROVAL FOR Ph.D.

This thesis entitled ***"Development of Artificial Intelligence Models Based on Genomics Data for Precision Therapy of Glioma"*** by Ms. Sana Munquad, Roll No: 718063, is approved for the degree of Doctor of Philosophy.

Examiner

Supervisor
Dr. Asim Bikas Das
Associate Professor,
Department of Biotechnology,
National Institute of Technology Warangal.

Chairman
Prof. Prakash Saudagar,
Associate Professor and Head,
Department of Biotechnology.
National Institute of Technology Warangal

Date: Place:

NATIONAL INSTITUTE OF TECHNOLOGY

WARANGAL – 506004.

DEPARTMENT OF BIOTECHNOLOGY

CERTIFICATE

This is to certify that the thesis entitled **"Development of Artificial Intelligence Models Based on Genomics Data for Precision Therapy of Glioma"** that is being submitted by **Ms. SANA MUNQUAD (Roll No.718063)** in partial fulfilment for the award of Doctor of Philosophy **(Ph.D.)** in the Department of Biotechnology, National Institute of Technology, Warangal, is a record of bonafide work carried out by her under my guidance and supervision. The results embodied in this thesis have not been submitted to any other Universities or Institutes for the award of any degree or diploma.

Dr. Asim Bikas Das
(Supervisor)
Associate Professor
Department of Biotechnology
NIT-Warangal

NATIONAL INSTITUTE OF TECHNOLOGY
WARANGAL – 506004.
DEPARTMENT OF BIOTECHNOLOGY

DECLARATION

This is to certify that the work presented in the thesis entitled **"Development of Artificial Intelligence Models Based on Genomics Data for Precision Therapy of Glioma"**, is a bonafide work done by me under the supervision of Dr. Asim Bikas Das was not submitted elsewhere for the award of any degree.

I declare that this written submission represents my idea in my own wordsand where other's ideas or words have been included, I have adequately cited and referenced the sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misinterpreted or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be a cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not taken when needed.

Date:

Place: Warangal

Ms. Sana Munquad
(Roll No. 718063)

ACKNOWLEDGMENT

First of all, I would like to express my deepest gratitude to my supervisor ***Dr. Asim Bikas Das****, Associate Professor, Department of Biotechnology, NIT Warangal for the continuous support &inspiration throughout my Ph.D. study and research. It is because of his immense knowledge, motivation and scientific freedom has always encouraged me to continue work with new ideas and writing of this thesis. Under his able guidance, I have learned a lot, overcame many difficultiesand have been inspired for which I would always be indebted. I could not imagine having a betteradvisor and mentor for my Ph.D. study.*

I would like to thank my Doctoral scrutiny committee: ***Prof. Sourabh Roy,*** *Department of Physics****, Dr. R. Satish Babu,*** *Department of Biotechnology and* ***Dr. P. Srinivasa Rao****, Department of Biotechnology, National Institute of Technology, Warangal for their insightful comments and encouragement, but also for the hard question which incented me to widen my research from various perspectives.*

I sincerely thank, ***Prof. BIDYADHAR SUBUDHI****, present Director, and* ***Prof. N.V. RAMANA RAO*** *past Director National Institute of Technology, Warangal and other authorities who gave me this opportunity to carry out my research work.*

I wish to extend my heartfelt thanks to ***Dr. Onkara Perumal and Dr. R. Satish Babu****, previous HOD, Department of Biotechnology, for providing computational facility and also understanding and guiding me properly in this journey. I also thank to* ***Dr. Prakash Saudagar,*** *present HOD Dept. of Biotechnology, for infrastructure facilities provided for my study.*

I also thank my collaborator, ***Dr. Tapas Si,*** *University of Engineering & Management Jaipur,* ***Dr. Saurav Mallik,*** *University of Arizona,* ***Dr. Zhongming Zhao****, The University of Texas Health Science Center at Houston and* ***Dr. Aimin Li****, The University of Texas Health Science Center at Houston for lending me their expertise and intuition to my scientific and technical problems.*

I sincerely thank ***Science and Engineering Research Board (Mathematical Research Impact Centric Support)****, DST, Government of India for providing me the research facilities.*

I would like to extend my thanks to ***all the faculty members*** *in the* ***Department of Biotechnology****who has directly or indirectly helped me during my work. I also acknowledge all the* ***supportingand technical staff*** *of the* ***Department of Biotechnology,*** *for providing research environment and necessary facilities when required.*

I am blessed to meet ***Diksha, Snigdha, Bidesh, Momina, Jyotsna, Ganesh, Upendra & Dheeraj,*** *who havetaken a lot of care and efforts during my stay here at NIT Warangal. I cannot count on the help and affection but express my heartfelt thanks.*

I thank and appreciate all my friends, and scholars for all those fun, support and advice. Special thanks to all my wonderful hostel-mates who have all made this experience enjoyable. Most of all, my heartfelt gratitude and indebtedness to my parents ***Late Munquad Ali*** *&* ***Smt. Naghma Khatoon,*** *my Husband* ***Mr. Mohd Zeeshan Khan*** *and my daughter* ***Miss Mabrura Zeeshan Khan.*** *Whatever I'm today is all because of their immeasurable sacrifices, endless love and support. Even in the most difficult situations, unflinching courage & encouragement of my mother & husband have inspiredme to continue my research work, without which this thesis would not have been possible. Words will not suffice to express my reverence and thankfulness, but I dedicate this thesis to you all. I also owe special thanks to my sister* ***Darakhshan Munquad*** *and my mother in-law,* ***Smt. Anjum Ara*** *for their sincere prayers & blessings.*

*Finally, I thank, "****ALLAH****" whose divine light and warmth showered upon me the perseverance and enough strength to keep the momentum of work high even at tough moments ofresearch work.*

(Sana Munquad)

SUMMARY

Classifying lower-grade gliomas (LGGs) and glioblastoma multiforme (GBM) is a crucial step for accurate therapeutic intervention. The histopathological classification of various subtypes of LGG and GBM suffers from intraobserver and interobserver variability, leading to inaccurate classification and greater risk to patient health. Accurate diagnosis of glioma subtypes and identification of specific molecular features are crucial for clinicians for systematic treatment. The efficient machine learning and deep learning-based classification frameworks were designed to diagnose subtypes and grades of glioma using transcriptome and methylome data. The frameworks achieved >90% accuracy in diagnosing the subtypes. To evaluate the biological and clinical applicability of the classification, weighted gene correlation network analysis, co-expression, gene set enrichment, and survival analysis of the feature genes were performed, and subtype-specific prognostic biomarkers were identified. Furthermore, a biologically and clinically interpretable deep learning-based model was developed by integrating transcriptome and methylome data using an autoencoder for glioma subtype classification. The method of precision therapy is more expansive than just subtype classification, and the accurate selection of drugs is a major challenge. The poor prognosis of glioma patients brought attention to the need for effective therapeutic approaches for precision therapy. Here, algorithms relying on network medicine and artificial intelligence were deployed to design the framework for subtype-specific target identification and drug response prediction in glioma. Subtype-specific disease modules in each subtype of glioma were identified by utilizing a network-based approach, and drugs for which the disease module has a target gene were identified. However, the efficacy of anti-cancer drugs depends on the molecular profile of the cancer and varies among cancer patients due to intratumor heterogeneity. To overcome this limitation, the present thesis designed an AI-based drug response prediction model for different subtypes of glioma. Results showed that subtypes of gliomas respond differently to the drug, highlighting the importance of subtype-specific drug response prediction. Overall, the thesis shows how personalized therapies may be developed using AI models based on genomic data, which can result in cancer-specific treatments and better patient care.

Contents

List of Abbreviations

1D-CNN	One-dimensional convolutional neural network
AB	Adaboost
ABC	Artificial Bee Colony
AI	Artificial intelligence
ANN	Artificial neural network
AUC	Area under the curve
BiGRU	Bidirectional gated recurrent unit
BRCA	Breast Cancer
BRR	Bayesian Ridge Regression
CCLE	Cancer Cell Line Encyclopaedia
CGGA	Chinese Glioma Genome Atlas
CNC-AE	Autoencoder with concatenated inputs
CNN	Convolutional neural network
CNS	Central nervous system
CNV	Copy number variation
COAD	Colon Cancer
COSMIC	Catalogue Of Somatic Mutations In Cancer
CV	Cross-validation
DEDG	Differentially expressed driver genes
DEDGN	Differentially expressed driver gene network
DEGs	Differentially expressed genes
DFNForest	Deep Flexible Neural Forest
DIAMOnD	Disease Module Detection
DMRs	Differentially methylated regions
DNA	Deoxyribonucleic acid
DNN	Deep neural network
DT	Decision Tree
EGF	Epidermal growth factor
EGFR	Epidermal growth factor receptor
ELU	Exponential Linear Unit
FC	Fully connected
FDA	Food and Drug Administration
FN	False negative
FP	False positive
FPR	False positive rate
GA	Genetic Algorithm
GBM	Glioblastoma multiforme
GC	Gastric cancer
GDSC	Genomics of Drug Sensitivity in Cancer
GEO	Gene Expression Omnibus
GGA	Grouping Genetic Algorithm

GM	Geometric mean
GNB	Gaussian Naïve Bayes
GO	Gene ontology
GP	Genetic Programming
GPC	Gaussian process classification
GWAS	Genome-wide association studies
HCC	Hepatocellular carcinoma
HGG	High grade glioma
HNSC	Head and neck squamous cell cancers
HPA	Human Protein Atlas
HSD	Honestly significant difference
IC50	50% inhibitory concentration
IDH	Isocitrate dehydrogenase
IG	Information Gain
KEGG	Kyoto Encyclopedia of Genes and Genomes
KIRC	Kidney Cancer
KNN	k-nearest neighbors
KW	Kruskal-Wallis
LADC	Lung adenocarcinoma
LASSO	Least absolute shrinkage and selection operator
LCC	Largest connected component
LDA	Linear discriminant analysis
LGG	Lower-grade gliomas
LIHC	Liver Cancer
LR	Logistic regression
LUAD	Lung adenocarcinoma
LUSC	Lung squamous cell cancer
MCC	Matthews correlation coefficient
MCDM	Multi-Criteria Decision Making
MCFS	Monte Carlo Feature Selection
MIC-BQPSO	Maximum information coefficient binary quantum particle swarm optimization
MKL	Multiple Kernel Learning
MOFA	Multi-Omics Factor Analysis
MRI	Magnetic resonance imaging
mRMR	Maximum Relevancy and Minimum Redundancy
mRNA	Messenger Ribonucleic Acid
MSE	Mean squared error
NCA	Neighbourhood Component Analysis
NF-1	Neurofibromin 1
NGS	Next-generation sequencing
OS	Overall survival
OV	Ovarian cancer
PCA	Principal components analysis
PCC	Pearson correlation coefficient
PDGF	Platelet-derived growth factor

PPI	Protein-Protein Interactions
PRAD	Prostate Cancer
P-value	Probability value
QDA	Quadratic discriminant analysis
R2	Coefficient of determination
RELU	Rectified linear activation function
RF	Random forest
RLRs	RIG-I-like receptors
RMSE	Root Mean Square Error
RNA	Ribonucleic acid
RNA-seq	RNA sequencing
ROC	Receiver Operating Characteristics
RSEM	RNA-Sequencing by Expectation-Maximization
RWCE	Random Walk based cluster ensemble
SAE	Stacked Autoencoder
SCLC	Small cell lung cancer
SD	Standard Deviation
SNP	Single-nucleotide polymorphisms
SCLC	Squamous cell lung cancer
SVM	Support Vector Machine
SVM-RFE	Support vector machine recursive feature elimination
TCGA	The Cancer Genome Atlas
THCA	Thyroid cancer
TN	True negative
TOM	Topological overlap matrix
TOPSIS	Technique for Order of Preference by Similarity to Ideal Solution
TP	True positive
TPM	Transcripts Per Million
TSS	Transcription start site
UCSC	The University of California, Santa Cruz (genome browser)
UTR	Untranslated Region
VAE	Variational autoencoder
VEGF	Vascular endothelial growth factor
WCSRS	Wilcoxon sign rank sum
WGCNA	Weighted gene co-expression network analysis

List of Figures

List of Tables

CHAPTER 1

INTRODUCTION

Chapter 1: Introduction

Brain cancer is a destructive complex genomic disease with a low survival rate. It arises from an accumulation of genetic and epigenetic changes in somatic cells. Although brain cancer comprises only 2% of all human cancer, the treatment of brain cancer is challenging due to molecular heterogeneity and late diagnosis, which leads to an increase in the mortality rate [1]. The latest WHO documentation indicates over 100 distinct forms of brain tumors [2]. The most common form of brain tumor is gliomas, a category of primary brain tumors that arise from glial cells in the Central Nervous System (CNS), are highly heterogeneous, and exhibit a wide range of morphological, molecular, and clinical characteristics. This heterogeneity poses significant challenges to the diagnosis, treatment, and prognosis of gliomas [3]. Accurately classifying Gliomas type and grade is vital to improving brain cancer patients' prognosis. Due to the enormous complexity at the molecular level, the critical molecular driver of glioma is poorly understood. Gliomas are the most prevalent type of brain tumor, accounting for approximately 33% of all cases. Gliomas are classified based on histologic types and malignancy grades. Most gliomas are infiltrative and diffuse gliomas [4]. Gliomas are classified according to how rapidly or slowly the cells divide. Slower-growing gliomas are known as lower-grade glioma (LGG), whereas more aggressive or rapidly-growing gliomas are named glioblastoma multiforme (GBM). LGG occurs more frequently in younger people, whereas GBM is more common in older patients. The LGG is a grade II and III tumor with three subtypes: astrocytoma, oligoastrocytoma, and oligodendroglioma. Astrocytomas arise from astrocytes, and oligodendrogliomas arise from oligodendrocytes, whereas oligoastrocytomas are mixed glioma, including oligodendroglioma and astrocytoma cells. Therefore, the pathological classification of oligoastrocytoma remains controversial due to its resemblance to both subtypes [5]. Some of these LGG turn into GBM, the grade IV tumor, but others stay in this stage for a long time [6,7]. Glioblastoma (GBM) is characterised by its significant invasiveness and is recognised as the most lethal form of brain tumour in the adult population. Similarly, there are three subtypes of GBM, i.e., classical, proneural, and mesenchymal [8].

The prognosis for patients with GBM is poor, and median survival is 12 months [9]. The molecular mechanisms of GBM tumorigenesis is unknown. This leads to ineffective therapeutic intervention, and many patients relapse. However, with the current treatment options for both LGG and GBM, i.e., surgery, radiotherapy, and chemotherapy, patient life expectancy can be

increased, but these are not curative. Understanding the molecular features and identification of LGG and GBM subtypes is crucial to find the remedial solution. Due to distinct molecular characteristics, the subtypes of glioma have different clinical outcomes and responses to treatment, highlighting the importance of personalized medicine for brain cancer treatment [10]. To find the curative solution, understanding the molecular features and identification of glioma subtypes is crucial. Therefore, there is an urgent need to identify the subtype-specific molecular marker for personalized therapy. Each of the subtypes of LGG and GBM has distinct molecular features, and they can be classified using genomics, epigenomics, and mutational profiles for clinical diagnosis. The rapid progress in high throughput genomics technology has resulted in the generation of a substantial volume of data pertaining to various molecular layers that can be used to detect features and find genomic connections between them.

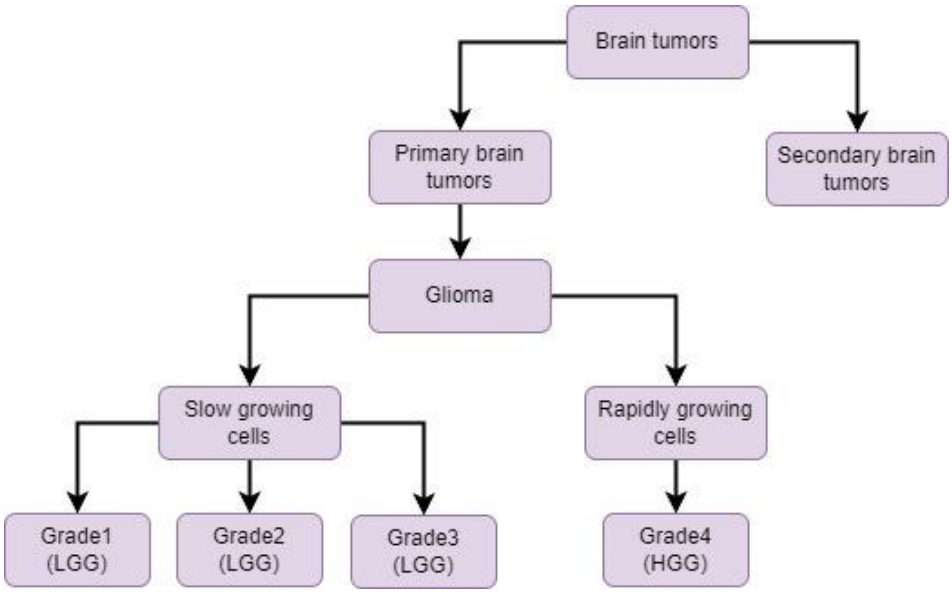

Figure 1.1: *Division of glioma based on slow growing and rapidly growing cells. Lower grade glioma (LGG) comes under in slow growing glioma and high-grade glioma (HGG) comes under the rapidly growing glioma.*

Data from sequencing experiments reveal that cancer initiation, progression, and maintenance are caused by perturbations in multiple genomics, epigenomics, and mutational factors. Gene expression and methylation are strongly interlinked processes; methylation levels in promoter regions influence gene expression by regulating transcription factor binding [11]. Similarly, gene expression and mutations are closely interconnected processes; mutations have the ability to modify gene expression within a protein coding sequence, hence impacting the

functionality of the protein and disrupting cellular pathways. Therefore, classification using multiple omics data, i.e., transcriptome, methylome, and mutation data can provide optimal features for the clinical diagnosis of cancer subtypes. Analysis of high-throughput omics data from different molecular layers can decipher the link between molecular signatures and cancer phenotype. Indeed, multi-omics data integration can elucidate how the molecular alterations at different layers contribute to disease formation and provide a global view of the molecular signature of disease. The biologically relevant diagnostic model can be developed by integrating gene expression, methylation, and mutation data because these data are biologically interlinked. Therefore, integration of multi-omics is essential to develop efficient AI-based diagnostic tools for accurate classification of cancer subtypes.

The method of precision therapy is more expansive than just subtype classification, and the accurate selection of drugs is a major challenge. However, anti-cancer drugs frequently do not work effectively. Molecular heterogeneity is a major contributor to cancer drug resistance, as it can create subpopulations of cancer cells with different mutations or molecular characteristics that allow them to survive even in the presence of the drug [12,13]. Hence, cancer patients with the same pathological conditions differ greatly in treatments. Therefore, the prediction of drug response, i.e., resistance or sensitivity, is essential for improving the efficacy of chemotherapy. Both accurate subtyping and drug response prediction model are crucial for the precision therapy of glioma. A DL-based model can improve the overall precision and efficacy of diagnostic processes using large-scale omics data. However, it is essential to design a biologically and clinically relevant AI-based diagnostic model to increase the reliability of diagnosis. In the present thesis, an AI-based diagnostic tool and drug response prediction model was developed for the precision therapy of glioma.

In this thesis, our focus has been directed towards the utilisation of artificial intelligence (AI), specifically machine learning (ML) and deep learning (DL) algorithms, to analyse data. These algorithms dive into the data, finding patterns, extracting a relationship between complex features discovering properties in genomics data such as transcriptome, methylome, and mutational data that the human brain cannot perceive. AI integration in brain cancer care could enhance brain cancer diagnosis and prognosis, stimulating the drug discovery and the development of effective therapies, aid clinical decision-making, and result in better health outcomes. Next, developed a biological interpretable model for glioma subtype classification and identify subtype-specific biomarkers of glioma. Further, a framework was developed by

combining network medicine and AI-based approaches to systematically integrate omics data to identify subtype-specific disease modules for precision therapy of glioma and drug response prediction models. Therefore, this thesis aims to contribute a novel perspective on enhancing the accuracy of cancer diagnosis, prognosis, and treatment with the help of AI.

Organization of the Thesis

The thesis presents the work in seven chapters, and the following section gives the outline.

Chapter 1: Presents a general introduction to brain cancer: presents diagnosing methods of brain cancer, data integration of different molecular levels of genomics in glioma, and precision therapy of glioma.

Chapter 2: Presents literature review on brain cancer classification, genomic alternations in brain cancer such as gene expression data, mutational profiles, DNA methylation data. Artificial intelligence techniques in cancer classification. Artificial intelligence techniques in genomics data types and Integration of genomic data for brain cancer classification. Cancer classification by utilizing machine learning and deep learning methods are also discussed, and finally, the Aim of the Work is enlisted. This chapter underlines gaps in the present knowledge and the objectives framed for the present study.

Chapter 3: Presents a detailed description of the Development of a machine learning-based framework for subtyping and grading of lower-grade glioma (LGG) using transcriptome data and the identification of biomarkers (objective 1). An efficient machine learning-based classification framework to diagnose LGG subtypes and grades using transcriptome data is presented. The development of an integrated feature selection method based on correlation and support vector machine (SVM) recursive feature elimination was done. Then machine learning models, i.e., Support Vector Machine (SVM) k-nearest neighbors (kNN), Gaussian Nave Bayes (GNB), Decision Tree (DT), and Random Forest (RF), were developed. This chapter also shows a 6-class classification model to predict grades and subtypes simultaneously. Furthermore, several predictive biomarkers using co-expression, gene set enrichment, and survival analysis were identified.

Chapter 4: Presents a detailed description of the Development of Deep learning and machine learning frameworks based on genomic data for subtyping glioblastoma multiforme (GBM) and identification of biomarkers (objective 2). In this chapter, a biologically interpretable and highly efficient deep learning framework based on a convolutional neural network for subtype

identification was developed. The classifiers were generated from high-throughput data at different molecular levels, i.e., transcriptome and methylome. An integrated subsystem of transcriptome and methylome data was also used to build the biologically relevant model. Furthermore, to evaluate the biological and clinical applicability of the classification, weighted gene correlation network analysis was performed, gene set enrichment, and survival analysis of the feature genes.

Chapter 5: Presents a detailed description of the Implementation of a deep learning embedding system for multi-omics data integration for the subtyping of Glioma (objective 3). Here, the transcriptome and methylome data of glioma patients were preprocessed, and differentially expressed features from both datasets were identified. Subsequently, a Cox regression analysis was performed to determine the genes and CpGs associated with survival. Gene set enrichment analysis was carried out to examine the biological significance of the features. Further, CpG and gene pairs were mapped based on the promoter region. The methylation and gene expression levels of these mapped CpGs and genes were embedded in a lower-dimensional space with an autoencoder. Next, ANN and CNN were used to classify subtypes using the latent features from embedding space. This chapter shows that multi-omics data integration performed better than mono-omics data for subtype classification.

Chapter 6: Presents a detailed description of the Identification of subtype-specific disease modules and development of drug response prediction models by combining network medicine and AI-based approaches (objective 4). The algorithms relying on network medicine and artificial intelligence were deployed to design the framework for subtype-specific target identification and drug response prediction in glioma. The driver mutations that were differentially expressed in each subtype of lower-grade glioma and glioblastoma multiforme were identified. Differentially expressed driver mutations were subjected to subtype-specific disease module identification. The drugs from the drug bank database were retrieved to target these disease modules. Next, a deep-learning-based drug response prediction framework was developed using the experimental drug screening data.

Chapter 7: Presents the conclusions drawn from the results. Potential future work and the scope of this work are also summarized.

CHAPTER 2

REVIEW OF LITERATURE

Chapter 2: Review of Literature

2.1 Brain cancer epidemiology

Brain tumors are relatively rare but deadly cancers that preferentially arise in the cerebral hemispheres of the central nervous system (CNS). According to the 2021 report by the World Health Organisation (WHO), the death rate of central nervous system (CNS) brain cancer exhibits the highest prevalence in Asia (https://gco.iarc.fr/). The 2020 global cancer statistics show that the number of new cases of brain cancer worldwide is about 308,102. The number of brain cancer deaths is 251,329 accounting for 2.8% of the total new cancer deaths [14]. The mortality rate among male brain cancer patients is 138,277, representing 3.2% of the overall new cancer-related deaths. Similarly, the mortality rate among female brain cancer patients is 113,052, accounting for 2.4% of the total new cancer-related deaths. The number of new patients with brain cancer in India is 31.5k, and the number of deaths is 26.7K, accounting for 43.9% and 48.6% of global cases, respectively. The mortality rate of brain cancer is relatively high because most patients are already at an advanced stage when they are detected. Therefore, finding effective biomarkers of early brain cancer is a vital way to reduce the high mortality rate of brain cancer. It is difficult to cure brain cancer because of its protected location. Nowadays, brain tumors can only partially cure by surgery, radiation, chemotherapy, and targeted therapy, having the risk of long-term patient morbidity. For targeted therapy, cancer grading is essential, as a cancer diagnosis is highly invasive, time-consuming, and expensive. There is a requirement for the development of affordable, and effective technologies for classifying and grading brain cancer and the advancement of targeted therapeutics involves the utilisation of molecularly targeted drugs that specifically target the cellular alterations responsible for the transformation of normal cells into cancerous cells, by which cancer should be detected at the earliest stage, so that many lives can be saved. However, when cancer is advanced, and the chances of survival are minimal, treatment becomes quite challenging.

2.2 Classification of brain cancer

Brain tumors can be classified either benign or malignant. Benign brain tumor cells have defined borders, seldom spread to nearby healthy cells, and develop slowly. Malignant brain tumor cells readily attack nearby cells, have hazy borders, and develop quickly. There are exist various forms of brain tumours, including choroid plexus tumours, embryonal tumours,

meningiomas, gliomas, and pituitary tumours. Glioma is the most prevalent type of brain tumour among all patients [1]. According to the growth rate of cancer cells, brain cancer can be classified into different grades, varying growth rates from low aggressive metastasis to high aggressive metastasis grade, i.e., low-grade glioma (LGG) and high-grade glioma (HGG). Grade I, II, and III come under the LGG, and Grade IV comes under the HGG, i.e., glioblastoma multiforme (GBM). The LGG are classified as astrocytoma, oligodendroglioma, and oligoastrocytoma. Astrocytomas arise from astrocytes, and oligodendrogliomas arise from oligodendrocytes, whereas oligoastrocytomas are mixed glioma, including oligodendroglioma and astrocytoma cells. Therefore, the pathological classification of oligoastrocytoma remains controversial due to its resemblance to both subtypes [5]. However, several attempts have been made to classify the oligoastrocytoma subtypes based on the genetic profile of individual markers [15–17]. GBM could be classified into four subtypes based on transcriptional features, i.e., classical, neural, proneural, and mesenchymal. However, recent findings suggest that the neural subtype probably arises due to the contamination of normal neuronal tissue tumor margins [8]. Therefore, GBM is currently classified into three subtypes. Histopathological-based diagnosis is the most common method for subtype identification. However, it often leads to inaccurate classification of subtypes due to inter-observer variability [18]. Accurate pathological subtype diagnosis is pivotal for optimal patient management. Because glioma subtypes are histologically and genetically heterogeneous, they differ in gene expression, mutation, and epigenetic states, which lead to different therapeutic response and clinical outcome [19,20].

2.3 Genomic alterations in brain cancer

A series of genetic abnormalities affect brain cancer at a molecular level, and impact signaling cascades that lead to the cancer initiation. The onset of brain cancer is triggered by modifications in the genome or DNA sequence, which disrupt the expression of genes, cell proliferation, and cellular behaviour. These alterations in gene sequences lead to the uncontrolled growth of cells. The anomalies include alterations in gene expression, DNA mutations, and variations in genome methylation profiles. Gene Expression data is primarily generated by two high throughput methods: RNA-sequencing (RNA-seq) and microarray. These techniques are efficient in capturing the genome-wide gene expression level. Cancer is a multifactorial disorder; therefore, studying all genes' expressions helps identify the critical player associated with cancer formation. Gene expression refers to the mRNA level for

particular genes at a given time point in the cells. Therefore, alteration of mRNA expression causes the change in protein level, thereby affecting the normal operations of the cells. Change of gene expression is a genome-level alteration in cancer; such alterations cause changes in cellular functions, resulting in a disease phenotype (Figure 2.1). Capturing such alteration from high dimensional gene expression data will aid in identifying the disease-causing gene and subsequently facilitate the discovery of novel biomarkers [21,22]. Furthermore, it is essential to track gene expression patterns to monitor the cancer progression from lower to a higher stage or understand the effectiveness of therapies [23–25]. This type of investigation required multiple comparisons of data from different time points.

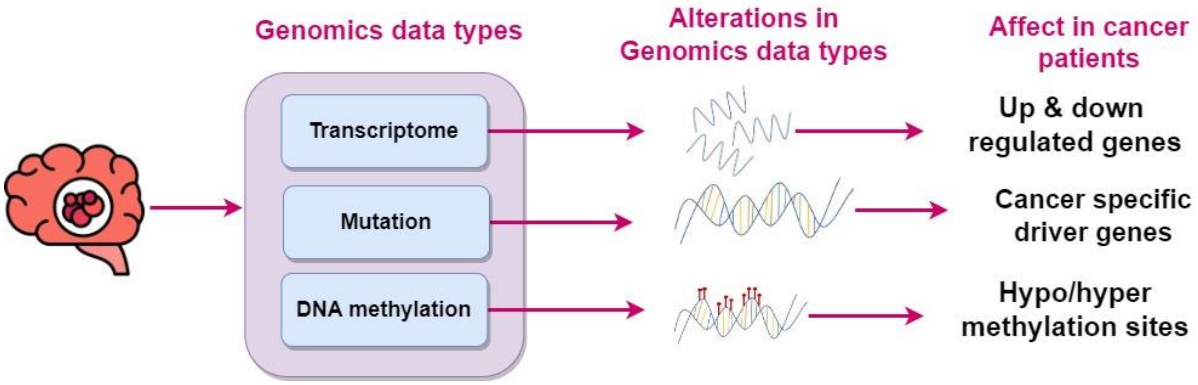

Figure 2.1: *Overview of genomics alternations in brain cancer.*

Altered expression of genes such as epidermal growth factor (EGF) [26], platelet-derived growth factor (PDGF) [27], vascular endothelial growth factor (VEGF) [28] and their receptor involved in the development of cancer progression. Several other genes are involved in brain tumors that exhibit irregular expression or have genetic changes have been discovered recently, including chromosomal irregularities of 1p19q have been reported in the oligodendroglioma [29].

The analysis of microarray data from various studies has revealed the presence of unique molecular profiles in high grade and low-grade glioma. The implementation of microarray technology enables the concurrent examination of alterations in the expression of numerous genes, hence facilitating the identification of gene sets that has the potential to predict glioma [30]. Differentially expressed genes (DEGs) have been found to be linked with the grade of tumours and the patient prognosis of glioma [31]. A comparison was conducted on 45 astrocytic tumours, consisting of 21 glioblastomas (GBMs) and 19 pilocytic astrocytomas. Through the

examination of a set of 360 genes, a distinct molecular signature was identified, enabling differentiation between GBMs and pilocytic astrocytomas [32]. The prognosis of IDH-mutant astrocytoma grade II to IV was found to be inversely correlated with the high expression of several specific genes. These genes include HOTAIRM1 [33], MCM6 [34], IRX1 [35], and MPC2 [36].

Several prior research have been conducted on bioinformatics analyses to examine the expression patterns of genes that are differentially expressed (DEGs) in patients with glioblastoma (GBM). These studies have also explored the functions of these DEGs in various pathways, molecular activities, and biological processes. Zou et al 2019, analysed the microarray data and reported that CDK1, BUB1B, NDC80, NCAPG, BUB1, CCNB1, TOP2A, DLGAP5, ASPM and MELK were significantly associated with carcinogenesis and the development of GBM [37].

Another most reported genetic alternations are mutations in glioma. Mutations in the coding genes alter the expression of mRNA; subsequently, proteins participate in the various biological processes inside the cells. Genetic mutations alter proteins in manners that induce the transformation of normal cells into malignant cells. A combination of mutated genes determines the deadliness of cancer (Figure 2.1). Several mutations in cancer contribute to the heterogeneity and complexity of the disease. Mutations greatly vary between the patients of the same cancer and pose a daunting obstacle to cancer treatment [38]. There are two types of mutations in cancer, driver mutations and passenger mutations. The driver mutations participate in uncontrolled cell growth, whereas passenger mutations usually do not involve oncogenesis. Driver mutations in the gene affect the protein structure and perturb normal biological processes. Mutation (driver mutations) in the tumor suppressor genes or oncogene can transform normal cells into cancer cells. Due to advancements in high throughput sequencing technologies, a large amount of tumor data has been generated that provide an opportunity to combine the gene mutations data and phenotypic information of cancer patients. However, the relationship between these mutations and clinical symptoms is still not revealed, this creates an obstacle to designing genomic medicine.

It has been determined that the primary genetic causes of gliomas are mutations in the isocitrate dehydrogenase (IDH) 1/2 enzymes [39–42]. In astrocytoma and oligodendroglioma IDH gene and PTEN mutations are frequently mutated and identified as a molecular marker in glioma [43]. The IDH1 gene is of great interest due to its association with alterations shown in both glioblastoma and low-grade gliomas. These mutations have been found in over 70% of

cases, encompassing the whole protein coding genes [44–46]. Genomic DNA samples, obtained from both tumour and normal tissues of glioma, were subjected to whole mutational data in order to determine the prevalent mutations in genes EGFR, ERBB2, IDH1, NF1, PIK3CA, PIK3R1, PTEN, PTPRD, RB1, and TP53 [47]. From the literature, it is observed that mutations in PI3K are novel prognostic markers in gliomas [48]. Mutations in FUBP1 and CIC are shown in astrocytomas oligoastrocytomas, and oligodendrogliomas [49]. ATRX [50], CDKN2A/B [51,52], EGFR [53,54], BRAF [55], H3 histone, family 3A (H3F3A) mutations [56] are reported in astrocytoma. Similarly, TERT promoter mutation was observed in oligodendroglioma [57]. Frequent mutations in some genes, such as TP53 and PTEN, have been observed in glioblastoma. However, it has been shown that these mutations do not significantly impact on prognosis [58,59]. The most commonly observed mutations in glioblastoma (GBM) subtypes are alterations in neurofibromin 1 (NF1), as well as epidermal growth factor receptor (EGFR) mutations [53]. Additionally, frequently observed genetic alterations in GBM include mutations in PIK3R1, PIK3CA, RB1, and IDH1, as reported in the data obtained from The Cancer Genome Atlas (TCGA) [60]. The presence of TERT promoter mutation has also been reported in glioblastoma multiforme (GBM) [57]. Multiple molecular markers are commonly observed in different subtypes of glioblastoma (GBM). For example, in the classical subtype, molecular markers such as PTEN, CHKN2, PDGFRA, TP53, and EGFR are frequently identified. In the mesenchymal subtype, NF-κB, NF1, PTEN, and in the proneural subtype, TP53, PI3K, IDH1, PDGFRA, and EGFR are commonly observed molecular markers [61]. Mutations occurring in these genes result in the activation of the PI3K/Akt and Ras/MAPK signalling pathways, hence presenting potential targets for therapeutic intervention [62].

Genome-wide association studies (GWASs) have additionally demonstrated that the heritable risk of glioma is influenced by common genetic variations. GWASs have successfully identified single-nucleotide polymorphisms (SNPs) at eight specific loci that have been found to influence glioma risk. These loci include 3q26.2 (near TERC), 5p15.33 (near TERT), 7p11.2 (near EGFR), 8q24.21 (near CCDC26), 9p21.3 (near CDKN2A/CDKN2B), 11q23.3 (near PHLDB1), 17p13.1 (TP53), and 20q13.33 (near RTEL1) [63–67]. Kinnersley, B., et al., 2015 identified the risk loci for glioblastoma (GBM) at 12q23.33 (rs3851634, near POLR3B) and non-GBM at 10q25.2 (rs11196067, near VTI1A), 11q23.2 (rs648044, near ZBTB16), 12q21.2 (rs12230172) and 15q24.2 (rs1801591, near ETFA) by using 1,490 cases and 1,723 controls [68]. The genes influenced by the risk single nucleotide polymorphisms (SNPs) that we have

identified is anticipated to result in enhanced understanding of the pathogenesis of this particular malignancy. GWASs have identified several genetic variants that are associated with glioma. These genetic variations have an impact on the DNA methylation levels of genes in close proximity and have a role in the susceptibility to cancer.

One of the most often observed genetic alterations in glioma is the changes the DNA methylation patterns. DNA methylation is a biological process that involves the addition of methyl groups to the DNA molecule without affecting the sequence. The level of DNA methylation can affect gene expression. When DNA methylation occurs at the promoter regions of a gene (also known as hypermethylation), it usually suppresses gene transcription and subsequently lowers gene expression levels [69]. Whereas a decrease in methylation level, known as hypomethylation, can elevate the gene expression level. Cancer pathogenesis is often caused by hypermethylation of tumor-suppressive genes and hypomethylation of oncogenes. Therefore, the methylation level of the promoter region is recently established as a promising biomarker in cancer (Figure 2.1). Methylation level not only influences the gene expression but also contributes to several other critical processes, such as X-chromosome inactivation, including genomic imprinting.

The MGMT, is a prominent epigenetic biomarker in glioma and its alterations play a central role in classification, treatment, and survival outcomes [70]. Wang et al. (2016) performed an analysis on three prognostic genes, specifically formyl peptide receptor 3, IKBKB interacting protein, and S100 calcium binding protein A9. These genes were selected from the comprehensive mRNA expression profile of the Chinese Glioma Genome Atlas (CGGA) and the RNAseq data obtained from The Cancer Genome Atlas (TCGA). They have conducted both univariate and multivariate Cox regression analyses on the entire genome mRNA expression in order to forecast the survival outcomes of patients with comparable MGMT methylation status. The expression of the three genes exhibits variation between glioblastoma multiforme (GBM) samples and non-cancerous tissues, and all three genes possess prognostic significance. The concurrent presence of these three genes holds predictive significance for individuals diagnosed with MGMT promoter-methylated glioblastomas [71]. DNA methylation may cause somatic mutations in driver genes, which would activate carcinogenesis. Additionally, DNA methylation can be utilized to categorise the molecular subtypes of glioma and might be more useful than gene expression changes. It has been determined that changes in epigenetic regulator genes are the primary cause of particular glioma subtypes with distinctive clinical

characteristics [72]. In the case of lower grade glioma, IDH1 or IDH2 mutations are associated with a specific pattern of DNA methylation while histone 3 mutations are commonly observed in paediatric high grade gliomas. These mutations are often accompanied by distinct DNA methylation patterns [73].

Here we have observed that the glioma subtypes of glioblastoma multiforme (GBM) and lower-grade glioma (LGG) have distinct genetic alterations. The utilisation of integrated analysis of genetic changes across many molecular levels can be employed to find the cause of cancer. The emergence of the big-data in the field of cancer genomics can be attributed to the widespread accessibility of genetic information facilitated by next-generation sequencing technology. The advancement of high-throughput genomics technologies, the size of genomics data is increasing exponentially. Simple statistical tests are inadequate for analyzing high-dimensional genome-wide data. The utilisation of artificial intelligence (AI) methodologies, including machine learning, and deep learning, is increasingly being employed to address the issues of scalability and high dimensionality of data. Currently, artificial intelligence (AI) is extensively employed for the purpose of cancer classification.

2.4 Artificial intelligence methods in cancer classification

Machine learning (ML) and deep learning (DL) is a subfield of artificial intelligence (AI) that involves the utilization of algorithms to acquire knowledge from datasets through training. These algorithms then utilize the acquired knowledge to draw inferences about outcomes based on the patterns and rules identified during the training process. ML, and DL algorithms and statistical modeling tasks have been found to enhance the efficiency and speed of processing complicated datasets in the field of cancer research. Due to the vast array of genomes data, the manual and rule-based analysis of such data poses significant challenges. Consequently, ML and DL approaches have gained prominence in this field, as they possess the capability to effectively handle the complexity inherent in genomics data and offer ease of implementation. Machine learning algorithms are employed to identify the grades and subtypes of the cancer. Classification is a form of supervised machine learning in which a model endeavors to accurately predict the appropriate label for a given set of input data. For example, in cancer subtype classification, models are trained on the training data (cancer samples) and predict the accurate output on the test data (new cancer samples or unseen data). In the figure 2.2 classification models are built on cancer samples used as a training data and new unseen cancer

samples are taken for model prediction. The aim of ML algorithms are to categorize the samples into different cancer subtypes. The models are able to predict the given samples are belong to which subtypes of cancer.

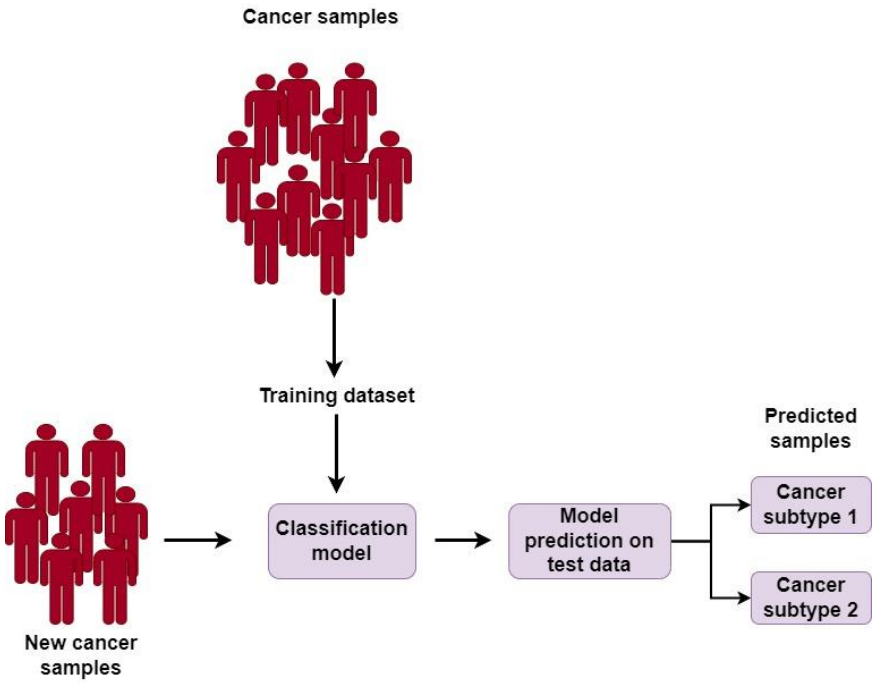

Figure 2.2: *Demonstration of ML based models for cancer subtype classification.*

There are two types of machine learning algorithms as: Un-supervised, and Supervised learnings. Unsupervised machine learning does not require the output label during the training phase. These algorithms possess the capability to identify similarities, and differences within a dataset. Principal Component Analysis (PCA), is an unsupervised technique that are well-suited for tasks such as dimensionality reduction and clustering analysis. These unsupervised machine learning techniques are unable to classify the cancer data. Therefore, the supervised machine learning techniques was used to efficiently classify the cancer grades and subtypes. Supervised machine learning refers to a type of learning algorithm that requires the availability of output labels during the training phase. This provision of labels enables the algorithm to discern and categorize data accurately, or make predictions based on the given labels. For instance, in the context of a cancer classification problem, the algorithm would require the labels 'cancer' and

'non-cancer' to effectively carry out its classification task. Several examples of supervised machine learning algorithms include Support Vector Machines (SVM), k-Nearest Neighbors (KNN), Naive Bayes (NB), Decision Trees (DT), Random Forest (RF), Logistic Regression (LR), and others. A brief description of each algorithm is provided here. SVM uses support vectors that separate data points in different hyperplanes [74]. SVM selects optimal hyperplanes for classification. In SVM, the hyperparameters were tuned i.e., regularization parameter c (c = 10), and applied a linear kernel to achieve higher accuracy. K-nearest neighbors (KNN) is a non-parametric method, and it utilizes neighboring elements that are trained to measure the accuracy of classification. KNN has two phases: the first is finding the nearest neighbors, and the second is assigning the class of a new sample using those neighbors by the majority vote rule [75,76]. Gaussian Naive Bayes is a probabilistic machine learning classifier based on the Bayes theorem. It assumes that the data from each label is drawn from a simple Gaussian distribution and considers all the features are independent [77]. In Decision Tree, the main aim is to create a model that predicts the value of a target variable by learning simple decision rules. The decision tree is constructed by repeatedly splitting a node into two child nodes, beginning from the root node containing the whole learning sample. Random forest is an ensemble technique used for classification by several estimators (decision trees). A logistic regression classifier predicts the response based on one or more predictor variables. It measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function.

DL are also divided into supervised and unsupervised learning techniques. Supervised deep learning techniques include Convolutional neural networks (CNN) and artificial neural networks (ANN), whereas unsupervised learning includes tsne and autoencoders. Unsupervised deep neural networks and autoencoders are successfully applied for model building in cancer genomics. Autoencoders are most recent and widely used in the DL. It is feed-forward neural network where input is the same as the output [78,79]. Autoencoders are used for analyzing transcriptomic, methylomic and mutational cancer data [80]. Conversely, supervised deep learning techniques ie., ANNs, which imitate the human brain, are feed-forward neural networks. ANNs are represented by a weighted, directed graph connecting inputs to a series of interconnected “hidden” layers that are composed of multiple nodes called “neurons,” that are in turn connected to an output layer [81]. ANNs are trained to recognize and categorize complex patterns. There are one input layer, one output layer and one hidden layer in the network. The hidden layers lies between the input and output layers. Similarly, CNN's are fully connected

networks, i.e., each neuron in a layer is directly connected to all neurons of the next layer. CNN's have a kernel that convolves the input to extract localized features and aggregate those using a pooling layer, enabling the model to extract features at all levels [82]. Therefore, it is efficient in extracting the relevant features from multidimensional data. These ML and DL learning approaches are used in abundance for cancer diagnosis and prognosis prediction (Figure 2.3).

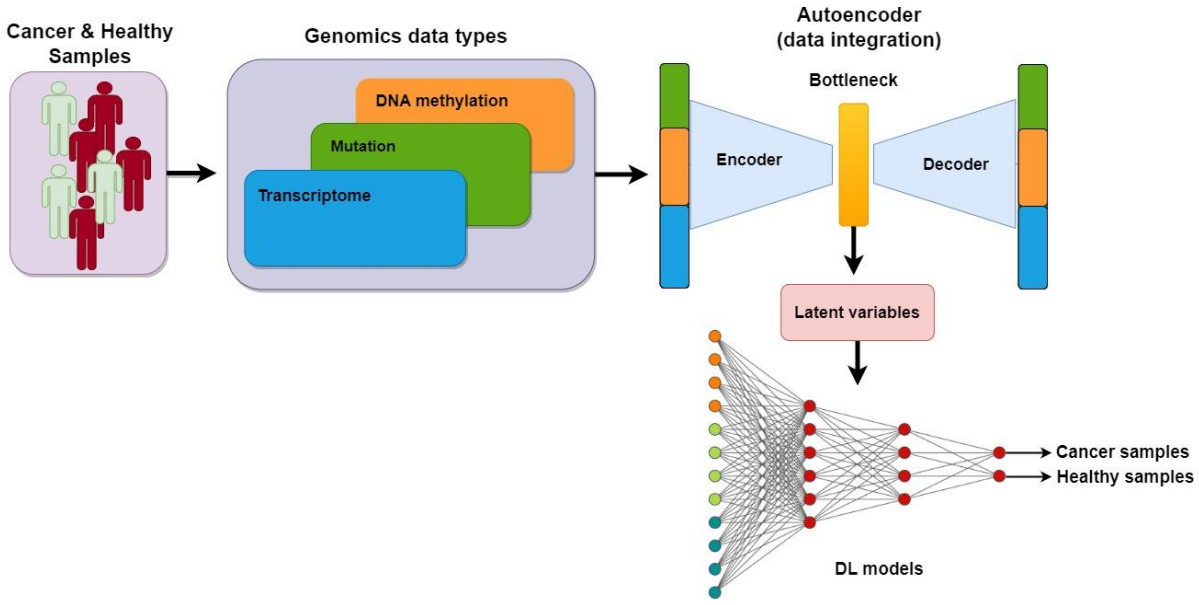

Figure 2.3: *Demonstration of DL based models for data integration and cancer classification.*

2.5 Artificial intelligence in genomics data types for cancer classification

2.5.1 Gene Expression Data for cancer classification

Several machine learning and deep learning techniques are applied to investigate cancer causing factors and help to cancer prediction and classification using gene expression data. The utilisation of machine learning (ML) approaches for cancer classification based on genomic data is driven by two key factors: cancer heterogeneity and the availability of various cancer genomic data. Machine learning models are utilised in the study of gene expression data to efficiently predict cancer categorization and diagnosis. Some unsupervised ML algorithms such

as Grouping Genetic Algorithm (GGA) and Bayesian latent (clustering algorithm) were used for multiclass cancer classification of gene expression RNA-Seq data [83,84]. In comparison to other machine learning algorithms, SVM exhibits a high degree of efficacy to find hidden patterns inside complex datasets [85]. SVM was incorporated to analyse the gene expression profiles of leukemia, gastric cancer, colon cancer, lung cancer, and prostate cancer samples for cancer classification [86–89]. Other machine learning models such as KNN were also used in the cancer classification and prediction of biomarkers by employing gene expression data that leads to improvement in the prognosis and treatment of cancer [90]. Random forest was used to classify lung cancer and esophageal squamous cell carcinoma [91,92]. Su, Y., et al. 2022 reported that random forest classifier was used to diagnose the colon cancer staging I, II, III and IV and predicted average accuracy of 99.81% and eight genes were selected as biomarkers such as GCNT2, GLDN, SULT1B1, UGT2B15, PTGDR2, GPR15, BMP5 and CPT2 [93]. Maniruzzaman, M., et al, 2019 developed a method of Statistical analysis of machine learning for classification of colon microarray gene expression data. They have used four statistical tests such as Wilcoxon sign rank sum (WCSRS), t-test, Kruskal-Wallis (KW) and F-test to identify the differential genes based on p-value. Further, identified differential expressed genes were employed in the ML models such as naïve Bayes (NB), support vector machine (SVM), linear discriminant analysis (LDA), Gaussian process classification (GPC), artificial neural network (ANN), quadratic discriminant analysis (QDA), decision tree (DT), random forest (RF), logistic regression (LR), and Adaboost (AB), to the classification of colon cancer. The mean accuracy of the machine learning system, encompassing all four statistical tests and all ten classifiers, was found to be 90.50% [94]. Kori, M. et al. 2022 applied CatBoost, Random forest, Decision Tree, KNN, Gradient boosting, MLP, LGBM, and XGB, classifier to classify gastric cancer into normal and cancer patients using microarray gene expression data in order to identify the novel biomarkers. The classification accuracy of eight ML models ranged from 92.6% to 89.4%. They have identified several novel biomarkers such as AES, CEBPZ, GRK6, HPGDS, SKIL, and SP3 for gastric cancer (GC), both in terms of diagnosis and prognosis [87]. WU et al. 2019 proposed a maximum information coefficient binary quantum particle swarm optimization (MIC- BQPSO) on SVM classifier method for brain cancer classification and achieved a classification accuracy of 74.64% [95]. Salem et al. 2017, developed a methodology that involved the utilization of the Information Gain (IG) technique for feature selection. Following this, the researchers deployed the Genetic Algorithm (GA) to perform feature reduction. Subsequently, they employed Genetic Programming (GP) to classify various types

of cancer, including brain cancer, into distinct categories of cancer and normal samples. The results of their analysis yielded a prediction accuracy of 86.67% [96]. The DL-based method outperformed the traditional statistical techniques. Yuan et al, 2020 reported the use of unsupervised deep autoencoder to extract features from high dimensional transcriptomic data. They implemented the supervised classifier, DeepC, using the extracted features to distinguish normal samples of different tissue origins. The authors have successfully diagnosed tumors in Pan-cancer with an accuracy of 90% or tissue-specific cancer with an average accuracy of 94% [97]. Shah et al. have employed a hybrid deep learning model based on Laplacian Score and CNN (LS-CNN) to classify brain cancer using the microarray gene expression data. They have shown that the LS-CNN model (average accuracy = 97%) outperformed the traditional machine learning model in terms of accuracy [98]. Mohammed et al. (2021) proposed a novel stacking ensemble deep learning approach utilizing a one-dimensional convolutional neural network (1D-CNN). The objective of their research was to conduct a multi-class classification of five prevalent cancers in women, including breast, lung, colorectal, thyroid, and ovarian, using RNA-seq data. The researchers employed the least absolute shrinkage and selection operator (LASSO) as a technique for selecting features. The researchers conducted a comparative analysis of the outcomes of the newly proposed model, with and without LASSO, with the outcomes of the single 1D-CNN and several machine learning approaches. The findings indicate that the proposed model with LASSO and without LASSO exhibit superior performance in comparison to alternative classifiers. The utilization of a one-dimensional convolutional neural network (1D-CNN) in conjunction with the least absolute shrinkage and selection operator (LASSO) yielded a prediction accuracy of 99.22%. In contrast, when LASSO was not employed, the prediction accuracy was somewhat lower at 98.06% [99]. Rezaee et al. (2022) proposed a novel approach that employs ensemble learning in conjunction with deep neural network (DNN) for the purpose of classifying three distinct forms of cancer, namely diffuse large cell lymphoma, leukemia, and prostate cancer. The prediction accuracies obtained were 97.51%, 99.6%, and 96.34%, respectively. Moreover, the researchers confirmed the model's generalizability by assessing its performance on brain tissue lesions associated with multiple sclerosis [100]. Almarzouki (2022) proposed a novel approach known as the Artificial Bee Colony (ABC) technique for the purpose of feature selection. This strategy was employed in conjunction with Convolutional Neural Networks (CNNs) to classify gene expression data from kidney, brain, and lung tissues into cancerous and normal states. To achieve this, the

researchers combined all available datasets. The Convolutional Neural Network demonstrates a high level of accuracy, i.e, 96.43% [101].

Further, several other deep learning models have been developed to predict the regulation of the gene expression, such as DEcode, which can predict the differential gene expression based on binding sites on RNAs and promoters [102]. Similarly, Deepdiff and DeepChrome predict the gene expression from histone modifications [103,104]. These DL-based tools can be explored for the diagnosis of cancer. Furthermore, it was observed that unsupervised DL was also implemented to generate the gene expression cluster in brain cancer and has been used to improve the model accuracy [105].

2.5.2 Mutational data for cancer classification

In several studies, artificial intelligence and machine learning have been successfully employed to draw the relationship between cancer mutations and clinical symptoms [106,107], including driver gene identification [108,109], drug development [110], and precision oncology [111]. However, challenges such as high data sparsity, and short sample size, are roadblocks for superior classification performance using the mutation-related genomic data in cancer. Somatic mutations provide us a great opportunity to investigate cancer classification using machine learning. Gene mutation profiles are used to classify, characterize and predict the subgroups of cancers. In breast cancer patients, somatic mutation profiles were used to classify the subgroup using machine learning methods such as Random Forest (RF), Support Vector Machine (SVM), C4.5, Naïve Bayes, and k-Nearest Neighbor (KNN). Among all classifier performances RF outperformed and achieved the average prediction accuracy of 70.86% than the other machine learning models [112]. In another case of breast cancer, machine learning methods such as naive bayes and KNN were used to classify the breast cancer patients and healthy patients. The K-nearest neighbors (KNN) algorithm had the highest classification accuracy, with a rate of 97.51%, while the Naive Bayes (NB) classifier displayed a classification accuracy of 96.19% [113]. Furthermore, Li, Y. et al, (2020), proposed an ensemble machine learning model including five classifiers for cancer classification of fourteen types of cancer utilizing mutation data. They achieved an overall accuracy of 71.46% [114]. Chen et al. investigated the distribution of 1,760,846 somatic mutations observed in 230,255 cancer patients. They have employed a Support Vector Machine (SVM) approach to analyze these mutations in conjunction with gene function information, in 17 types of cancer including glioma. They conducted a multiclass

classification experiment employing the gene symbol, somatic mutation, chromosome, and gene functional pathway. The prediction performance of primary sites in terms of accuracy was reached to 57% by using genes as features, by including the genes, mutation and chromosome information, it was improved to 62% [115]. Similarly, a machine learning model was developed for distinguishing between driver and passenger mutations in GBM using sequence based features and physiochemical properties, named GBMDriver; showed accuracy of 73.59% and AUC score of 0.82. The accuracy was 81. 99% and AUC was 0.87 on 10-fold cross validation. By this method driver mutations in glioblastoma are prioritized and therapeutic targets are identified [116].

Palazzo, M., et al (2019) developed a pipeline based on an unsupervised deep learning method known as autoencoder. This pipeline aims to uncover concealed patterns within lower dimensional space using somatic mutation data derived from a diverse range of 40 tumor types and subtypes. In order to assess the effectiveness of the acquired somatic mutation embedding, a combination of kernel learning and hierarchical cluster analysis was employed. This approach yielded an accuracy rate exceeding 75% across various types of cancer, with the exception of stomach, colorectal, and liver malignancies [117]. Furthermore, DNN-Boost model was also developed to classify the tumor and normal samples by employing mutation data [118]. Yuan et al., developed DeepGene, an advanced cancer type classifier based on deep learning and DNA point mutation data. DeepGene was designed to extract the critical features between combinatorial point mutations and cancer types [119]. Furthermore, Zeng et al., proposed deep learning-based model DeepCues that utilizes CNN to find features from DNA sequencing data for cancer classification. DeepCues uses whole-exome sequencing, germline variants, and somatic mutations, including insertions and deletions, for feature extraction and classification. The overall accuracy of DeepCues is 77.6% [120].

2.5.3 DNA methylation data for cancer classification

The investigation of methylation patterns assumes a crucial role in comprehending the progression of diseases. Therefore, methylome data is used in cancer classification and diagnosis [121]. Several machine learning approaches are developed to accurately classify the cancers such as lung cancer [122], breast cancer [123], and head and neck squamous cell cancers (HNSCs) [124] by utilizing DNA methylation data. Ren, J., et al, (2022) utilized DNA methylation data to identify potential biomarkers in different subtypes of sarcoma. They employed an

unsupervised machine learning algorithm, specifically boruta, for feature filtration. This was followed by the use of LASSO, Light Gradient Boosting Machine (GBM), and Monte Carlo Feature Selection (MCFS) for feature selection. To develop a classification model, they employed supervised machine learning methods including decision trees (DT) and random forests (RF). The random forest (RF) model demonstrated superior predictive accuracy compared to the decision tree (DT) model. The prediction accuracies of LASSO with RF, Light GBM with RF, and MCFS with RF were found to be 98.70%, 99.10%, and 98.70% respectively. The present study employed a specific approach to identify biomarkers that exhibit gene expression patterns derived from the annotation of methylation site features that are strongly connected. Notably, the genes PRKAR1B, INPP5A, and GLI3 were found to be associated with these biomarkers. They were found to be linked with sarcoma [125]. Cai, Z.,(2015) developed a ML based method to classify the lung cancers types into small cell lung cancer (SCLC), lung adenocarcinoma (LADC), and squamous cell lung cancer (SQCLC) using DNA methylation data. RF and Maximum Relevancy and Minimum Redundancy (mRMR) were used to classify LADC, SQCLC and SCLC and achieved a prediction accuracy of 86.54% [126]. Moreover, ML models such as XGBoost, SVM, RF, NB and KNN were employed to classify the different types of cancers by employing DNA methylation data [127]. These studies have contributed novel perspectives on cancer detection from an epigenetic standpoint, and may lead to personalized and therapeutic approaches. Interestingly, survival of the patients were also predicted by the machine learning models in different types of cancer by DNA methylation data [128].

Eissa et al. (2022) constructed a deep neural network (DNN) model based on DNA methylation data for the purpose of classifying various types of cancer, including Breast Cancer (BRCA), Ovary Cancer (OV), Stomach Cancer (STOMACH), Colon Cancer (COAD), Kidney Cancer (KIRC), Liver Cancer (LIHC), Lung Cancer (LUSC), Prostate Cancer (PRAD), and Thyroid cancer (THCA). The classification was based on DNA methylation data obtained from the TCGA database. The system that was developed also shown exceptional performance in terms of receiver operating characteristic area under the curve (ROC AUC) values, ranging from 0.85 to 0.89 [129]. A few studies have attempted to uncover DNA methylation indicators that can be used to diagnose various cancer types using deep learning techniques such as MethylNet [130], MRCNN [131], deep neural network (DNN) [132], and deep autoencoder [133]. DeepCpG is another CNN-based approach for predicting methylation states and has accurately identified the changes in methylation levels [134]. A DNA methylation-based cancer classification tool, MethPed [135] was developed for pediatric brain tumors. The present methylome data

consists of more than 800000 CpG sites; therefore, extracting the relevant features is challenging. Further, exploration of DL algorithms and methylation data may contribute to understanding the complex mechanism of gene regulation to identify the brain cancer-specific markers. The above literature showed that DNA methylation data have a potential to serve as a biomarker for several cancers.

It is observed that artificial intelligence techniques are used for cancer classification using genomics data. Apart from cancer classification, molecular subtyping of cancer is also important step towards the personalized therapy.

2.6 AI in cancer subtype classification

The precise identification of the specific subtype of cancer is of utmost importance in order to get an accurate diagnosis and effective therapy for patients. This is because the cancer subtype plays a critical role in improving clinical outcomes. Many human cancers have multiple subtypes with unique molecular signatures, and these subtypes also show different prognosis and treatment responses. Choi, J.M., et al. 2023 proposed a semi-supervised method for classifying breast cancer subtypes using DNA methylation profiles. The accuracy of the subtype classification was determined to be 82.3% [136]. Yuan, F., et al, 2020 applied ML algorithms such as SVM and RF on lung cancer data to classify the subtypes of lung cancer into Lung adenocarcinoma (LUAD) and lung squamous cell cancer (LUSC) by employing the gene expression profiles. They observed that SVM outperformed RF to classify the lung cancer subtypes; and achieved classification accuracy of 96.7% [137]. Similarly, DL-based methods were employed for the classification of lung cancer subtypes. For example, XGBoost algorithm was used to classify the subtypes of lung cancer into LUAD and LUSC using gene expression data. The models showed excellent subtype classification accuracy of 97.1% [138]. Tao, M., et al, 2019 applied Multiple Kernel Learning (MKL) on breast cancer to classify the breast cancer subtypes using gene expression, methylation and copy number data. They obtained a classification accuracy of 79.8% [139]. Shen, J., 2022 introduced a novel methodology that integrates a convolutional neural network (CNN) with a bidirectional gated recurrent unit (BiGRU) as a deep learning strategy named DCGN. This approach was used to classify the cancer subtypes of breast cancer and bladder cancer using high dimensional gene expression data. They compared the DCGN performance with seven other methods and DCGN outperformed among

all. The DCGN showed the subtype classification accuracy in breast cancer was 96% and in bladder cancer 95.5% [140].

It is observed that there are only few reports are available for subtype classification of cancers using genomics data. Based on whole genomics data, cancer subtyping studies is carried out and demonstrated that it is efficient approach for dissecting cancer heretogeneity. The advent and swift progress of high-throughput sequencing technologies, including next-generation sequencing technology, RNA sequencing (RNA-seq), DNA methylation arrays, and a lots of mutations in entire genome, have facilitated the exploration of disease mechanisms at the genome, transcriptome, epigenome and mutational levels. However, the use of single omic data is limited to examining only one component of omics data, and it lacks the ability to elucidate the intricate relationships among genetic alterations, such as mutations, gene expression, and methylation. On the other hand, the integration of multi-omics data from different genomic levels provides a more extensive comprehension of intricate disease modifications and helps to understand the cancer initiation, facilitating cancer detection, and improved therapy strategies.

2.7 Integration of genomic data for cancer classification

To predict cancer from single-omics data such as genome, transcriptome, methylome or mutational data are widely used. However, these single layers of genomics data individually do not explain every aspect of cancer. Although integration of all genomics layer interaction collectively explains the complex relationships between molecular layers that leads to cancer. With only a single type of omics data, tumour occurrence and development cannot be effectively predicted. Accurate multi-omics integration techniques are required to combine data from diverse patients because multi-omics data typically come from entirely distinct sources. Consequently, a pressing issue in precision medicine is how to rationally integrate the current chaotic multi-group data to increase the accuracy of disease diagnosis.

Genomics data from different molecular levels are linked to one another. Such as, mutation changes the mRNA expression level of the genes or methylation level in the promoter region determines the depleted or elevated expression of the genes. Therefore, recently researchers have tried to integrate multiple genomics data or multi-omics data to develop powerful ML and DL-based tools. Moreover, to capture crucial cellular mechanisms or interactions between

biomolecules, it is essential to analyze the multi-omics data, which can facilitate the discovery of the new diagnosis and therapeutic approach for cancer treatment. In order to find the new patterns in cancer patients, multi-omics data was used by employing machine learning methods [141]. Many multi-omics integration studies for various cancers have been conducted in recent years. However, in the field of diagnosis of brain cancer, there are very few studies on multi-omics integration. Yang et al. (2019) proposed a novel approach for data integration and cancer subtyping, specifically targeting seven forms of cancer, including GBM. The approach is based on a Random Walk based cluster ensemble (RWCE) method that incorporates mRNA, miRNA, and methylation data. This study provides evidence that it possesses the capacity to identify subtypes that hold clinical and biological relevance [142].

Recently a deep neural network (DNN) learning model was proposed to effectively integrate the omics datasets of copy number alteration and gene expression data. The objective of this integration was to accurately predict the molecular subtypes of breast cancer. The researchers showed that an integrative deep learning model provided good prediction accuracy. The study demonstrated an accuracy rate of 79.2% [143]. Furthermore, another multimodal DL tool, MultiSurv, was designed to estimate the long-term survival prediction of cancer patients [144]. MultiSurv integrates clinical, imaging, and multi-omics data (mRNA, miRNA, DNA methylation, CNA data) to predict patient survival with high accuracy. Zhang et al. designed a multi-view multi-task deep learning framework, OmiEmbed, to integrate the high dimensional multi-omics data. OmiEmbed can be used for demographic and clinical feature reconstruction and survival prediction [145]. The authors also explained that OmiEmbed could facilitate accurate and personalized treatment for cancer. This evidence shows that implementing a deep learning-based framework for integrating and analyzing the various omics data could revolutionize the clinical diagnosis of cancer.

Autoencoders are used for data integration of multi-omics data to identify disease states and cancer subtyping. Subtype identification is a challenging task, therefore identifying the particular patient subgroup necessitates the integration of multi-omics data. Two patient subgroups with significant survival differences have been found using supervised and unsupervised learning on transcriptomics, and DNA methylation data of hepatocellular carcinoma (HCC) [146]. Xu et al, 2019 proposed a hierarchical integration approach called HI-DFNForest, which utilises deep flexible neural forest data to effectively integrate multi-omics data for the purpose of cancer subtype classification. The researchers employed a Stacked

Autoencoder (SAE) technique to extract meaningful features, followed by the utilisation of a Deep Flexible Neural Forest (DFNForest) model for the classification of patients into breast cancer subtypes using data sets obtained from TCGA. This integration involved the incorporation of gene expression, miRNA expression, and DNA methylation data. The integration of multi-omics data of breast cancer demonstrated favourable predictive accuracy, with percentages of 84.6% [147].

Artificial intelligence can effectively manage high-dimensional genome-wide data and discern concealed patterns that may not be noticeable in individual genetic data. This integration aims to convert large datasets into clinically actionable knowledge, thereby serving as the basis for precision medicine.

2.8 Artificial intelligence and personalized medicine

Precision medicine is a method to develop personalized care for patients based on an individual patient's molecular profile. The approaches in precision medicine are designed to investigate the relationship between genomic alteration and its contribution to the risk of developing specific cancer or its effect on treatment. Inter- and intra-tumor heterogeneity causes the genotypic differences between patients, showing the necessity of personalized medicine for effective treatment. Due to the abundance of available data and high-throughput experimental techniques, DL can revolutionize decision support systems in oncology and decipher the hidden phenotypic and genotype patterns, as well as their correlations.

The deepProfile, a DL-based framework uses the unlabeled gene expression data to predict the complex disease phenotype. deepProfile can be implemented on gene expression data from brain cancer to find the phenotype-genotype relationship for personalized treatment [148]. The differences in Drug response occur due to inter-and intra-tumor heterogeneity. A deep variational autoencoder (VAE) model was demonstrated to predict the accurate drug response with higher efficiency with these heterogeneous data. In addition, the authors identified molecular features associated with drug response in 33 cancer types, including the brain [149]. Identifying the genotype to phenotype relationship is a crucial step to finding the molecular signature of the disease. GenNet, a deep learning framework, can predict phenotypes from genetic variants [150]. Another biologically interpretable tool Varmole [151] embeds multi-omics networks data into a deep neural network framework and prioritizes variants, genes, and

regulatory linkages, subsequently predicting genotype to phenotype relationships. For complex diseases like brain cancer, personalized medicine based on individual molecular signatures is essential for targeted therapy. Furthermore, to avoid the adverse effects of drugs and to increase the life expectancy of brain cancer patients, a DL-based support system will be most desirable in modern medicine. The utilisation of artificial intelligence (AI) presents an opportunity to leverage genomic information across many molecular layers. This has the potential to facilitate prognostic predictions regarding patient outcomes, including the probability of a positive response to a cancer treatment intervention.

2.9 Integration of multi-omics data for Drug Development

The integration of multi-omics data, which encompasses information on biomolecules from several levels, has great promise in facilitating a comprehensive and systematic understanding of complicated biological processes. Argelaguet, R., et al., 2018 introduced a computational approach called Multi-Omics Factor Analysis (MOFA) to identify hidden components within a multi-omics dataset that capture both biological and technical sources of variability [152]. Integrated approaches aid in the evaluation of the transfer of information between different omics levels, hence facilitating the connection between genotype and phenotype. There are exists substantial evidence indicating that modifications in the genomes of cancer cells can significantly impact the efficacy of anticancer treatments in clinical settings. There are various cases in which genetic variations have been utilised as molecular biomarkers to identify individuals who are most likely to derive therapeutic advantages from a specific treatment. The utilisation of integrative analyses that effectively synthesise and establish connections between molecular data and treatment sensitivity is of utmost importance in order to comprehensively capture the intricate biological complexity that underlies precision medicine. The primary objective of precision medicine is to administer the right drug to the right patient at the right time. Different patients respond differently for the same drug due to intratumor heterogeneity. The main challenge in the field of oncology research is the prediction of individual response for different treatments. To overcome this problem, AI based techniques are widely used. Artificial intelligence-based discovery has gained attention recently since it drastically cuts the time and money needed to produce novel drugs. To identify drug response on cancer to therapies based on molecular profiles of multi-omics data, deep learning models can be used, this can lead to profiling of the modern era of precision medicine and yield the clinical

relevance. Chiu et al. 2019 introduced a pair of deep neural networks, wherein one network was designed to handle gene expression data and the other network was tailored for gene mutation data. Subsequently, the two networks were integrated to collectively forecast drug response [153]. Wang, C., et al., 2021 developed deep neural network architecture to integrate the multi-omics data encompassing gene expressions, copy number variations, gene mutations, reverse phase protein array expressions, and metabolomics expressions from cancer cell lines data available in CCLE and GDSC. They employed a graph embedding layer to incorporate the interactome data and attention layer to combine different omics features and achieved the drug response prediction accuracy was 98% [154]. Almutiri et al. (2023) proposed a novel methodology that integrates Bayesian Ridge Regression (BRR) with Deep Forest. The BRR method was employed for the purpose of integrating several omics datasets, while the DF approach was utilised for drug response prediction. The Cancer Cell Line Encyclopaedia (CCLE) dataset was utilised to integrate gene expression, copy number variation, and single nucleotide variation. The evaluation criteria employed in this study included Root Mean Square Error (RMSE), Pearson Correlation Coefficient (PCC), and the coefficient of determination (R2). The model obtained an RMSE value of 0.175, a PCC value of 0.842, and an R2 value of 0.708 [155]. Malik et al. (2021) introduced a comprehensive methodology that integrates multi-omics data including copy number variation (CNV), mutation, methylation, miRNA, RNA, and protein expression. To accurately assess the survival outcomes and medication responsiveness in individuals diagnosed with breast cancer. The Neighbourhood Component Analysis (NCA) algorithm, which is a supervised feature selection method, was utilised to identify pertinent features from multi-omics datasets obtained from The Cancer Genome Atlas (TCGA) and Genomics of Drug Sensitivity in Cancer (GDSC) databases. The survival prediction framework shown a high level of effectiveness in classifying patients into risk subtypes, with an accuracy rate of 94% [156]. Recently several other are also avalaible such as MOLI [157], AGMI [158], and DrDimont [159] for drug response prediction. These computational models of drug sensitivity prediction help to aid in the selection and prioritization of candidate compounds for pre-clinical research.

It has been observed that none of the reports presented thus far provide information on the subtype-specific drugs for the treatment of glioma. Through the integration of multi-omics data with pre-existing knowledge of molecular interactions, artificial intelligence (AI) has the capability to identify potential drug targets that play a critical role in the advancement of cancer and can be potentially influenced by therapeutic treatments. For the advancement of precision

medicine and the development of tailored cancer treatment, individualized medication response prediction is essential. Large-scale multi-omics profiles provide unprecedented opportunities for precision cancer therapy.

Lacunas

Brain cancer classification was done by histopathological methods that suffer from intraobserver and interobserver variability [6], which causes poor clinical outcomes. Introducing genetic markers (WHO classification, 2016), such as a mutation in either the Isocitrate dehydrogenase IDH1 or IDH2 gene and co-deletion of 1p and 19q chromosomes, provides more persistent diagnosis options and better clinical management [160]. IDH1 mutation and 1p/19q co-deletion are diagnostically and prognostically significant. However, this may not always provide accurate classification, as IDH mutation was reported in all types of LGGs [161]. Therefore, 1p/19q co-deletion testing may lead to a false positive (FP) result [162]. Alternatively, imaging techniques, such as magnetic resonance spectroscopy and positron emission tomography, are used for grading the LGG [163]. However, these techniques do not provide the genetic basis of cancer grade. Consequently, several studies suggest the requirement of additional clinical variables to increase the sensitivity of current treatment [164]. Most of the research papers have reported AI–based binary classification methods to classify cancer and healthy samples using image data or mono omics data (i.e., gene expression). None of the reports shows the subtyping of brain cancer with multi-omics data till now. As previously mentioned, LGG can be classified into three subtypes: astrocytoma, oligoastrocytoma, and oligodendroglioma, of grade 2 and grade 3. Similary, GBM has three distinct subtypes, namely classical, mesenchymal, and proneural. To yet, the classification of grade and subtype based on genomes data has not been undertaken. To date, there has been a lack of grade and subtype-specific classification utilising genomes data. Most of the biomarkers for brain cancer until this point have been discovered using inconclusive, low-throughput techniques without taking omics data into account. Most observations from low-throughput experiments fail to provide therapeutic solutions since they cannot provide a comprehensive perspective of the complex systems of cancer. With the aim of resolving this issue, omics data analysis is a rapidly growing area of research to effectively capture complex relationships from multiple omics layers, i.e., genomics, transcriptomics, and epigenomics. Because of the advancement of high-throughput technologies, the size of these omics data is increasing exponentially. In order to extract new insights from vast amounts of data, powerful computing approaches are needed. In this context,

machine learning (ML) and deep learning (DL) algorithms have emerged as one of the most successful techniques because of their capacity for handling high dimensional data, effective data integration, efficient dimensionality reduction, stability, and higher prediction accuracy. While various techniques exist for integrating multi-omics data, such as the utilisation of autoencoders, there is still a lack of research demonstrating the categorization of glioma subtypes using multi-omics data. Numerous machine learning (ML) models have been developed for the purpose of cancer classification, exhibiting superior accuracy. However, these models are limited in their ability to identify the cancer causing genes that trigger the tumorigenesis process. Integrating the methylome and transcriptome is crucial in finding the genetic and epigenetic features that cause cancer, which is also important for making biologically relevant models. Univariate cox analysis facilitates identifying the biologically important and cancer-associated features, which can lead to the development of a clinically relevant DL model. Another relevant task for precision medicine is to find the targets and develop a drug response prediction model for cancer subtypes as we know that every individual has different genetic makeup. They respond differently to the same drug and same tumor type due to their inter and intra-tumor heterogeneity. Considering the complexity of glioma, network medicine-based approaches should be implemented to find the subtype-specific drug targets. Advances in sequencing techniques and genome-wide association studies have revealed that accumulated genetic variations associated with an increased risk for cancer are distributed throughout the genome. Further studies illustrate that genes affected by genomic variations are not randomly distributed in molecular networks. Indeed, genes associated with the same disease are more likely to interact with each other. As a result, a disease module forms, a subnetwork linked to a disease. Numerous genes that are known to be relevant to disease are found in disease modules. Utilizing disease modules in each subtype of glioma, subtype-specific target can be identified. Identification of cancer-specific disease modules can help to identify novel biomarkers for therapeutic targets. Therefore, network medicine and rational drug-designing approaches recognize these modules as pharmacological targets as opposed to the individual genes or proteins in the network. However, the therapeutic efficiency of drugs in cancer is highly context-dependent; often, drug resistance reduces the effectiveness of chemotherapy. Therefore, the prediction of drug response, i.e., resistance or sensitivity, is essential for improving the efficacy of chemotherapy. Therefore, AI-based drug response prediction models can be developed using genomics data for precision therapy. Based on the present lacuna, the

present thesis develops AI-based models to support the clinical diagnosis of glioma subtypes and drug response prediction.

Objectives

The objective of the work is to develop a machine-learning and deep-learning based framework for subtype classification of glioma and identify the biomarkers in each subtype of glioma. AI-based diagnostic tool and drug response prediction model for the precision therapy of glioma is developed. This thesis focuses on the following objectives.

1. Development of a machine learning-based framework for subtyping and grading of lower-grade glioma (LGG) using transcriptome data and identification of biomarkers.
2. Development of Deep learning and machine learning framework based on genomic data for subtyping the glioblastoma multiforme (GBM) and identification of biomarkers.
3. Implementation of Deep learning embedding system for multi-omics data integration for subtyping of Glioma.
4. Identification of subtype-specific disease modules and development of drug response prediction models by combining network medicine and AI-based approaches.

CHAPTER 3

OBJECTIVE 1

Chapter 3: Objective 1
Development of a machine learning-based framework for subtyping and grading of lower-grade glioma (LGG) using transcriptome data and the identification of biomarkers

3.1 Introduction

Classifying lower-grade gliomas (LGG) is a crucial step for accurate therapeutic intervention. The histopathological classification of various subtypes of LGG, including astrocytoma, oligodendroglioma, and oligoastrocytoma, suffers from intraobserver and interobserver variability leading to inaccurate classification and greater risk to patient health. The accurate classification of glioma types and grades is vital to improving brain cancer patient's prognosis. Due to the enormous complexity at the molecular level, the critical molecular driver of gliomas is poorly understood. Therefore, there is an urgent need to identify the subtype-specific molecular marker for personalized therapy. Over the past decade, advances in sequencing technology have provided the opportunity to understand complex disorders holistically and have contributed to designing effective therapeutic approaches. RNA sequencing technology provided the opportunity to study genome-wide expression patterns. Changes in gene expression patterns are a prominent feature of any cancer cell, which have been successfully implemented to explain the mechanism of cancer. However, whole-genome expression data or transcriptomes are barely used to classify the brain cancer type and grade. In this chapter, a comprehensive analysis was performed to develop an interpretable machine learning (ML) framework using the transcriptome data of LGG to diagnose subtypes and grades.

To develop a model for subtype and grade classification, both unsupervised and supervised learning techniques was applied. However, unsupervised methods were unable to separate the subtypes and grades. Therefore an ML framework was developed based on supervised learning techniques. In brief, correlation, support vector machine recursive feature elimination (SVM-RFE), and Boruta algorithm was implemented for feature selection. Subsequently, the classification using Support Vector Machine (SVM), k-nearest neighbors (KNN), Gaussian Naïve Bayes (GNB), Decision Tree (DT), and Random forest (RF) was performed and compared their performance. Most published reports focus on the two-class normal vs. cancer cell classification. In comparison, our approach has efficiently categorized multiple classes, i.e.,

astrocytoma, oligodendroglioma, and oligoastrocytoma cells, including grades. Further to find the biological relevancy of feature genes, the gene expression pattern among the LGG of different subtypes and grades was compared using co-expression analysis. Additionally, subtype-specific prognostic markers for diagnosis and treatment was identified.

3.2 Methodology

3.2.1 Data Collection and Balancing of the Dataset

Healthy brain tissue (n = 93) gene expression data of GTEx and mRNA expression data of LGG (n = 281) patients were obtained from UCSC Xena [165] (https://xena.ucsc.edu/). Based on the clinical information, LGG samples were divided into specific subtypes and grades (Table 3.1 and results section, Figure 3.2 A). The external data set (GSE74462 and GSE43378) was collected from the Gene Expression Omnibus (GEO) repository for validation. In table 3.1, there are unequal number of samples in different subtypes and grades of LGG; due to unequal number of samples the model can become biased towards one class, leading to poor performance of the model. Hence, random sampling was performed to select the equal number of patients in each subtype before the feature selection. The oversampling technique was used to reduce the bias and variance of the classifier. Dataset balancing was done using the *imbalanced-learn* package in Python, and the minority class was randomly oversampled to obtain a balanced dataset. The oversampling technique was used to reduce the bias and variance of the classifier.

Table 3.1: Details the tumor, normal samples in LGG

	Transcriptome		
Type	**Grades**	**Subtypes**	**Samples**
LGG	Grade II	Astrocytoma (G II)	30
		Oligoastrocytoma (G II)	42
		Oligodendroglioma (G II)	67
	Grade III	Astrocytoma (G III)	66
		Oligoastrocytoma (G III)	33
		Oligodendroglioma (G III)	43
Healthy	Normal	—	93

3.2.2 Principal Component Analysis

Principal Component Analysis (PCA) is a statistical technique used for analysing high dimensional datasets containing a high number of features per instance helps to preserve the maximum amount of information while converting into high dimensional space to low dimensional space. PCA is widely used for reducing the dimension of the features and visualization. In PCA, data is linearly transformed into new coordinates where most of the variation in the data can be described in fewer dimensions than the initial input dataset. Most of the studies use the first two principal components because it explains most variance; in order to plot the two dimension data for the visualization of data points that are clustered and closely related. PCA is used when many variables are highly correlated, and it is desirable to reduce the dimension of the variable into independent variables. Further, these independent variables are taken for making predictive models [166]. The principal component analysis (PCA) was used to observe the gene expression patterns in different subtypes and grades of LGG. Principal components analysis (PCA) of the gene expression data of LGG was performed using the ggfortify package in R. PCA was done on scaled data. A cancer subtype-wise cluster was generated using the cluster package in R.

3.2.3 Correlation-based Feature Selection

The feature subset with low feature-feature correlation avoids redundancy. The feature sets with high predictive power contain highly correlated features with the class but are uncorrelated with each other. Genes with the same expression pattern in different subtypes are highly correlated and redundant because they cannot distinguish different classes. Therefore, highly correlated genes between the subtypes were removed to improve classification accuracy. We measured the correlation coefficient between the gene expression values in the different class labels (subtypes) separately for grade 2, grade 3, and mixed grade. In this approach, the correlation was measured on gene expression data using the Pearson correlation coefficient (PCC). The correlation coefficient is a statistical metric that quantifies the magnitude and direction of the association between two variables, with values ranging from -1 to 1. The correlation was calculated using the NumPy package in Python using formula shown in the below. PCC > 0.7 is set as the threshold, and genes that had PCC > 0.7 in between the classes were dropped, and the remaining features were taken for model development.

$$r = \frac{n(\sum xy) - (\sum x)\,(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2]\,[n\sum y^2 - (\sum y)^2]}}$$

Where,

r = Pearson Coefficient

n= number of pairs of genes

∑xy = sum of products of the paired genes

∑x = sum of the x scores

∑y= sum of the y scores

$\sum x^2$= sum of the squared x scores

$\sum y^2$= sum of the squared y scores

3.2.4 Machine Learning-based Feature Selection

We have applied supervised machine learning-based feature selection methods, support vector machine recursive feature elimination (SVM-RFE), and Boruta. These algorithms were used to find the features that optimized the classifier's performance.

3.2.4.1 Support Vector Machine Recursive Feature Elimination (SVM-RFE)

Support vector machine recursive feature elimination (SVM-RFE) is a supervised machine learning-based feature selection method [167]. It is a potent feature selection algorithm. Avoiding overfitting. The aim of recursive feature elimination (RFE) is to select features by recursively considering smaller and smaller sets of features. In brief, SVM-RFE initializes the data set for all features and trains the SVM using the dataset, and then it ranks the features. Feature selection is done only on the training dataset by use of SVM-RFE. SVM-RFE deletes features having the minimum weight to obtain the optimum rank list of the features. Next based on ranking it screens the optimum features and eliminates the lower-ranked features. The process of selecting feature sets for SVM-RFE may be broken down primarily into three steps: (1) input of the datasets to be classified; (2) computation of each feature's weight; and (3) deletion of the feature with the lowest weight to determine the ranking of features. This way, recursively deleting the least important features from the ranked list and selecting the optimum gene set improved

classification accuracy [168,169]. SVM-RFE was used to find the features that optimized the classifier's performance. SVM-RFE was implemented using the *Scikit-Learn (*https://scikit-learn.org) package in Python. After selecting the features by RFE, the best features were identified by rank, and features were selecting according to the highest rank. The top-rank features (20, 50, 100, 200, and 500) were selected as variables for classifications.

3.2.4.2 Boruta

The Boruta algorithm iteratively removes the statistically less relevant features than the shuffled copies of the features. This algorithm selects the important features by comparing the Z-scores of the shuffled features with the original features [170]. We have used the Boruta algorithm for feature selection using the Boruta package in R.

3.2.5 Machine Learning Algorithms

Supervised machine learning is a type of machine learning algorithm that needs the labeled dataset to train the algorithms that to classify the data and predict the outcomes accurately for unforeseen data. It means a small set of data is already tagged with the correct label. For example, ‘cancer’, and ‘non-cancer’ for a cancer classification problem. Classification-based supervised learning methods are probability-based and find the category of outcome (discrete values); the algorithm finds the highest probability of a set of data items belonging to. In the classification approach, discrete values of a particular class are predicted and evaluated based on the accuracy of the model. This is either binary classification or multiclass classification. In binary classification, the model either predicts cancer and normal (0, or 1), whereas in multiclass classification model predicts more than one class, for example, cancer subtypes astrocytoma, oligoastrocytoma and oligodendroglioma (0, 1, and 2). Here, in this thesis, we classified the subtypes of LGG using several ML algorithms. We used the *sklearn* library in Python to build the ML models. Supervised ML methods which are implemented in the present work are Support Vector Machine (SVM), k-nearest neighbors (kNN), Gaussian naïve bayes (GNB), Decision Tree (DT), and Random Forest (RF). A brief description of each algorithm is provided here.

3.2.5.1 Support Vector Machine (SVM)

Support Vector machine (SVM) is a popular supervised machine learning algorithm, which is used for classification problems. This method is based on statistical learning theory. The main objective of the SVM algorithm is to maximize the margin or to create the best linear decision boundary that can segregate the n-dimensional space into the classes of data points on either side of the decision boundary, and the best decision boundary is called a hyperplane [171]. The training samples that are close to the hyperplane are called support vectors. The margin is calculated as the perpendicular distance from the line to the closest point. Therefore, SVM computes the maximum boundary that leads to a uniform split of all data points. If a dataset is noisy and messy, then it cannot be separated with a hyperplane. In some cases, a hyperplane or linear decision boundary cannot be found, and a kernel is used. SVM uses support vectors that separate data points in different hyperplanes [74]. SVM selects optimal hyperplanes for classification. In SVM, we tuned the hyperparameter, i.e., regularization parameter c (c = 10), and applied a linear kernel to achieve higher accuracy. SVM was implemented using the *SVC* package in Python.

3.2.5.2 k-nearest neighbors (kNN)

K-nearest neighbors (KNN) is one of the simplest supervised machine learning algorithms considered a lazy learner (it does not learn from the training set immediately; instead, it stores the dataset, and at the time of classification, it performs an action on the dataset) as there is no learning is required in the model. It is a non-parametric (it does not make any assumption on underlying data) algorithm that categorizes data points based on their proximity and association to other available data. This algorithm assumes that similar data points are nearby. As a result, Euclidean distance (which is calculated as the square root of the sum of the squared differences between a point a and b across all input attributes i, and which is represented as $d\,(a,b) = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2}$. Euclidean distance is a good distance measure to use if the input variables are similar in type). It is usually used to calculate the distance between data points and assign categories based on the most common category or average. For new data points, predictions are made by finding the K most similar instances (neighbors) across the training set and summarizing the output variables for those K instances [75,76]. KNN was implemented using the *KNeighborsClassifier* package in Python.

The following steps are to be followed in KNN:

1. Select the number of K of the neighbors.
2. Calculate the Euclidean distance of k number of neighbors of the sample that have to classify.
3. Among the k neighbors, count the number of data points in each category.
4. Assign the new data point to the class label for maximum number of neighbors.
5. The model is ready.

3.2.5.3 Gaussian naïve bayes (GNB)

Naïve bayes is a probabilistic classification approach based on Bayes theorem and used for solving classification problems [172]. It is a simple and effective classification algorithm to build the model for large datasets and make quick predictions, but it has high functionality. First, it is called Naïve because it assumes that a certain feature is independent of the occurrence of other features. Second, it is called Bayes because it depends on the principle of Bayes' Theorem (Conditional and Joint Probability). This implies that each predictor has an equivalent influence on the outcome, and the presence of one feature does not influence the presence of another in determining the probability of a specific event.

Here, Gaussian naïve bayes (GNB) are used for classification of LGG subtype. GNB is a generative model. It is an approach to create a simple model to assume that each datapoint follow the Gaussian distribution with no co-variance (independent dimensions) between dimensions. This model can be fit by simply finding the mean and standard deviation of the points within each label. This is all it takes to define such a distribution. This is how the GNB classifier works. Gaussian Naive Bayes is a probabilistic machine learning classifier based on the Bayes theorem. It assumes that the data from each label is drawn from a simple Gaussian distribution and considers all the features are independent [173]. GNB was implemented using the *GaussianNB* package in Python.

3.2.5.4 Decision Tree (DT)

Decision Trees (DT) are the supervised machine learning algorithm preferred to solve classification problems and predictions. It is a hierarchical tree-like structure where internal nodes denote a test on the features, branches represent the decision rules (outcome of the test),

and each leaf node holds the class label. It boosts the predictive model accuracy and ease in interpretation and stability. In Decision Tree, the main aim is to create a model that predicts the value of a target variable by learning simple decision rules. The decision tree is constructed by repeatedly splitting a node into two child nodes, beginning from the root node containing the whole learning sample. Decision tree learning follows a divide-and-conquer strategy by performing a greedy search to identify the optimal split points in the tree [174]. The process of splitting the nodes is repeated in a top-down manner until all data are classified into particular class labels homogeneously. DT was implemented using the *DecisionTreeClassifier* package in Python.

3.2.5.5 Random Forest (RF)

Random forest (RF) is a very popular and fast supervised machine learning algorithm. It is used for classification problems. Random forest is an ensemble technique used for classification by several estimators (decision trees). Classification is done using the majority vote among estimators. Random forest prevents overfitting and leads to higher accuracy because it contains many trees, leading to a more generalizable model. It is diverse, and more stable models are formed [175]. RF was implemented using the *RandomForestClassifier* package in Python.

3.2.5.6 K-fold Cross Validation

When enough validation set is not available to tune the hyperparameter, this k-fold cross-validation technique is used. A validation and test set may be burdensome when you have few training samples. You would instead train the model with more data. Then only divide the data into a training set and a test set. After that, mimic a validation set on the training set using cross-validation. K-fold cross-validation is the randomized subset of data. Here, stratified k-fold cross-validation (CV) was applied to each model's training dataset (70%). Stratified k-fold cross-validation was employed in cases when the dataset size is small. It described the reliability and stability of the models. The stratified k-fold data is split into k equal parts, where k-1 is used to train a model, and the remaining portion is used as a test data to evaluate the model's performance. This is an iterative process repeated up to k times. The final output is then computed by averaging over the obtained performance parameters from each test set. Figure 3.1 shows an example of 10-fold cross-validation. The calculation of the standard deviation was performed on the mean accuracy in order to get the error rate for classification. Additionally, a

statistical test was conducted using pairwise comparison for each machine learning method to assess the average accuracy scores obtained from the 10-fold cross-validation.

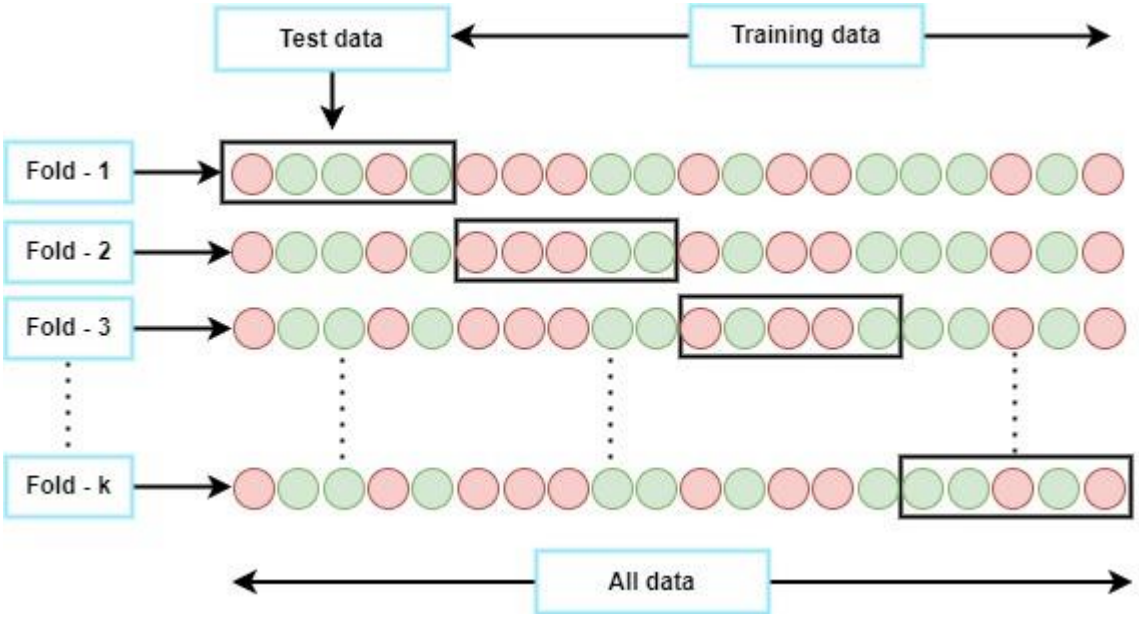

Figure 3.1: *Demonstration of k-fold cross-validation.*

3.2.6 Model Evaluation Metrics for Classification

Evaluating the performance of a model is one of the important tasks of machine learning. Before creating a model, datasets are divided into training and test datasets, and the model trained on the training datasets and model evaluation is done on the test datasets with labels. In which the predicted label is compared with the actual label, and measure the performance of the model using several evaluation metrics. Creating a precise model that can forecast previously unobserved data is essential. The model may perform exceptionally well on some criteria while doing poorly on others. Therefore, it is crucial to analyze the model using a variety of measures. Once the model is prepared, it should undergo an evaluation to assess its performance.

3.2.6.1 Confusion Matrix

A confusion matrix is in the form of a contingency table or matrix of output and describes the performance of the model. Sometimes, it is known as an error matrix. The matrix consists

of prediction results in summarized form, showing the total number of correct and incorrect predictions. The description of a matrix is given below:

	Actual positive	**Actual negative**
Predicted positive	**TP**	**FP**
Predicted negative	**FN**	**TN**

True positive (TP) — the number of samples with the absence of brain cancer predicted as an absence of brain cancer.

False positive (FP) — the number of samples with the presence of brain cancer predicted as an absence of brain cancer.

True negative (TN) — the number of samples with the presence of brain cancer predicted as the presence of brain cancer.

False negative (FN) — the number of samples that have an absence of brain cancer predicted as the presence of brain cancer.

3.2.6.2 Computation of Performance Measures

From the contingency table, several parameters are selected, and the performance of ML models was evaluated using accuracy, specificity, precision, sensitivity (recall), F1-score, Geometric mean (GM), and Matthews correlation coefficient (MCC). At first, a confusion matrix was generated to compute these performance scores. True positive (TP), true negative (TN), false positive (FP), and false negative (FN) was calculated from the confusion matrix. Then we calculated the accuracy or success rate as follows,

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

The sensitivity or true positive rate of a ML model was measured using the following equation.

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

The specificity or true negative rate of a ML model was measured using the following equation.

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

The precision or positive predicted value was measured using the following equation.

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

A measure of model performance that combines precision and recall into a single number is known as the F-measure or F1-score. The following equation was used to compute the F1-score.

$$F1 - score = \frac{TP}{TP + \frac{1}{2}(FP + FN)}$$

Geometric mean (GM) is the average value or mean, which signifies the central tendency of the set of numbers by taking the nth root of the product of their values. The higher value of GM indicates better balance classification.

$$Geometric\ mean(GM) = (x_1.x_2 \ldots\ldots.x_n)^{1/n}$$

Matthews correlation coefficient MCC measures the correlation of the true classes with the predicted labels.

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

We used *sklearn.metrics* library in Python to calculate the above measures.

AUC-ROC curve

ROC curve stands for Receiver Operating Characteristics Curve [176] and AUC stands for Area Under the Curve. It is a graph that shows the performance of the classification model, the probability of true positive results against the probability of false positive results for a range of different cut-off points. The formula of TPR and FPR are given below:

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$

AUC-ROC curve is used for the visualization of the model. The higher the area under the ROC curve, the model will be better. A classifier that exhibits superior performance is characterized by AUC value that surpasses 0.5. The AUC of a perfect classifier would be 1. Typically, assuming model performance, an effective classifier can be obtained by selecting a threshold value that maintains a low false positive rate (FPR) and a high true positive rate (TPR). Further, we also visualized the model performance across a wide range of conditions using receiver operating characteristic (ROC) plots.

3.2.7 Ranking of the Models

Multiple Criteria Decision Making (MCDM) [177] was used to select the best model. MCDM was implemented in Python, and the Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) was used to find the rank. The performance parameters, i.e., accuracy, recall, precision, F1-score, GM, and MCC, were used for ranking.

3.2.8 Survival Analysis

Survival analysis of feature genes having correlation > 0.5 ($p < 0.05$) of LGG patient samples was conducted using Gene Expression Profiling Interactive Analysis (GEPIA) databases (http://gepia.cancer-pku.cn/). GEPIA utilises the log-rank test to perform survival analysis by employing TCGA clinical data. The overall survival of feature genes was generated based on

the high and low expression of genes. The cut-off value set as quartile (upper-quartile = 75% and lower-quartile = 25%) and $p < 0.05$ was considered as statistically significant.

3.2.9 Biological Pathway and Process Enrichment Analysis

Pathway and process enrichment analysis was carried out using the Metascape tool with the following ontology sources: GO Biological Processes, KEGG Pathway, and Reactome Gene Sets [178]. Metascape combines functional enrichment, interactome analysis, and gene annotation. If the adjusted p-value < 0.05, the biological process or pathway was considered significantly enriched.

3.2.10 Statistical Analysis

One-way ANOVA followed by a post-hoc Tukey-HSD test was performed using Sigma Plot 11.0. A hypergeometric test was conducted using R.

3.3 Results

3.3.1 Development of Machine Learning-based Classifier for Diagnosis of the LGG Subtypes

Genome-wide mRNA expression data of LGG (n = 281) patients were obtained from UCSC Xena (https://xena.ucsc.edu/). Based on clinical information, LGG patients were segregated into grade 2 and grade 3 of astrocytoma (n = 30 and 66), and oligoastrocytoma (n = 42 and 33) oligodendroglioma (n = 67 and 43) (Figure 3.2A).

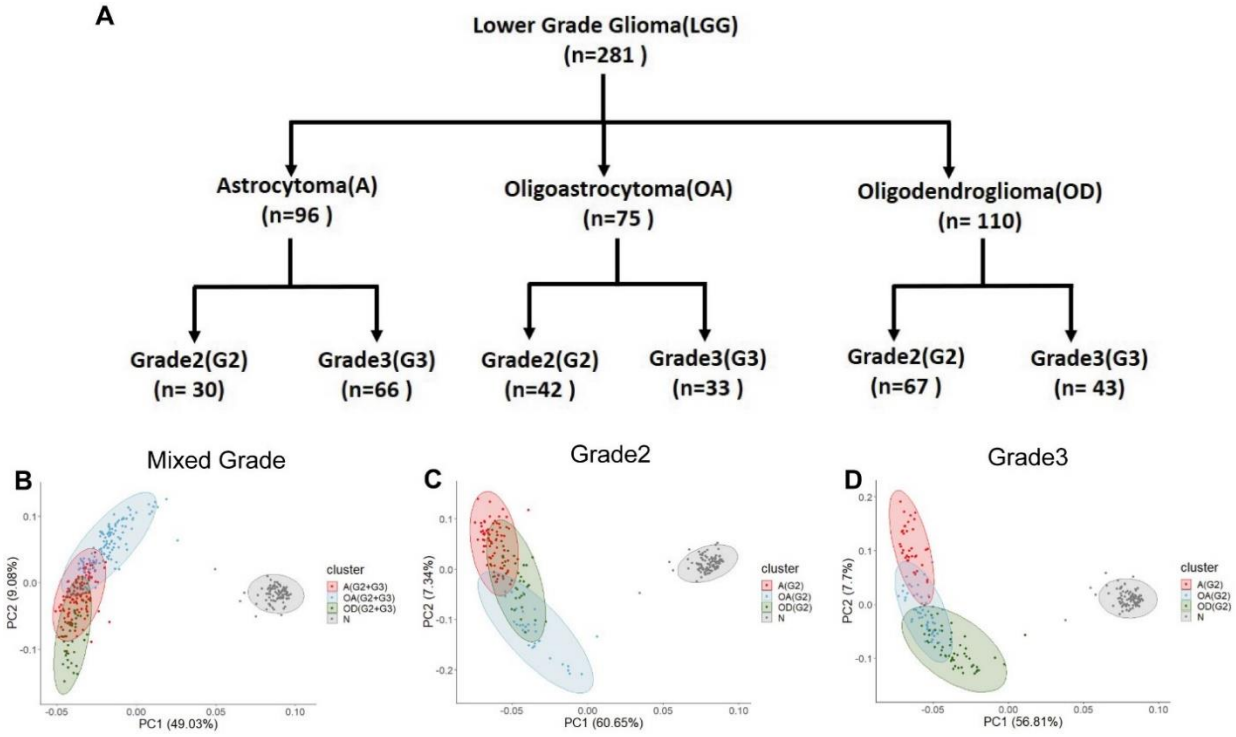

Figure 3.2: *Division of sample and clustering of patients. (A) The flow diagram shows the histological classes of LGG and the scheme of sample division. PCA plots show the clustering of the patients using gene expression data of (B), mixed grade (C), grade 2 and (D), grade 3 of LGG. A: astrocytoma, OA: oligoastrocytoma, OD: oligodendroglioma, N: healthy, G2: grade 2, G3: grade 3, G2+G3: mixed grade.*

Next, the gene expression data was pre-processed to remove the merely expressed genes, $\log_2$ (RSEM +1) < 0.1 in 90% samples, to implement the machine learning algorithms, reducing the computing time. Finally, 14,517 genes expressed in cancer and healthy tissue were subjected for further analysis. Next, the principal component analysis (PCA) was performed to observe the gene expression patterns in different subtypes and grades of LGG. We also wanted to observe the clustering of LGG patients based on the information contained in gene expression data. The first two principal components (PCs) was focused, as they captured the most variations in the data set [179,180]. The PCA was performed using mixed grade (without considering the cancer grade), grade 2, and grade 3 gene expression data to separate the subtypes. The resulting projection of PC1 and PC2 is shown in Figure 3.2 (B, C, and D), representing the clear separation between LGG and healthy cells. However, PCA could not

separate LGG subtypes. PCA successfully clustered the healthy and LGG patients due to distinct gene expression patterns. However, it failed to separate the LGG subtypes, which may be due to highly heterogeneous gene expression within the cancer patients. These results indicated a need for efficient computational tools to diagnose LGG subtypes to support the clinician.

Hence, supervised machine learning approaches were implemented to develop a framework for patient classification (Figure 3.3). The subtype classification of patients were performed using mixed grade (without considering cancer grade), grade 2, and grade 3 transcriptomics data. From the previous step, 14,517 pre-processed gene expression data of cancer cells were taken for analysis. This expression data was high-dimensional, and the number of genes was much larger than the number of patient samples; therefore, to improve the classification accuracy, we performed feature selection or gene selection before applying the supervised machine learning algorithm [181]. Besides, feature selection removes the irrelevant genes and identifies the discriminatory genes, which facilitates the improved performance of the classifier. The feature selection was done separately on grade2, grade3 and mixed grade patients. A two-step process was applied for feature gene selection. At first, a correlation-based approach was used to eliminate redundant and irrelevant features. The Pearson correlation coefficient (r_s) was computed and genes with $r_s > 0.7$ were dropped. The remaining genes, i.e., 5,943 genes in grade 2, 7,007 genes in grade 3, and 7,375 genes in the mixed grade, were taken for further analysis. Due to an unequal number of patients in different subtypes and grades of LGG (Figure 3.2A), the model can become biased towards one class, leading to poor performance of the model. Hence, we performed random sampling to select the equal number of patients in each subtype before the feature selection. We divided the balanced data into training (70%), and test (30%) datasets. Next, we have performed supervised ML-based methods for feature gene selection, i.e., SVM-RFE and Random Forest-based Boruta algorithm to select the most variable features among the classes (see Materials and methods).

Linear SVM-RFE was computed on the training data set with 5-fold cross-validation (CV). We performed the CV to avoid the issue of overfitting [182,183]. The separate rank list of genes were generated in grade2, grade3, and mixed grade cancer, and from these lists, the top 20, 50, 100, 200, and 500 feature genes were selected for further analysis. Next, different machine learning (ML) algorithms were applied to classify subtypes of LGG. Support vector machine (SVM), k-nearest neighbors (KNN), GaussianNB (GNB), Decision Tree (DT), and Random

forest (RF) was used on top 20, 50, 100, 200, 500 feature genes for classification (Figure 3.4). Initially, our objective was to determine the best set of features with the best accuracy. Therefore, we compared the accuracy score of each model for each feature gene set. The 100 feature genes had shown the best prediction accuracy by all ML models (Figure 3.4A, B, and C). The other performance parameters, i.e., recall, precision, and F1 score were also evaluated, and it is observed that 100 feature genes provided the overall highest score (appendix table I.1) in test dataset (30%). Furthermore, the PCA was performed using expression data of 100 feature genes to examine the clustering of patients. It is observed that improved subtype-specific separation between patients using 100 feature genes compared to pre-processed data, indicating that the integrated feature selection method efficiently extracted most variable features from the transcriptome data (appendix figure I.1).

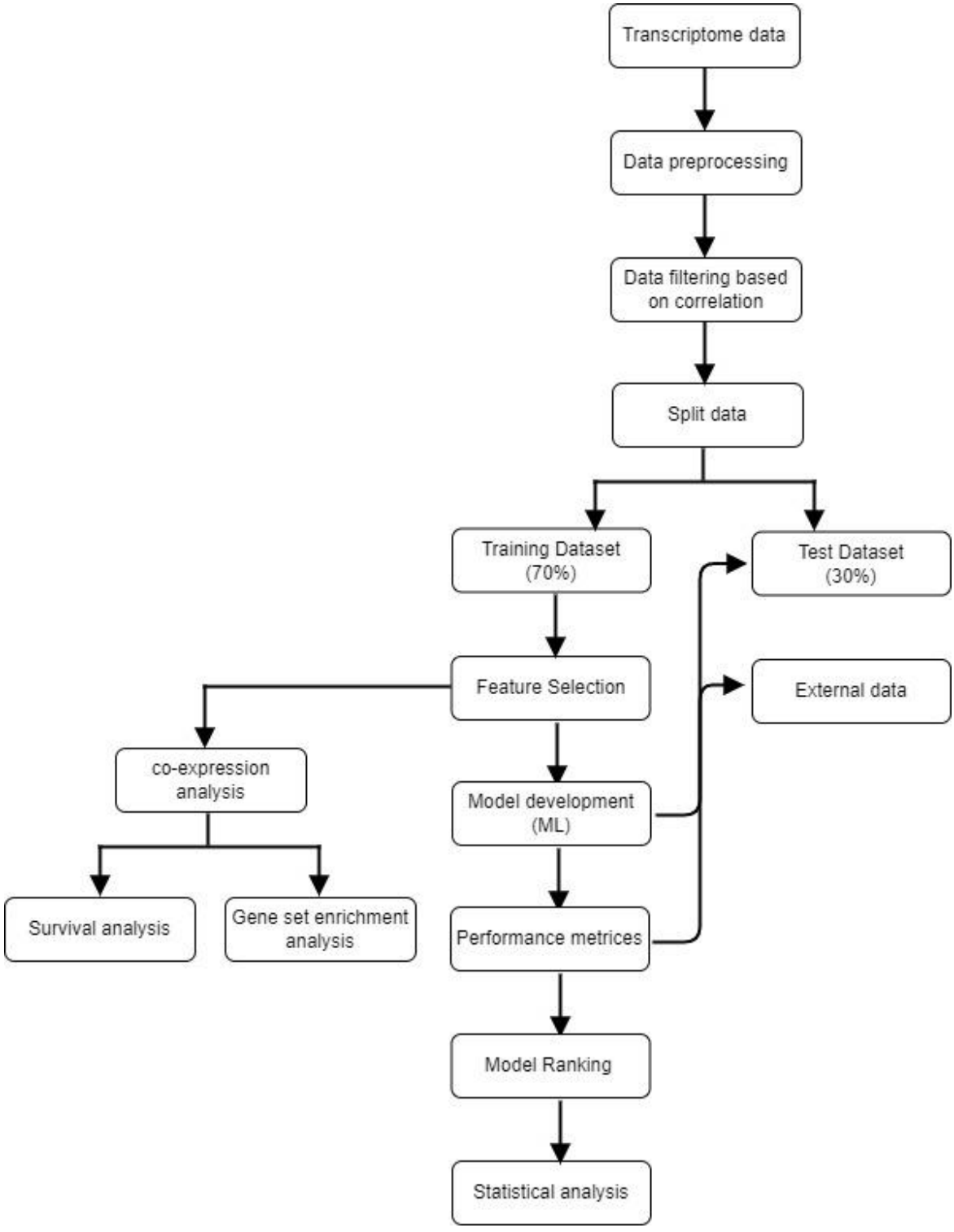

Figure 3.3: *The machine-learning framework and classification accuracy with a different set of features. The flow chart shows the machine-learning pipeline using transcriptomics data to classify the subtypes and grades of LGG.*

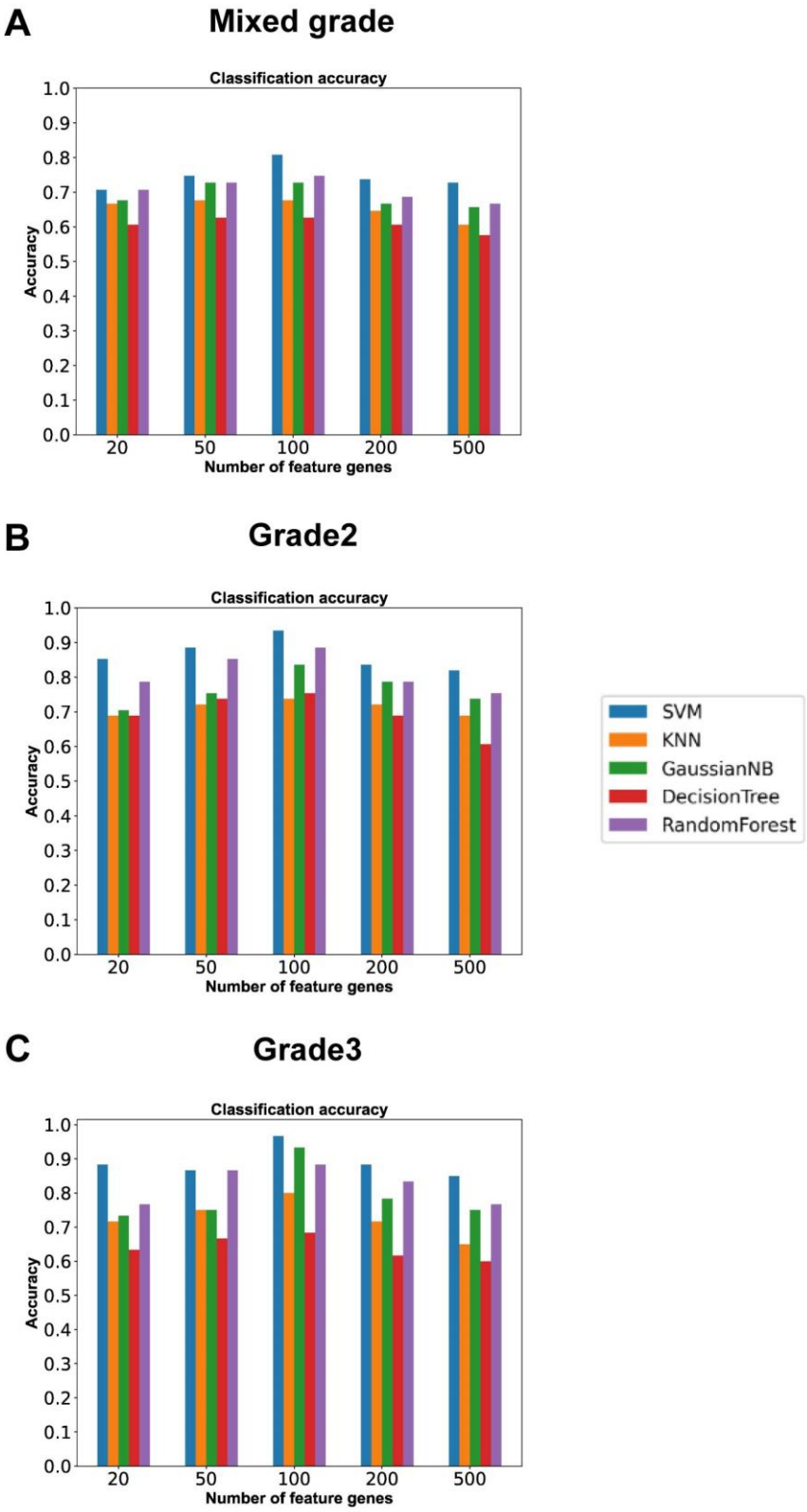

Figure 3.4: *(A, B, and C) bar plots show the accuracy of subtype prediction using different feature genes and ML algorithms.*

Next, stratified k-fold CV (k = 10) was performed using all ML algorithms on 100 feature genes. CV was performed for estimating the true accuracy of a given model; in turn, it described the reliability and stability of the models. In cross-validation, the dataset was divided into a training set and a test set. This process was repeated ten times, and in each split model's performances were measured. Then the average performance was calculated, such as average accuracy. Here, along with the average accuracy, the recall, precision, F1-score, GM, and MCC were calculated (Table 3.2). Then the overall performance of subtype classification of the ML models were compared in mixed grade, grade 2, and grade 3. It is observed that the average prediction accuracy of SVM was superior compared to the other ML models, i.e., 82% (±0.08) in mixed grade, 90% (±0.08) in grade 2, and 94% (±0.03) in grade 3. Moreover, Table 3.2 shows a key finding, the classification of subtype is always high in the specific grade of cancer compared to mixed grade. Except for SVM, for all classifiers, the accuracy of subtype classification drops below 80% with mixed grade transcriptome data. The MCC score, which represents the correlation between the observed and predicted classifications, was less in mixed grade; for SVM, the MCC score was 0.6. Whereas in grade 2 and grade 3, MCC scores were 0.79 and 0.87, respectively. A similar observation was made (appendix Table I.1 and Figure 3.4A, B, and C), that irrespective of ML-method and the number of features genes, the subtype classification accuracy was always less with mixed grade data compared to grade 2 and grade 3. These results showed that prediction of the subtype was more accurate in a particular grade of cancer. This also indicates ML- algorithms are sensitive to cancer grade. Next, the SVM performance was statistically validated using one-way ANOVA followed by Tukey-HSD test. The pairwise comparison was performed and compared the average accuracy score from the k-fold CV of SVM with other machine learning algorithms. In the previous section, it is observed that the accuracy score of SVM was always high compared to the other algorithms. Therefore, we constructed the null hypothesis as follows:

H0: The accuracy of all models is equivalent.

The alternative hypothesis is

H1: The accuracy of all models is not equivalent.

The ANOVA test result on the accuracy is provided in Table 3.3, and it is observed that SVM performance was significantly high ($p < 0.05$) compared to KNN and DT. At the same time, RF and GNB showed no significant difference in classification accuracy. A ranking was

done in order to find the best classifier. The multiple-criteria decision-making (MCDM) was performed using TOPSIS on each k-fold result [184]. All performance measures mentioned in Table 3.2 were considered for the ranking, and SVM topped the overall ranking. It is always desirable to have a highly sensitive and specific model for diagnosis. Therefore, the relationship was visualized between sensitivity and specificity using the ROC curve. ROC curve represents the probability of a true positive result or the test's sensitivity against the probability of a false-positive result for a range of different cut-off points. Figure 3.5 (A, B, and C), shows the area under the ROC curve (AUC) of all ML models, and it is observed AUC for SVM was higher in all three cases, i.e., 0.87 in mixed grade, 0.98 in grade 2, and 0.951 in grade 3. The overall performance of SVM was superior compared to other ML models. The classification was performed using two external datasets from GEO (GSE74462 and GSE43378) to validate this observation. The random sampling method was used on external datasets to equalize number of patients in each subtype before classification (appendix Table I.2) [185,186]. The prediction accuracy of SVM for a mixed grade, grade 2, and grade 3 was 89% (±0.11), 90% (±0.08), 94% (±0.12), respectively (Table 3.4). It is also observed that the MCC score was ≥ 0.80 in all three classifications. Hence, SVM was the best classifier for subtype classification for transcriptome data from model building to validation. To compare the efficiency of the present framework, the subtype classification was performed again using the features computed by the Boruta algorithm. The performance of all five ML algorithms were compared, i.e., SVM, KNN, GNB, DT, and RF, using the Boruta features gene. Results are summarized in appendix Table I.3 and indicate that features computed using SVM-RFE were substantially better than the Boruta in terms of classification accuracy. Therefore, correlation and SVM-RFE for feature gene selection and then subtyping using SVM can be efficient tools for clinical diagnosis of LGG subtypes. However, interestingly the feature genes from Boruta showed a similar trend of classification, i.e., classification accuracy was high with only grade 2 or grade 3 gene expression data compared to mixed grade.

Table 3.2: Models' performance and ranking

	Methods	Performance measures (Average of 10 fold cross-validation)						Rank
		Accuracy	Precision	Recall	F1-score	Gmean	MCC	
Mixed grade	**SVM**	0.82%(±0.08)	0.7263	0.7439	0.7237	0.8196	0.6	1
	RF	0.79%(±0.05)	0.6768	0.7044	0.6724	0.784	0.53	2
	DT	0.66%(±0.05)	0.4925	0.5303	0.4853	0.6567	0.25	5
	KNN	0.69%(±0.05)	0.5382	0.5645	0.5381	0.6904	0.32	4
	GNB	0.77%(±0.05)	0.6504	0.6733	0.6511	0.7664	0.49	3
Grade2	**SVM**	0.90%(±0.08)	0.8607	0.8676	0.8558	0.9035	0.79	1
	RF	0.86%(±0.05)	0.7967	0.808	0.791	0.8598	0.69	3
	DT	0.79%(±0.08)	0.6929	0.7231	0.6722	0.7866	0.55	4
	KNN	0.75%(±0.09)	0.6321	0.6546	0.6245	0.7482	0.45	5
	GNB	0.88%(±0.06)	0.8186	0.8353	0.8119	0.8747	0.73	2
Grade3	**SVM**	0.94%(**±0.03)**	0.9035	0.9267	0.9016	0.9368	0.87	1
	RF	0.86%(±0.05)	0.795	0.827	0.7882	0.8575	0.7	3
	DT	0.78%(±0.07)	0.6729	0.671	0.6603	0.7797	0.52	5
	KNN	0.80%(±0.04)	0.6938	0.6847	0.6606	0.7923	0.56	4
	GNB	0.89%(±0.07)	0.8469	0.8496	0.8383	0.8924	0.77	2

Table 3.3: ANOVA followed by Tukey-HSD test

SI.No.	Comparison	Mixed Grade	Grade2	Grade3
		p-value<0.05	p-value<0.05	p-value<0.05
1	SVM vs. KNN	Yes	Yes	Yes
2	SVM vs. DT	Yes	Yes	Yes
3	SVM vs. RF	No	No	No
4	SVM vs. GNB	No	No	No

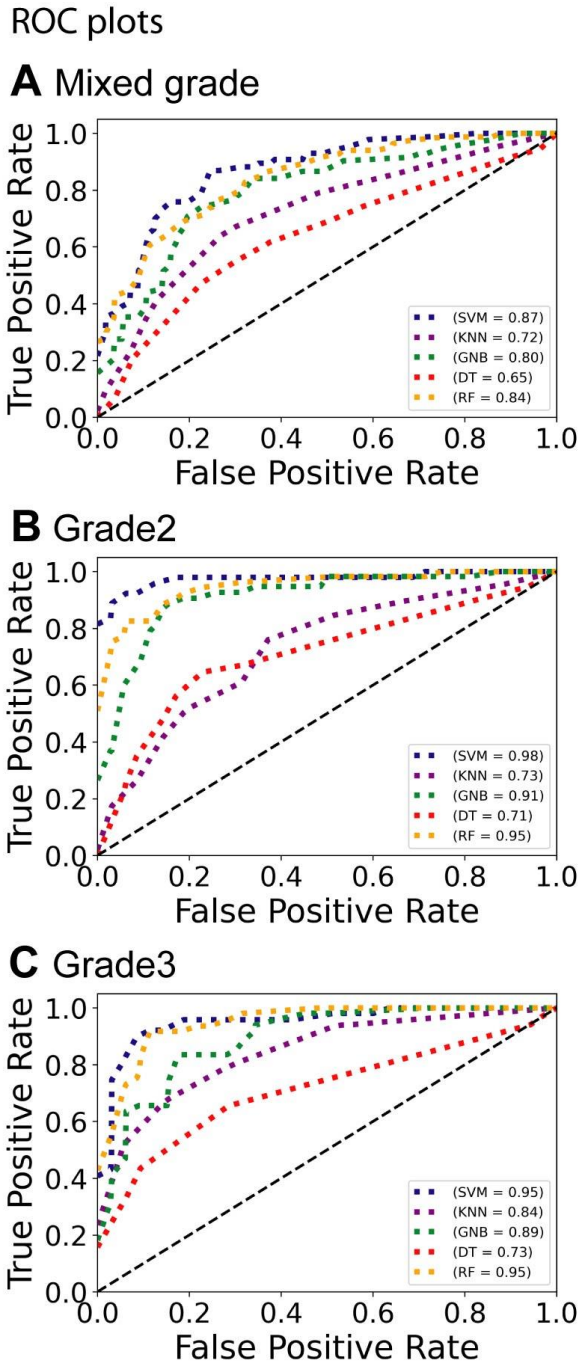

Figure 3.5: *ROC of various prediction models. (A–C) ROC plots were generated using an independent dataset. SVM, support vector machine; KNN,* k*-nearest neighbors; GNB, Gaussian Naïve Bayes; DT, Decision Tree; RF, Random Forest.*

Table 3.4: Performance of SVM with independent datasets

	Performance measures (Average of 10 fold cross-validation)					
	Accuracy	**Precision**	**Recall**	**F1-score**	**GM**	**MCC**
Mixed Grade	0.89(±0.11)	0.8583	0.8700	0.8420	0.8869	0.80
Grade 2	0.90(±0.08)	0.8350	0.8558	0.8199	0.9025	0.81
Grade 3	0.94(±0.12)	0.9300	0.9350	0.9216	0.9453	0.88

3.3.2 Simultaneous Subtyping and Grading of LGG using SVM

In the previous section, it was observed that it was necessary to identify grades to achieve higher accuracy in subtype classification. Additionally, the grade of cancer determines the malignancy level. The correct grade stratification has significant implications in determining the patient treatment plan [187]. Therefore, simultaneous identification of grade and subtype will greatly support clinicians in deciding accurate treatment strategies. Henceforth, we decided to classify the six classes, and subsequently divided the whole LGG data set into astrocytoma grade 2 (n = 30), astrocytoma grade 3 (n = 66), oligoastrocytoma grade 2 (n = 42), oligoastrocytoma grade 3 (n = 33), oligodendrogliomas grade 2 (n = 67), and oligodendrogliomas grade 3 (n = 43) (Figure 3.2 A). For classification, the top 100 feature genes of grade 2 and grade 3 were used from previous steps. The grade 2 and grade 3 features were combined and screened the unique list of 178 features out of 200. It is worth mentioning that only 11 % of feature genes were common between grade 2 and grade 3, which is a nonsignificant overlap (p =1.0, hypergeometric test). This shows that these feature genes could be used as grade and subtype-specific biomarkers. Next, the SVM was implemented using the gene expression data of 178 genes and calculated the performance measures as previously described (Table 3.5). The average accuracy of the model in k-fold CV (k = 10) was 91% (±0.02), indicating that SVM efficiently classified the six classes and showed stable performance. To further confirm, the model was executed on the independent test dataset, and accuracy was 93.39%. It is also examined the accuracy of prediction of individual class, that is, astrocytoma grade 2 (accuracy = 93.38%), astrocytoma grade 3 (accuracy = 95.04%), oligoastrocytoma grade 2 (accuracy = 87.60%), oligoastrocytoma grade 3 (accuracy = 94.21%), oligodendrogliomas grade 2 (accuracy = 95.86%), and oligodendrogliomas grade 3 (accuracy = 94.21%). It was noted that the prediction accuracy was > 90% in maximum classes.

Furthermore, we analyzed the performance using the ROC plots (Figure 3.6). The average AUC for the six classes is 0.88 and the individual class AUC varies from 0.79 to 0.94. This indicates that the model was highly specific and sensitive. The simultaneous classification of grade and subtype with higher accuracy can be a major support and breakthrough for clinicians for patient management.

Table 3.5: Performance of SVM for multi-class (six class) classification to predict the grade and subtype simultaneously

	Performance measures					
	Accuracy	**Precision**	**Recall**	**F1-score**	**GM**	**MCC**
Training data (k=10)	0.91(±0.02)	0.7646	0.7502	0.7444	0.8372	0.70
Independent data	0.9339	0.8168	0.8024	0.8024	0.8714	0.76

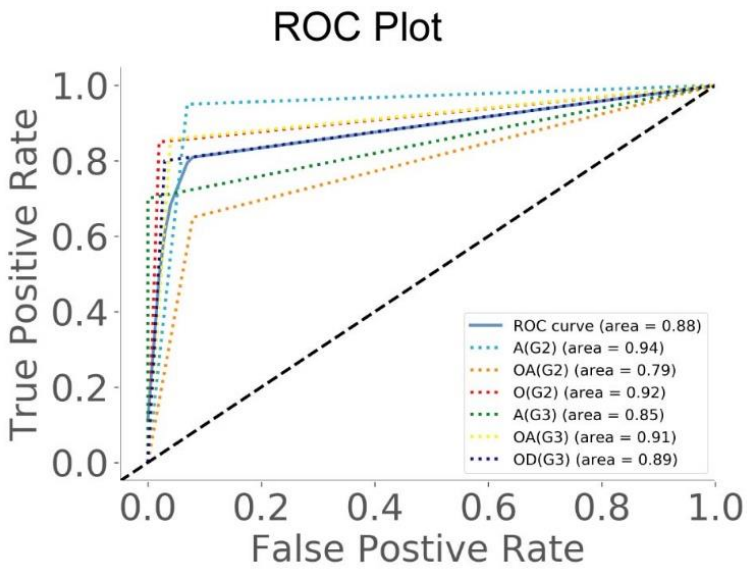

Figure 3.6: *Model performance. ROC plot for multi-class (six class) classification.*

3.3.3 Grade and Subtype-specific Co-expression Pattern of Feature Genes and Biological Relevance

In our approach, feature selection has a crucial role in achieving higher prediction accuracy. However, ML algorithms were restricted from constructing an accurate prediction model as a black box because they are rarely linked with biological processes and functions. However, the association of these features with biological processes would allow us to explore them as potential biomarkers. Furthermore, if these biomarkers are specific to a subtype in a particular grade, then it will be an advantage for precision therapy. The co-expression network was constructed and analyzed between the feature genes to explore such a possibility. The gene to gene correlation is linked with specific disease states because gene expression patterns are not the same in different cellular conditions. Again, the gene expression level often determines the corresponding protein's functional activity, which is directly linked with the biological processes and molecular functions [21]. Hence, our study focused on the examination of the co-expression pattern of the feature gene and its association to specific-subtype in a grade. The Pearson correlation coefficient (r) was computed for each pair of feature genes using the gene expression data of a specific subtype in a grade. $r > 0.5$ ($p < 0.05$) was selected as the threshold to screen the statistically significant co-expressed gene pairs. A total of 82, 36, and 73 correlated gene pairs in grade 2 (Figure 3.7A, B, and C), and 27, 116, and 98 correlated gene pairs in grade 3 (Figure 3.7D, E, and F) were present in astrocytoma, oligoastrocytoma, and oligodendroglioma, respectively. Next, the list of co-expressed genes was compared between the subtypes using a hypergeometric test. It was found there was a nonsignificant ($p > 0.05$) overlap between the genes in subtypes except astrocytoma and oligodendroglioma ($p = 0.004$) (Table 3.6). The nonsignificant overlap indicates that many co-expressed genes are specifically associated with a particular subtype in a grade.

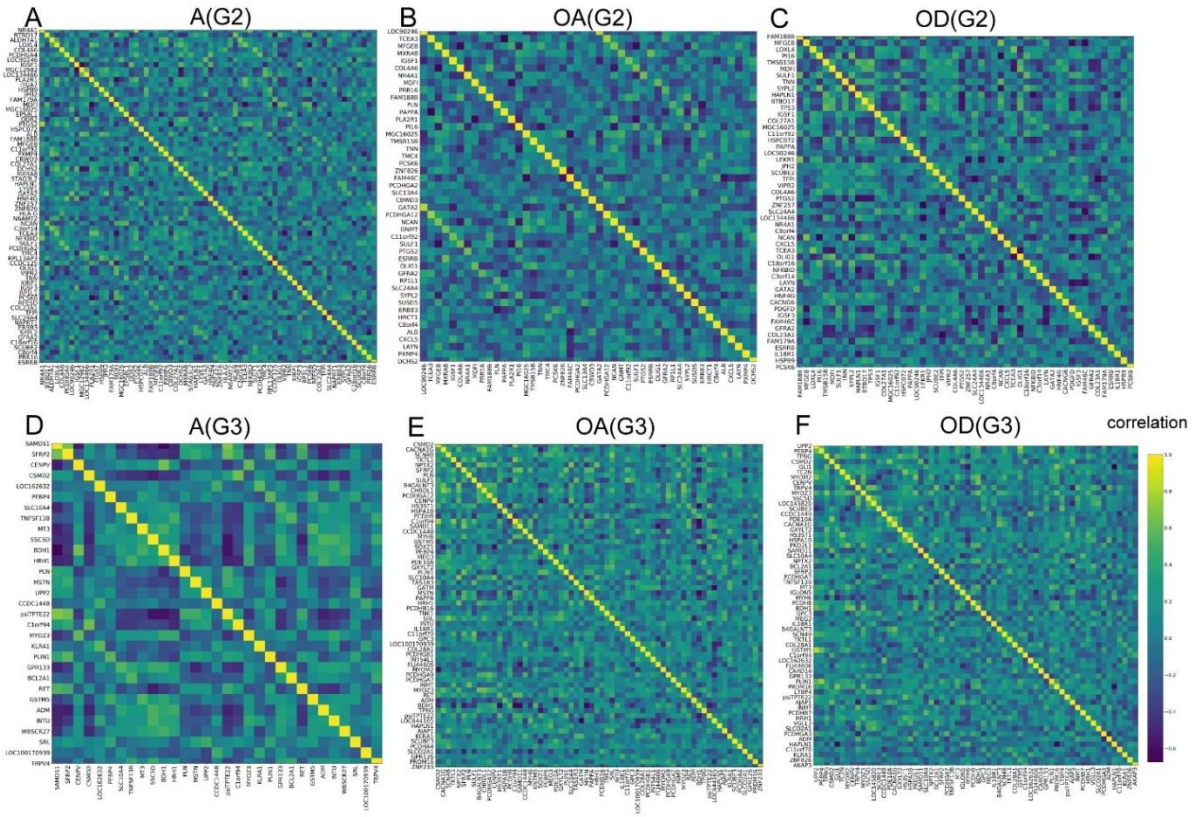

Figure 3.7: *Biological relevance of feature genes. (A-F) The correlation heat maps of feature genes in different subtypes and grades as mentioned in the figure. A, astrocytoma; OA, oligoastrocytoma; OD, oligodendroglioma; G2, grade 2; G3, grade 3.*

Table 3.6: Statistical significance of the overlap between two groups of feature genes

Hypergeometric test	
Grade 2	***p*-value**
Astrocytoma Vs oligoastrocytoma	0.0824
Astrocytoma Vs oligodendroglioma	0.004
Oligoastrocytoma Vs oligodendroglioma	0.032
Grade 3	***p*-value**
Astrocytoma Vs oligoastrocytoma	0.7771
Astrocytoma Vs oligodendroglioma	0.2437
Oligoastrocytoma Vs oligodendroglioma	1

Next, the gene set enrichment analysis was performed to understand the biological relevance of co-expressed feature genes in a specific subtype and grade using the Metascape tool [178]. It was observed that enriched biological processes and pathways were linked to oncogenic events (Figure 3.8). Such as extracellular matrix organization in astrocytoma (grade 2), TNF signaling pathways and PI3k-Akt signaling pathway in oligodendroglioma. These biological processes are generally activated in glioma, affecting the biological behavior of tumors, and linking to patient prognosis and survival [188–190]. Furthermore, hemophilic cell adhesion molecules in oligoastrocytoma are involved in the growth and progression of glial tumors [191]. Similarly, the calcium signaling pathway in astrocytoma (grade3), chemical synaptic transmission in oligoastrocytoma, and NF-kappa β signaling pathway in oligodendroglioma are the prominent signature of brain cancer formation and progression [192–194].

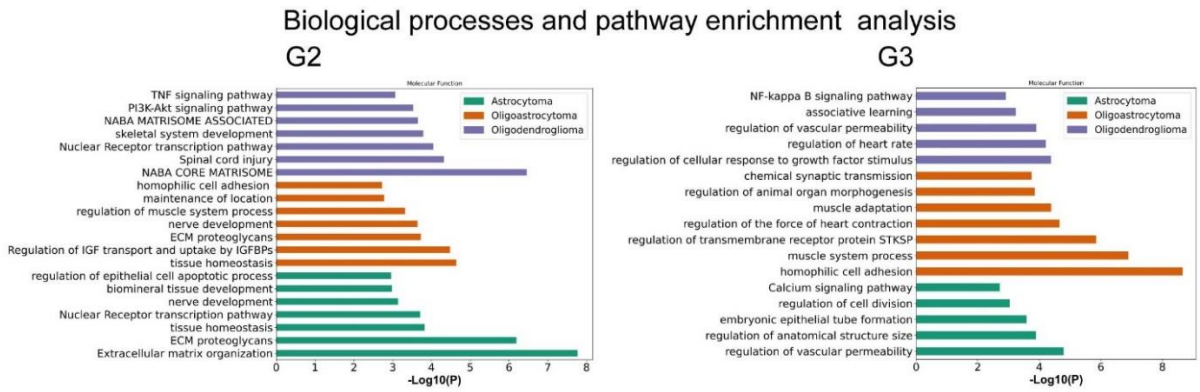

Figure 3.8: *Biological processes and pathway enrichment analysis of co-expressed feature genes. (G2), grade 2; (G3), grade 3.*

It was noticed that various biological processes and pathways had been enriched, which differed among the cancer subtype in a grade (Figure 3.8). These results represent that a distinct gene expression pattern and biological processes are linked with the subtypes within a particular grade of LGG. There are many divergences between grade 2 and grade 3, although both grades are considered lower-grade cancer. This observation again illustrates that identifying grade and subtype is crucial for finding a proper therapeutic intervention.

3.3.4 Identification of Prognostic Biomarkers of LGG for Diagnosis and Treatment

Correlation analysis and hypergeometric test in the previous section revealed that many of the co-expressed genes (features) were unique to the specific subtype in a grade. Importantly, biological processes and pathway enrichment analysis showed that these co-expressed genes were also associated with oncogenic processes. This allows further evaluation of these genes as a potential biomarker for cancer therapy and diagnosis. Next, survival analysis and log-rank test was performed using GEPIA web tools of the genes unique to a specific subtype in a particular grade (PCC>0.5, and $p < 0.05$). Survival analysis revealed that several genes were linked to patient survival. Here, the Kaplan–Meier survival plot has been shown, illustrating the genes whose higher expression is associated with poor prognosis (Figure 3.9). Furthermore, several experimental pieces of evidence showed that many genes are involved in oncogenic processes in brain cancer. Such as in grade 2 cancer, MGC12982 (FOXD2-AS1) [195], ITGA7 [196], EPS8L1 [197], DDR2 [198], STAG3L2, and HLA-G [199,200] in astrocytoma, TP53 [201,202], and PDGFD [203,204] in oligodendroglioma, were significantly ($p < 0.05$) associated with worse patient survival. In grade 3 cancer, GATM [205], PCDHGA9 [206] and GPR126 (ADGRG6) [207] in oligoastrocytoma, and GLI1 [208], TC2N [209], PKD2L1 [210], LTBP4 [211], and VGLL3 [212], in oligodendroglioma, were linked to worse patient survival. Additionally, we identified several new genes, i.e, which are not reported before in LGG, such as N6AMT2, CCDC125, and RFESD in astrocytoma, HRCT1 in Oligoastrocytoma, LEKR1 in oligodendroglioma of grade 2, and in grade 3, WBSCR27 in astrocytoma, PCDHB16 in oligoastrocytoma, and PCDHB7 in Oligodendroglioma. The higher expression of these genes affects patient survival. The co-expression of these genes in LGG subtypes and association with patient survival shows the possibility to identify them as grade and subtype-specific prognostic biomarkers.

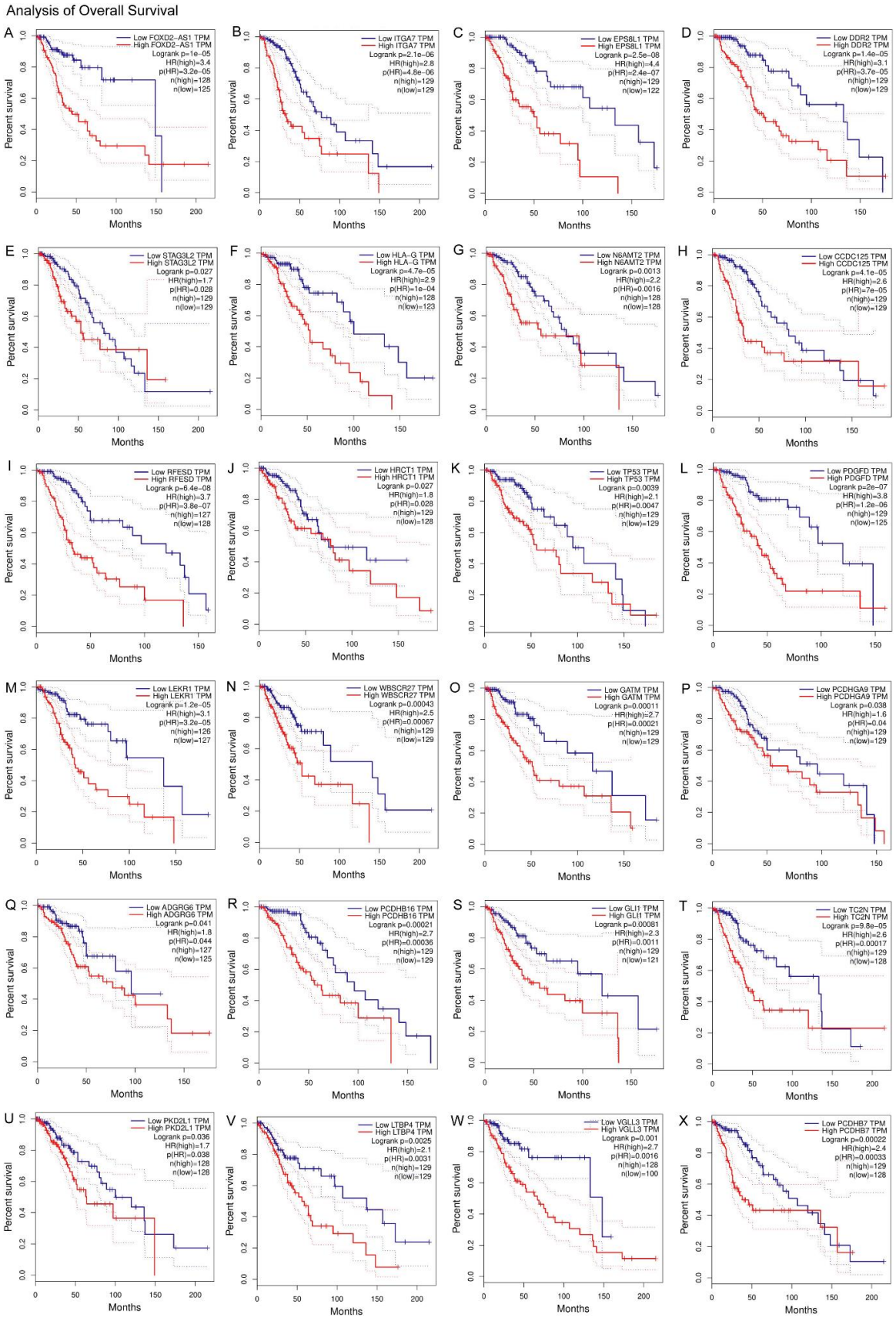

Figure 3.9: *Survival analysis of feature genes., (A-I) astrocytoma grade2, (J) oligoastrocytoma grade2, (K-M) oligodendroglioma grade2, (N) astrocytoma grade3, (O-R) oligoastrocytoma*

grade3, (S-X) oligodendroglioma grade3.Overall survival was analyzed based on the clinical information of the patients from TCGA and quartile method of 75 % cut-off of higher and 25% cut-off of lower limit.

3.4 Discussion

In this chapter, publicly available transcriptomes of LGG for subtyping and grading was explored. Heterogeneity in the tumour is the main issue for molecular subtyping and precision treatment of brain cancer patients. We present a comprehensive and precise ML-based approach for cancer grading and subtyping. It was observed that a subtype of LGG was not separable using PCA. This result led us to design an ML-based framework for the accurate prediction of LGG subtypes. It is found an integrated approach consisting of correlation and SVF-RFE algorithm for feature gene selection, and then computation of SVM using those feature genes (n = 100) had shown superior performance (accuracy > 90%). **We found that the accuracy of subtype classification is always good using the gene expression data of a specific grade of cancer rather than a mixed grade.** We repeatedly observed the same with other ML techniques. This gave us clues that cancer grading is essential to achieve higher accuracy for subtype prediction. Further, the performance of the SVM was statistically verified through the one-way ANOVA followed by a Tukey-HSD test. A pairwise comparison was conducted to evaluate and contrast the average accuracy score obtained from 10-fold cross-validation of SVM with that of other machine learning techniques. The results indicate that the performance of SVM was significantly higher ($p < 0.05$) when compared to other machine learning techniques. Then, **six-class classification was performed for simultaneous grading and subtyping using the same ML framework and attained an overall accuracy of 91.0% (±0.02) and AUC=0.88.** Therefore, the findings of this study strongly strengthen the fact that grading and subtyping are both required to achieve a higher accuracy of prediction. **Indeed, cancer grade and subtype are the essential clinical parameters to design the treatment plan and determine the patient's prognosis.**

The correct set of feature genes and their discriminative ability play a crucial role in the superior performance of ML algorithms. In addition, the biological relevance of these features could lead to finding the formation and therapeutic targets. It is analyzed that the expression data of feature genes and their biological significance in a similar line of thought. We identified

the subtype and grade-specific co-expressed feature genes associated with the oncogenesis. Furthermore, survival analysis of these genes revealed several predictive biomarkers, which could be used as potential molecular indicators for diagnosis and treatment. Therefore, **we conclude that gene expression data of a subtype of LGG without considering the grade is more heterogeneous than data of a specific grade.** The higher heterogeneity in the data resulted in lower accuracy of subtype prediction. **Lastly, the findings of the present study and ML-based framework can offer new avenues for developing subtype- and grade-specific therapeutic strategies.** To promote the further development for building more accurate biological relevant models and identification of novel therapeutic marker multi-omics data analysis is essential, which has grown in popularity in cancer research in recent decades. Moreover, the integration of transcriptomic, epigenomic, proteomic, and metabolomic data can reveal the intricate systemic dysregulation linked to the phenotype of lower-grade glioma.

CHAPTER 4

OBJECTIVE 2

Chapter 4: Objective 2
Development of Deep learning and machine learning frameworks based on genomic data for subtyping glioblastoma multiforme (GBM) and identification of biomarkers

4.1 Background

Glioblastoma multiforme (GBM), which is the grade IV of glioma, is a highly invasive and devastating primary form of brain cancer. The complexity and molecular heterogeneity of GBM pose the challenge for accurate diagnosis and therapy [20,61]. Because of enormous molecular heterogeneity and difficulty in early diagnosis, the molecular mechanisms of GBM tumorigenesis is not clear. Understanding molecular features that facilitate aggressive phenotypes in glioblastoma multiforme (GBM) remains a major clinical challenge. There are many other studies to find other subtypes using omics and clinical data [213]. Histopathological-based diagnosis is the most common method for subtype identification. However, it often leads to inaccurate classification of subtypes due to inter-observer variability [18]. To find the curative solution, understanding the molecular features and identification of GBM subtypes is crucial. GBM is currently classified into three subtypes i.e., classical, proneural, and mesenchymal. Accurate pathological subtype diagnosis is pivotal for optimal patient management. Because, GBM subtypes are histologically and genetically heterogeneous, differs in gene expression, mutation, and epigenetic states, which lead to different therapeutic response and clinical outcome [19,20].

Recent advances of sequencing technologies have helped generate massive omics data in cancer, leading to a deep understanding of the molecular mechanisms in both common and rare cancers [214,215]. The Genome-wide analysis revealed that changes in gene expression and methylation patterns in several positions in the genome are strongly associated with the GBM formation and progression [216–218]. Gene expression and methylation are both strongly interlinked processes; methylation levels in promoter regions influence the gene expression by regulating the transcription factors binding [11]. On many occasions, hypermethylation of CpG sites on promoter regions inhibits the gene expression, whereas hypomethylation causes higher

expression of genes [219]. Therefore, classification using multiple "omics" data, i.e., transcriptome and methylome, can provide optimal features for the clinical diagnosis of cancer subtypes. However, enormous amounts of genetic and epigenetic alterations pose challenges to finding the unique molecular marker for diagnosing GBM subtypes. Benefitting from recent advances in computational methods such as deep learning (DL) and traditional machine learning (ML), it is possible to scan the genome-wide transcriptome and methylome data to find the subtype-specific molecular feature for diagnosis [220].

In this chapter, ML and DL algorithms were implemented for the precise and accurate classification of GBM subtypes. Each data type (i.e., transcriptome and methylome) and their integrated data (patients having both transcriptome and methylome data of GBM) were separately used for classification. In addition, the biological relevancy of features were examined using weighted gene co-expression network analysis (WGCNA) and Gene Ontology (GO) analysis. Furthermore, these co-expression module genes were used to identify subtype-specific prognostic biomarkers for GBM diagnosis and treatment.

4.2 Methodology

4.2.1 Data collection, preprocessing, and integration

In this study, we analyzed TCGA glioblastoma multiforme (GBM) transcriptome (RNA-seq) and methylome (Illumina Infinium HumanMethylation450 platform) data. The dataset was retrieved from UCSC Xena (https://xena.ucsc.edu/) [165]. Log2 (RSEM +1) (RSEM: RNA-Seq by Expectation Maximization) values for transcriptome and β values for methylation were used for analysis. Next, the lowly expressed genes were removed from transcriptome data (log2 (RSEM +1) <0.1 in 90% sample), and data was scaled before analysis. Based on the clinical information, patients (n=155) were divided into three categories based on cancer subtype, i.e., classical (n=42), mesenchymal (n=55), and proneural (n=39) for transcriptome data (Table 4.1). Similarly, we have divided the methylome data (n=84) into a particular subtype, i.e., classical (n=29), mesenchymal (n=32), and proneural (n=23) (Table 4.1). Next, based on the clinical information, patients with both transcriptome and methylome profiles in TCGA were screened to integrate the transcriptome and methylome data. The total number of these patients with transcriptome and methylome data was 52, including classical (n=16), mesenchymal (n=22), and proneural (n=14) (Table 4.2).

Table 4.1: Details the tumor samples in transcriptome and methylome data of GBM.

	Transcriptome			Methylome		
Type	**Grades**	**Subtypes**	**Samples**	**Grades**	**Subtypes**	**Samples**
GBM	Grade IV	Classical	42	Grade IV	Classical	29
		Mesenchymal	55		Mesenchymal	32
		Proneural	39		Proneural	23

Table 4.2: Details the tumor samples having both transcriptome and methylome data of GBM.

	Patients with transcriptome and methylome samples		
Type	**Grades**	**Subtypes**	**Samples**
GBM	Grade IV	Classical	16
		Mesenchymal	22
		Proneural	14

Due to the unavailability of healthy patient data for both transcriptome and methylome, we used the Z-score to classify higher and lower expression of genes and hyper-and hypo-methylated CpG sites. We have calculated the Z-score for each gene or CpG site in a specific subtype using the following formula.

$$Z - score = \frac{\overline{x} - \mu}{\sigma}$$

Here, $\overline{x}$ represents subtype-specific average expression or methylation level of a gene/CpG site, while μ and σ represent the population mean and population standard deviation, respectively [221]. We have applied Z-score>1 for higher expression and hypermethylation and Z-score<-1 for lower expression and hypomethylated on each subtype of GBM. Next, we screened the higher and lower expressed genes whose promoter regions were differentially methylated, considering that the differential methylation in the promoter regions may alter the corresponding gene's expression. Finally, genes with both differential expression patterns and differential methylation promoter regions were used for further analysis [22,222]. We have collected the external dataset from the Gene Expression Omnibus (GEO) repository for

validation. GSE145645 was used to validate the model constructed using transcriptome and integrated data. GSE145645 contained all the subtypes of GBM, i.e., classical (n=9), mesenchymal (n=14), and proneural (n=9). Models built on methylome data were further validated using GSE128654, which consisted of classical (n=11), mesenchymal (n=8), and proneural (n=10) subtypes.

4.2.2 Clustering using Principle component analysis (PCA)

The subtype-specific clustering of patients using transcriptome, methylome, and patients having transcriptome and methylome samples named as integrated data was visualized by principal component analysis (PCA) (A detailed description was provided in the chapter 3.2.2). We used PCA for visualization of GBM subtypes; ggfortify and cluster package in R was used.

4.2.3 Features selection by Least absolute shrinkage and selection operator (LASSO)

The feature or variable selection was performed to improve the performance of ML and DL algorithms. The least absolute shrinkage and selection operator (LASSO) approach regularises model parameters by decreasing some regression coefficients to zero. After the shrinkage, comes the feature selection phase, during which all non-zero values are chosen to be incorporated in the model. LASSO regularization is a crucial idea that helps to prevent data overfitting. In order to attain a lower variance with the tested data, regularization is accomplished by adding a penalty term to the best fit produced from the trained data. Regularization also limits the effect of predictor variables over the output variable by compressing their coefficients. LASSO models provide good prediction accuracy. Since the method involves shrinking of coefficients, which lowers variance and minimizes bias, the accuracy increases. The LASSO was performed on all types of preprocessed data [223]. The default parameter values were used for lambda (tuning factor that controls the strength of penalty) and dropped those genes with a coefficient value 0. LASSO was implemented in the ScikitLearn (https://scikit-learn.org) package in Python.

4.2.4 Machine learning and Deep learning models for classification of GBM subtypes

Classification was performed on the subtype of GBM as a multi-classification problem using gene expression levels as covariates. Several machine learning and deep learning algorithms were used for classification: support vector machine (SVM), K-nearest neighbors (kNN), naïve bayes (NB), random forest (RF) (A detailed description of above classifiers were provided in the chapter 3.2.5 in details), logistic regression (LR), and convolutional neural network (CNN). LR and CNN models are discussed here:

4.2.4.1 Logistic Regression (LR)

Logistic regression (LR) is the most popular supervised machine learning algorithm. A logistic regression classifier predicts the response based on one or more predictor variables. It measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function. Logistic regression is basically used for solving classification problems. The logistic function curve indicates that, whether the cancer patients are healthy or cancerous. It has the ability to provide a probability to classify new data using discrete datasets [224]. LR was implemented using the *LogisticRegression* package in Python.

4.2.4.2 Convolutional neural network (CNN)

Convolutional neural network (CNN) is one of the deep feed-forward artificial neural network architectures that consist of the convolutional layer, activation function, and pooling layer. CNN's is a type of neural network that are fully connected networks extracts the features from the training dataset i.e., each neuron in a layer is directly connected to all neurons of the next layer. It is consist of sequence of layers, each has specific functions such as convolution, pooling and fully connected layer. Each layer takes the output from the previous layer as an input. This is followed by certain number of convolution layers composed of certain number of filters called kernel that are convolved with the input data to obtain feature maps. CNN's have a kernel that convolves the input to extract localized features and aggregate those using a pooling layer, enabling the model to extract features at all levels.

Convolution is one type of linear operation used instead of general matrix multiplication in convolution layers where filters are applied to original data or to feature maps in deep CNN. The convolution operation (denoted by an asterisk) is defined by:

$$f(t) = (x * K)(t)$$

Where the function $x\,(t)$ is referred to as input, $K(t)$ is referred to as kernel, and the $f(t)$ is referred to as output.

Next, convolutional layer followed by pooling layer that introduces non-linearity to the activations and performing down sampling reducing the number of parameters and select more salient features that the network needs to learn. Most common activation is rectified linear unit (ReLu). It allows for faster and more effective training by mapping negative values to zero and maintaining positive values. The activated features are carried forward into the next layer. Next, layer is fully connected layer (FC) that is the end of the CNN model. The input of the FC layer comes from the last pooling or convolutional layer. Finally FC layer convert the data into suitable form, flattened into the vector and fed into the feed forward neural network. After training of many epoch, cancer and normal or multiclass classes are identified by CNN using softmax classifier. Therefore, it is efficient in extracting the relevant features from multidimensional data. CNN is widely used for feature extraction to classify cancer using genomics data such as gene expression data [225,226], and methylation [227,228]. Here 1D-CNN was used to perform the classification of GBM subtypes using gene expression, methylation and integrated data. The 1D-CNN architecture is provided in the figure 4.1.

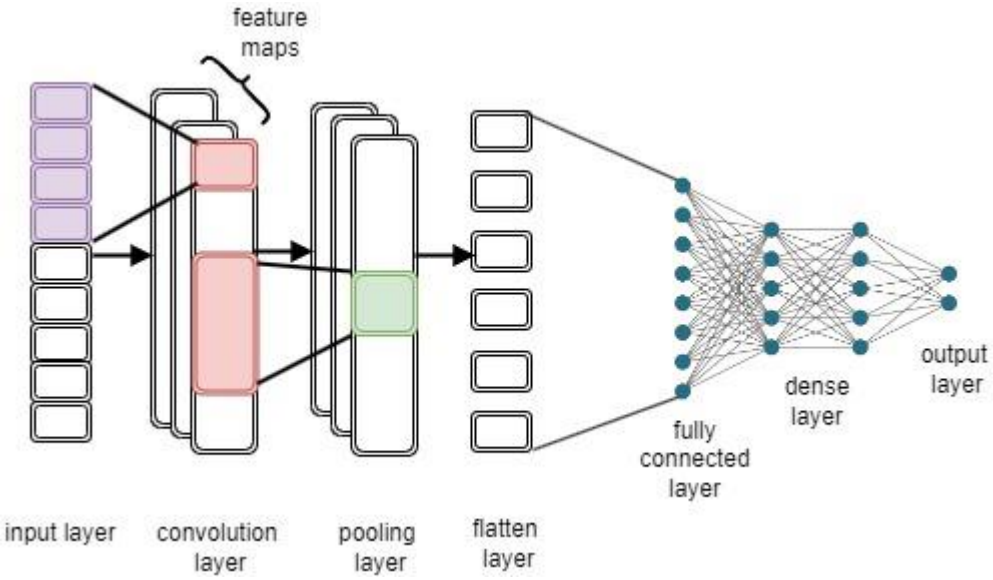

Figure 4.1: *Architecture of 1D-CNN used for GBM subtype classification.*

4.2.4.3 Hyperparameter tuning

Most machine learning and deep learning algorithms require optimum parameters to reach the robust performance of the model. Sometimes default parameters are not reached at that point. Only a hyperparameter is one way to select the best parameters while training the model. GridsearchCV from the sklearn library provides the best hyperparameters of the model, as it tries with all the given parameters using cross-validation. Hyperparameters selected from the grid search module for every model.

In this paper, all machine learning classifiers are built on the Python platform by using sklearn library. Keras library was used to construct the model architecture for CNN. Eight convolutional layers were used for obtaining the best result. All parameters for CNN were provided in table 4.3. After obtaining optimal features, stratified k-fold was applied on the 70% training dataset and average performance measures were recorded. In Stratified k-fold CV, the dataset is divided into k independent folds where k-1 folds were used to train the network, and the remaining one is reserved for the test purpose. This procedure is then repeated until all folds are used once as a test set. The final output is then computed by averaging over the obtained performance parameters from each test set (A detailed description was provide in the chapter 3 section 3.2.5.6 and figure 3.1 in details).

Table 4.3: Parameters of CNN

Parameters	Datasets		
	Transcriptome	**Methylome**	**Integrated**
Activation	relu	relu	relu
Batch_size	50	50	50
Dropout_rate	0.1	0.1	0.1
epochs	100	100	100
filters	32	1	3
Init_mode	uniform	uniform	uniform
Kernel_size	5	3	3
optimizer	RMSprop	RMSprop	Adam

4.2.4.4 Performance evaluation

The performance of ML and deep learning models was evaluated using accuracy, recall, precision, F1-score, FPR, GM, and MCC. As described in previous chapter, we generated a confusion matrix to compute these performance scores. True positive (TP), true negative (TN), false positive (FP), false negative (FN) were calculated from the confusion matrix [229]. Then, we calculated the accuracy or success rate as,

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

The sensitivity or true positive rate of a ML model was measured using the following equations

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

The specificity or true negative rate a ML model was measured using the following equations

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

The precision or positive predicted value was measured using the following equations

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

A measure of model performance that combines precision and recall into a single number is known as the F measure or F1-score. The following equation was used to compute the F1-score.

$$F1 - score = \frac{2 \times TP}{2 \times TP + FP + FN}$$

The FPR of a ML model was measured using the following equations

$$FPR = \frac{FP}{TN + FP}$$

Geometric mean (GM) is the average value or mean which signifies the central tendency of the set of numbers by taking the nth root of the product of their values.

$$Geometric\ mean(GM) = (x_1.x_2 \ldots\ldots.x_n)^{1/n}$$

Mattews correlation coefficient (MCC) measures the correlation of the true classes with the predicted labels.

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

We used the sklearn.metrics library in Python to calculate the above score by importing functions such as confusion_matrix and classification_performance. Finally, we visualized the model performance across a wide range of conditions using receiver operating characteristic curve (ROC) plots using the roc_curve function.

4.2.4.5 Ranking of the model

Algorithms performance was compared using Multi-Criteria Decision Analysis (MCDA)/Multi-Criteria Decision Making (MCDM). Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS), an established MCDM method, was used to rank.

Multiple criteria, such as accuracy, sensitivity, precision, G-mean. F-measure, FPR, and MCC were used in TOPSIS [177].

4.2.5 Weighted correlation network analysis

We identified co-expressed gene modules and analyzed the module-trait relationship using the WGCNA package in R [230] . First, the similarity matrix between each pair of feature genes in a specific subtype was measured based on Pearson's correlation coefficient. Next, we transformed the similarity matrix into an adjacency matrix. The soft power β value was calculated for building the proximity matrix so that the co-expression network conformed to a scale-free network based on connectivity. Subsequently, we computed the topological overlap matrix (TOM) and the corresponding dissimilarity (1-TOM) value. Next, a dynamic tree-cut algorithm was implemented to detect gene co-expression modules. The co-expression modules were constructed with a cut height of 0.6, and a minimum module size was set to 10 (transcriptome), 10 (methylome), and 5 (integrated) genes, respectively.

4.2.6 Gene set enrichment and survival analysis

The biological process and functional enrichment analysis was performed using Enrichr [231]. Enrichr uses Fisher exact test to rank terms from gene-set libraries. Terms were considered statistically significantly enriched if the adjusted *p*-value was less than 0.05. The gene list from each positively correlated module was used to examine the enrichment of GO Biological Processes and Molecular Function terms. Overall survival and log-rank test of a coexpressed module was performed using the survminer and survival package in R. We calculated the average expression of all genes in the module. Survival was compared between two groups: patients with higher ($\geq$ 75 percentile) and lower ($\leq$25 percentile) gene expression levels. Furthermore, we performed the overall survival analysis of specific gene using GEPIA [232]. GEPIA performs survival analysis based on The Cancer Genome Atlas (TCGA) gene expression levels and patient clinical information (A detailed description of GEPIA tool was provided in the chapter 3.2.8 in details). Here, the TCGA GBM dataset was used for survival analysis. GEPIA generates Kaplan-Meier plots and performs the log-Rank test to identify the genes associated with patient survival.

4.3 Results

The etiology of GBM is associated with the alteration of transcriptome and methylome patterns. Therefore, the multi-omics approach that combines genome-wide methylation with transcriptome (RNA-seq) data can provide novel insights into biological function and disease mechanisms. In this chapter, first the transcriptome and methylome data were separately analyzed and then integrated both data types were analyzed to classify the GBM subtypes. LASSO feature selection method was used to find the relevant features using transcriptome, methylome and integrated data. Next, ML and DL algorithms were employed to classify the GBM subtypes. WGCNA was performed to observe the association of subtypes and identify the molecular feature. Further, biomarkers were identified in each subtypes of GBM from transciptome, methylome and integrated data.

4.3.1 Classification of GBM subtype using transcriptome

The transcriptome data of the GBM at TCGA contained 20,531 genes. After removing the low expressed genes, a total of 14,125 genes were found expressed in all GBM subtypes, including classical (n=42), mesenchymal (n=55), and proneural (n=39). These genes were taken for further analysis. However, 14,125 genes could not be used as variables for prediction, as the data is high-dimensional, leading to the inaccurate classification of subtypes. Therefore, we performed the LASSO to reduce the dimension of data and subsequently for selecting top key feature genes to enhance the prediction accuracy of the DL and ML model. LASSO performs $L1$ regularization and adds a penalty to the loss function. A total of 201 feature genes were obtained after performing the LASSO analysis. Next, we performed PCA to examine the local structure of data, including 14,125 genes and 201 feature genes. We observed improved subtype-specific separation between patients using 201 feature genes compared to 14,125 genes, indicating that the LASSO feature selection method efficiently extracted most variable features having higher percentage of variability in principal component 1 (PC1) in PCA of 201 feature genes compared to the preprocessed data in the transcriptome data (Figure 4.3A-B). These results indicated that information contained in 201 feature genes could separate the subtype with higher accuracy upon implementing DL and ML algorithms. However, distinct clusters of subtypes were not formed in PCA.

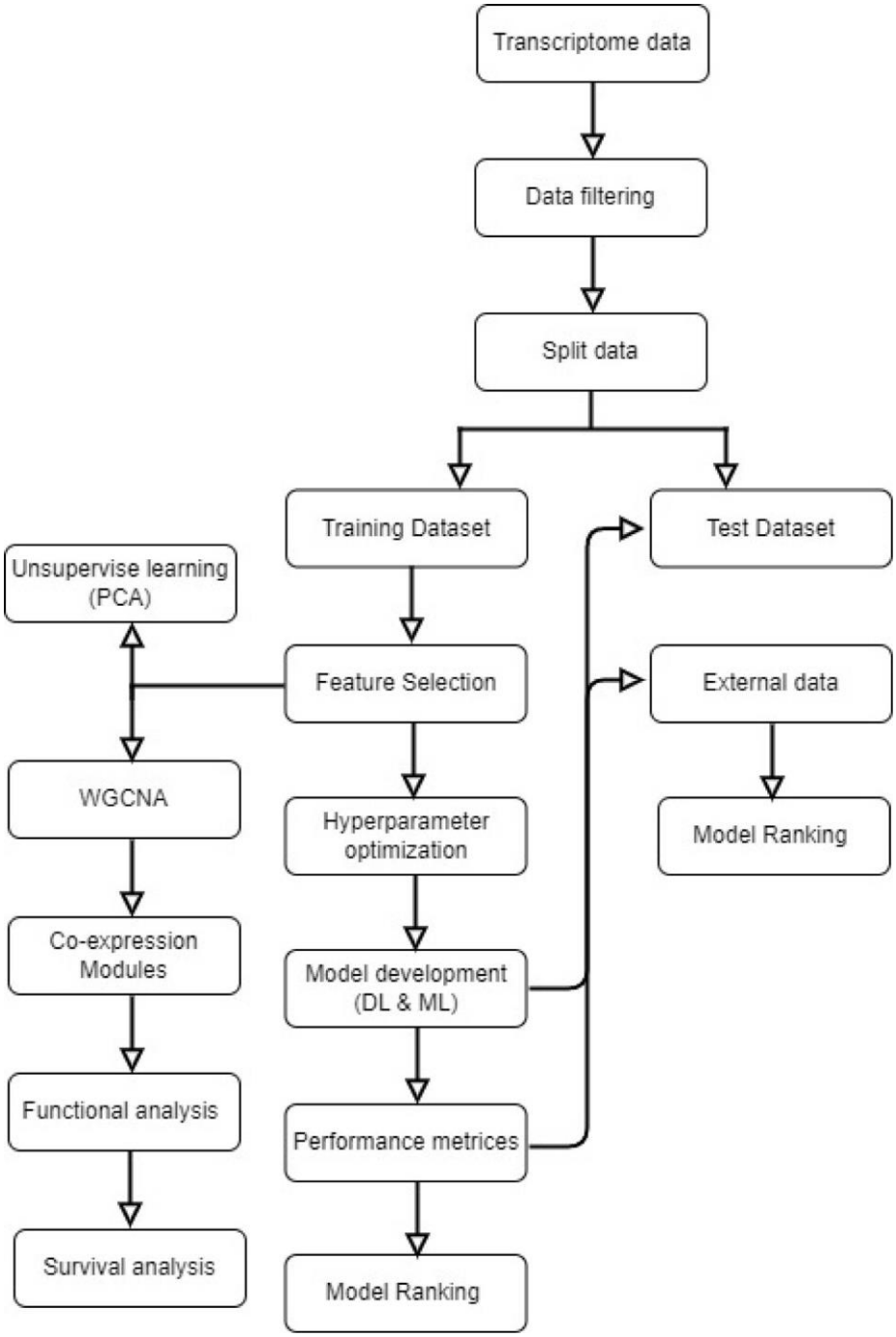

Figure 4.2: *Pipeline of GBM subtype classification using transcriptome data. The flow chart shows DL and ML pipelines using transcriptome data to classify the subtypes.*

PCA plots

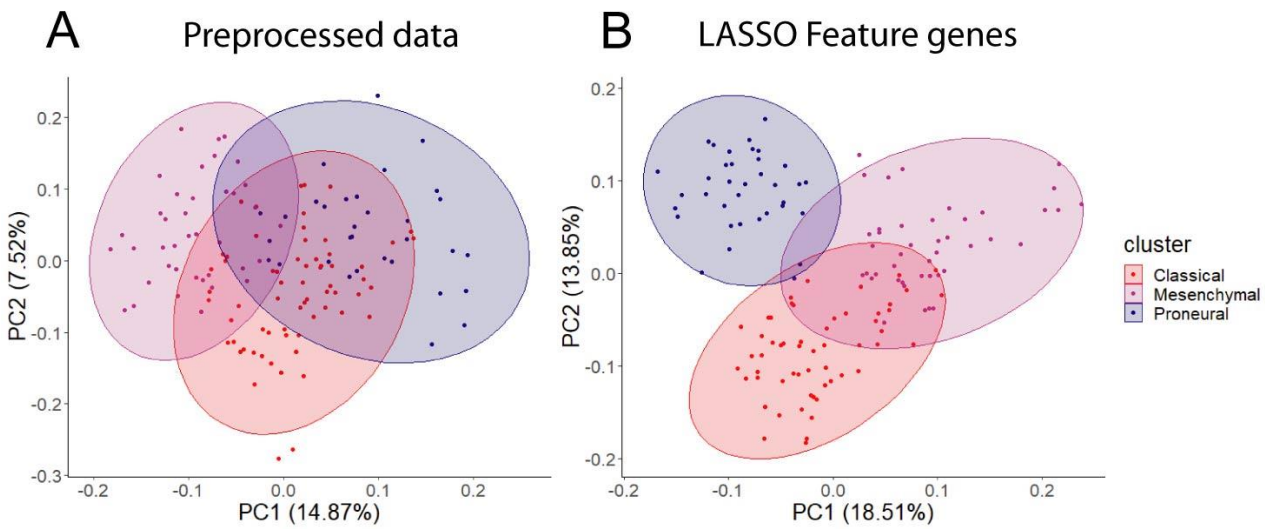

Figure 4.3: *In (A) and (B) PCA plots to visualize the subtype-specific clustering of patients using preprocessed data and feature genes.*

Next, DL (CNN) and ML algorithms (i.e., SVM, KNN, RF, NB, LR) were applied to classify subtypes of GBM using these feature genes as variables. The data was divided into training (70%) and test (30%) datasets. 70% of the data was used for parameter optimization and to assess the performance of each model. The remaining 30% of data was used for independent predictors. Additionally, an external dataset was also used for the final validation of models. In the model training step, 70% of the data was used to obtain the best combination of hyperparameters using the grid search method for each DL and ML model. Next, we performed the stratified k-fold cross-validation (k=10) on the training data using the optimal hyperparameters obtained from the grid search and recorded average performance measures of each model (Table 4.4). The performance of the models was evaluated using average accuracy, recall, precision, F1-score, FPR, GM, and MCC (see Methodology). It is observed that the prediction accuracy of CNN was superior (98.56%) compared to the other ML models. Even standard deviation (±0.03) and FPR (0.01) were minimum in the case of CNN. The MCC score is 0.97 for CNN, which represents the excellent correlation between the observed and predicted classifications. It is observed that the performance of other ML classifiers was also good (accuracy >90%). Therefore, to compare the overall performance, the multiple-criteria decision making (MCDM) was performed using TOPSIS [184]. All performance measures mentioned in

Table 4.4 were considered for the ranking, and CNN topped the overall ranking. To validate this observation, we have performed the classification using two datasets, i.e., 30% data as the test data (or independent data) and an external dataset from GEO (GSE145645). In test data, prediction accuracy (98.33%) of CNN was superior to other ML models and the MCC score was 0.96 (Table 4.5).

Table 4.4: Models performance and ranking for transcriptome data

Method	**Performance measures (Average of 10-fold cross-validation)**							**MCDM Rank**
	Accuracy	**Recall**	**Precision**	**F1-score**	**FPR**	**GM**	**MCC**	
SVM	91.42%(±0.08)	84.48	91.80	85.51	0.06	91.52	0.82	4
KNN	91.03%(±0.06)	85.78	90.59	86.06	0.07	91.44	0.82	5
RF	93.06%(±0.08)	88.52	93.04	89.15	0.05	93.02	0.85	3
NB	90.15% (±0.07)	86.08	87.16	85.38	0.08	90.52	0.80	6
LR	93.32 %(±0.05)	89.47	92.12	89.97	0.05	93.61	0.86	2
CNN	98.56%(±0.03)	97.86	98.36	97.81	0.01	98.64	0.97	1

Table 4.5: Models performance and AUC using test data (transcriptome)

Method	**Performance measures (on test dataset)**							
	Accuracy	**Recall**	**Precision**	**F1-score**	**FPR**	**GM**	**MCC**	**AUC**
SVM	95.12	93.14	92.82	92.57	0.04	95.12	0.89	0.94
KNN	93.46	90.70	90.70	89.80	0.05	93.49	0.85	0.92
RF	91.73	87.14	87.61	87.30	0.06	91.86	0.81	0.91
NB	96.61	95.58	94.80	94.92	0.02	96.74	0.92	0.96
LR	96.61	95.58	94.80	94.92	0.02	96.74	0.92	0.96
CNN	98.33	97.56	97.21	97.28	0.01	98.37	0.96	0.99

It is always desirable to have a highly sensitive and highly specific model for diagnosis. Therefore, we visualized the relationship between sensitivity and specificity using the ROC curve (Figure 4.4A-F). The ROC curve represents the probability of a true positive result or the test's sensitivity against the probability of a false-positive result for a range of different cut-off points. Figure 4.4A shows the area under the ROC curve (AUC) is 0.99 for CNN, indicating that CNN can classify the GBM subtype with high specificity and sensitivity for clinical diagnosis. Additionally, classification with the external dataset also represented a similar outcome, i.e., the performance of CNN was higher (Table 4.6). While validating with the external dataset, 10-fold cross-validation was implemented to calculate the average performance measure and compared the model performance by computing the rank. Furthermore, the classification accuracy of LASSO feature was compared with the features selected using the variance. Gene with higher variance may contain more useful information. To compare the performance with LASSO, we selected the top 201 variable genes according to the degree of variance across all samples. The CNN was performed using the same parameters and 10 fold cross-validation. The average accuracy was 84.02% (±0.08). Therefore, the accuracy of prediction was less than LASSO features (98.56%). Hence, model building to validation, it is observed that the feature genes from LASSO and CNN were the best for subtype classification for the transcriptome data. Therefore, we implemented this framework in subsequent analysis.

ROC plots

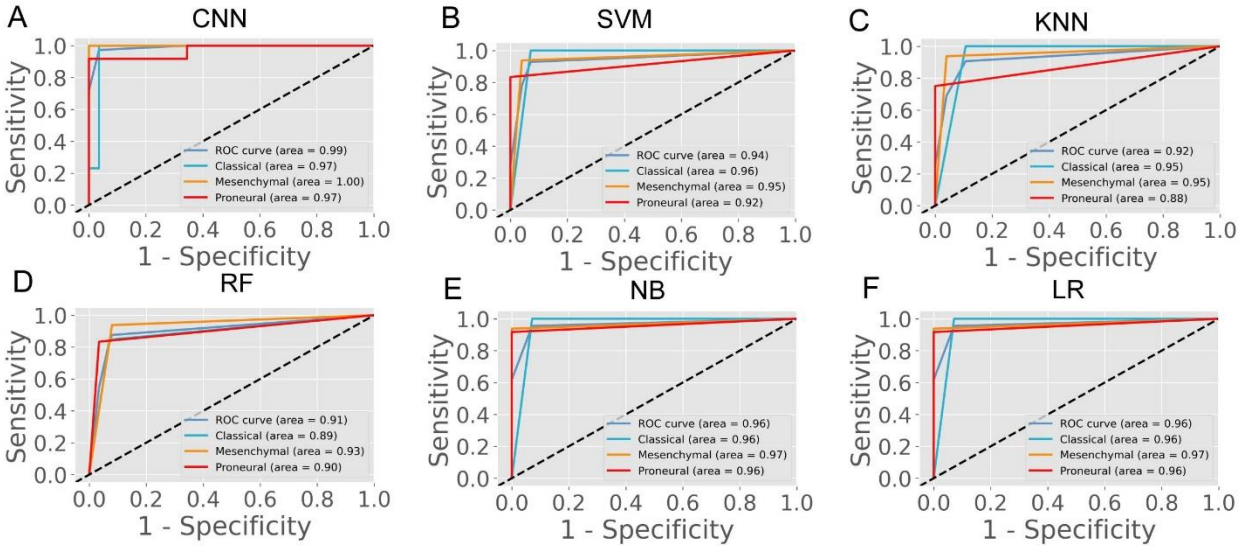

Figure 4.4: *In (**A – F**) ROC of various prediction models. ROC plots were generated using a test dataset.*

Table 4.6: Models performance and ranking for validation data (transcriptome)

Method	Performance measures (Average of 10 fold cross-validation)							MCDM Rank
	Accuracy	Recall	Precision	F1-score	FPR	GM	MCC	
SVM	79.14 %(±0.14)	71.33	63.57	65.68	0.11	84.07	0.71	4
KNN	79.15 %(±0.14)	71.33	63.57	65.68	0.11	84.07	0.71	5
RF	80.57%(±0.22)	71.38	65.75	67.54	0.10	85.85	0.66	3
NB	77.59 %(±0.17)	68.28	61.90	64.02	0.12	82.99	0.68	6
LR	81.20%(±0.15)	74.68	66.01	68.90	0.10	86.44	0.75	2
CNN	92.70 %(±0.12)	90.20	88.77	89.24	0.01	98.25	0.96	1

4.3.2 Classification of GBM subtype using methylome

In the previous section, the GBM subtype were classified using the transcriptome data (or gene expression data), because the alteration of gene expression is a hallmark of oncogenesis. However, the level of gene expression is regulated by DNA methylation. Therefore changes in DNA methylation patterns can play a crucial role in GBM development. Recent studies show that methylation biomarkers are essential for improving and designing cancer therapy [233]. Hence, the information contained in methylation data could possibly help to classify the GBM subtype. The genome-wide methylation or methylome data of 84 GBM patients were retrieved from the UCSC Xena database. The data from the Illumina Infinium HumanMethylation450 platform (450K array) were selected that has 4,85,577 probe sites. In this dataset, the methylation level is estimated using the beta value. The beta value ranges from 0 to 1, representing the ratio of the intensity of the methylated bead type to the combined locus intensity. Thus, higher beta values represent a higher level of DNA methylation, i.e. hypermethylation and lower beta values represent a lower level of DNA methylation, i.e. hypomethylation. The recent reports show that hypermethylation/hypomethylation level in the promoter region (e.g., defined as TSS1500 upstream to TSS200 downstream of TSS, 5'UTR, and first exon; TSS denotes transcription start site.) and gene body determine the gene expression level [234,235]. Therefore, we screened the promoter and gene body methylation data to perform classification because the alteration of methylation levels in these regions can influence the gene expression level and subsequently influence the biological processes [236]. The

CpG sites which include all promoter regions and gene body, were screened for feature selection. By using LASSO, we obtained 498 features CpG sites. Next, the subtype-specific clustering of patients were examined with these 498 features CpG sites using PCA. Results showed that there was slighter mixing among the different subtypes (Figure 4.6). Next, the DL and ML was performed using these 498 CpG sites as variables. We repeated the same methodology as described in the previous section. First, the methylome data were divided into training (70%) and test (30%). The hyperparameters were optimized using the grid search method, and 10-fold cross-validation was performed on the training data. The average performance measures were used to select the top-performing model using MCDM (Figure 4.5). The overall performance of CNN was superior compared to other ML models using methylation data as well (Table 4.7). Next, our observation was validated with the 30% test data set (Table 4.8) and an external data set (GSE128654) (Table 4.9). ROC plots (Figure 4.7A-F) showed that the performance of the CNN (AUC=0.98) was better when compared to other ML models. However, the accuracy value is 89.0%, which is lower than the ML models. The overall performance of CNN on external data is superior (Rank =1, see Table 4.9). These results indicate that CNN is the best classifier for predicting the GBM subtype using DNA methylation data.

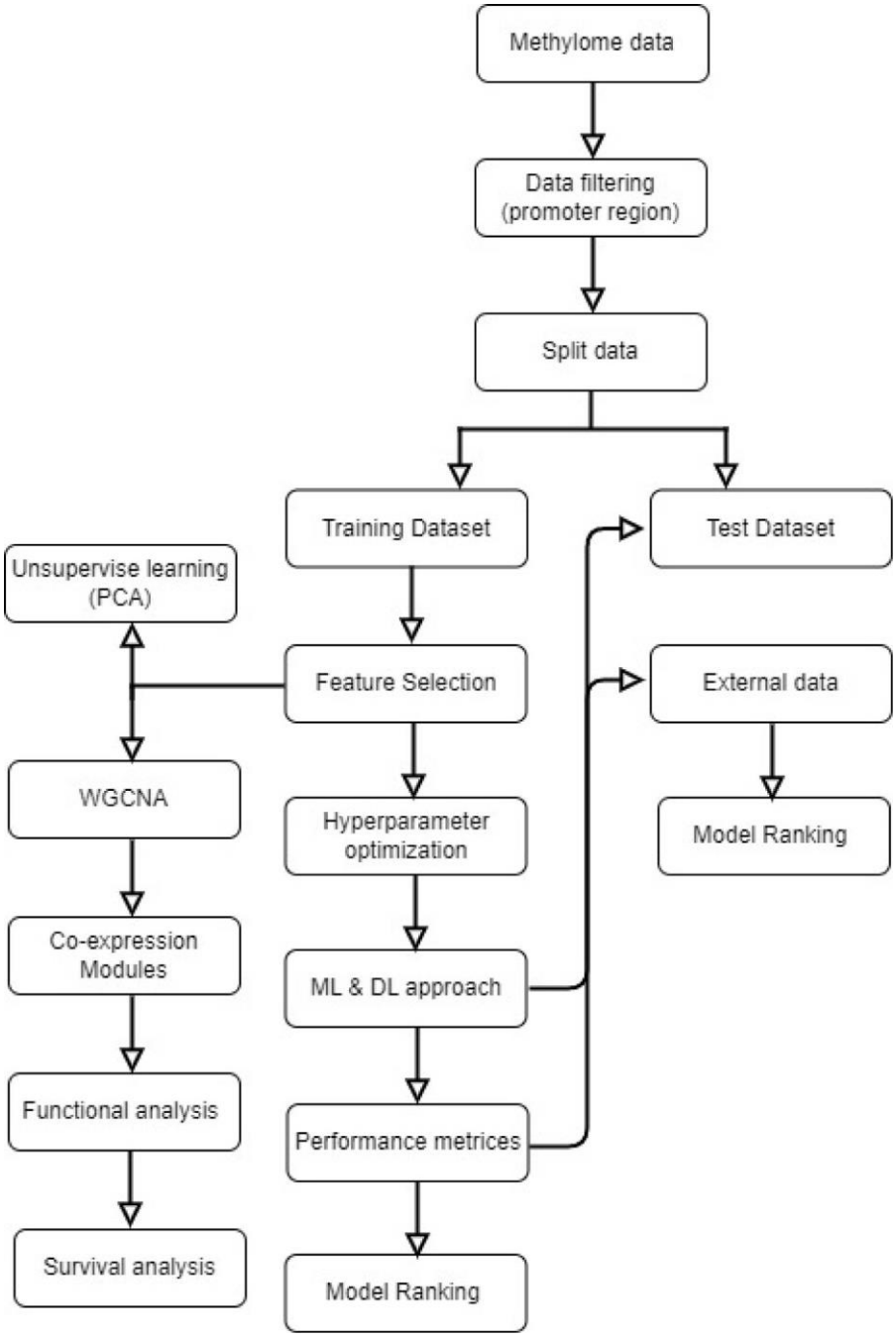

Figure 4.5: *Pipeline of GBM subtype classification using methylome data. The flow chart shows deep-learning and machine-learning pipelines using genome-wide DNA methylation data to classify the subtypes.*

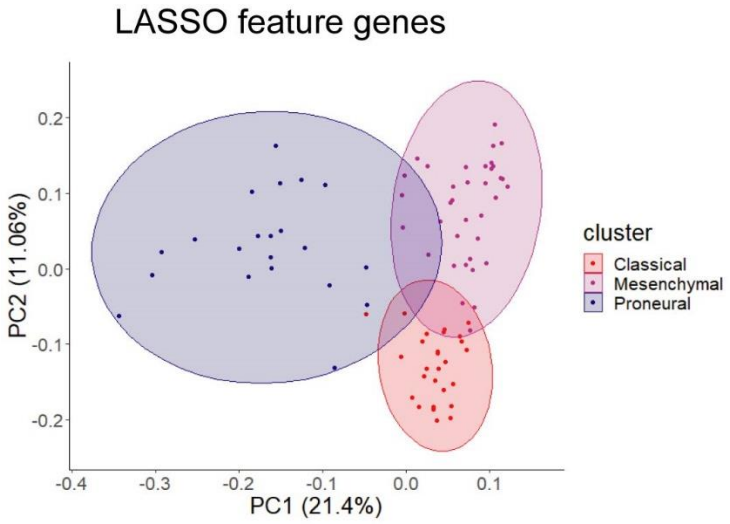

Figure 4.6: *PCA plots to visualize the subtype-specific clustering of the patients from features gene.*

Table 4.7: Models performance and ranking for methylation data

Method	**Performance measures (Average of 10-fold cross-validation)**							**MCDM Rank**
	Accuracy	**Recall**	**Precision**	**F1-score**	**FPR**	**GM**	**MCC**	
SVM	90.61%(±0.09)	86.40	87.67	84.49	0.07	90.55	0.81	4
KNN	90.72 %(±0.12)	85.86	88.10	84.90	0.07	90.36	0.81	5
RF	91.03 %(±0.10)	86.92	89.74	86.33	0.06	90.81	0.82	3
NB	92.34 %(±0.08)	88.85	92.63	88.46	0.05	92.03	0.84	2
LR	89.84%(±0.11)	83.71	82.70	81.80	0.08	89.46	0.78	6
CNN	97.54%(±0.05)	96.77	97.71	96.47	0.01	97.47	0.95	1

Table 4.8: Models performance and AUC from test data (methylation)

Method	**Performance measures (on test dataset)**							
	Accuracy	**Recall**	**Precision**	**F1-score**	**FPR**	**GM**	**MCC**	**AUC**
SVM	97.19	95.73	96.85	96.09	0.02	97.42	0.94	0.97
KNN	89.94	84.15	86.98	84.86	0.07	89.73	0.77	0.88
RF	89.50	84.96	84.62	84.62	0.08	89.73	0.76	0.89
NB	95.12	92.26	94.02	92.55	0.02	94.85	0.89	0.95
LR	94.82	92.26	93.17	92.48	0.03	94.85	0.88	0.95
CNN	89.50	85.38	86.54	84.55	0.08	89.73	0.78	0.98

ROC plots

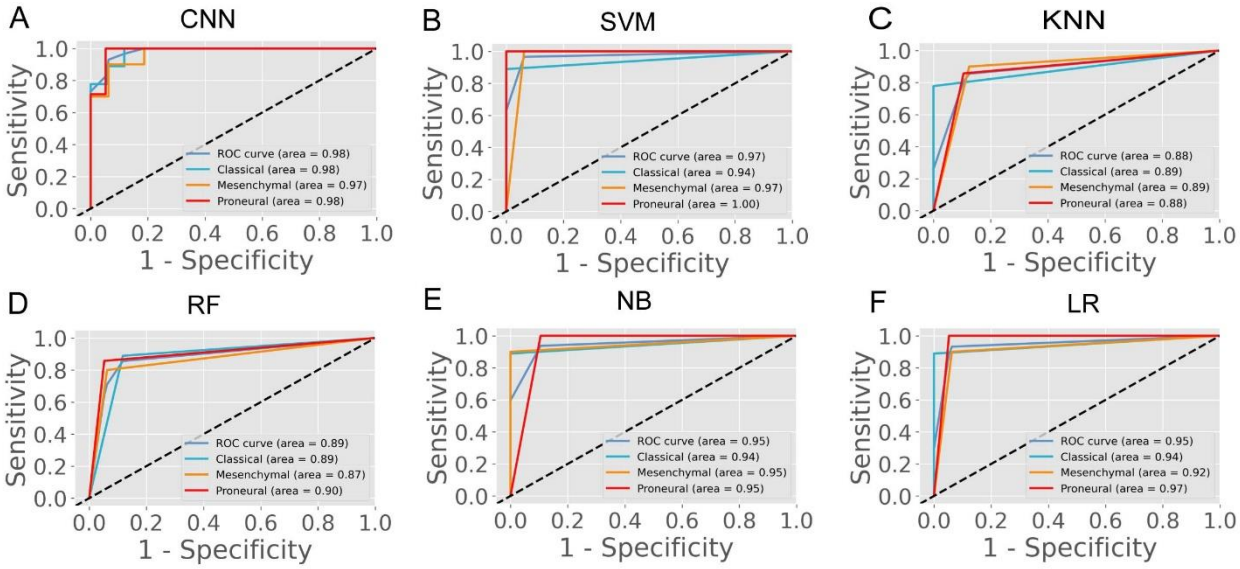

Figure 4.7: *In (A-F) ROC of various prediction models. ROC plots were generated using the test dataset.*

Table 4.9: Models performance and ranking for external data (methylation)

Method	Performance measures (Average of 10-fold cross-validation)							MCDM Rank
	Accuracy	Recall	Precision	F1-score	FPR	GM	MCC	
SVM	82.42%(±0.23)	76.65	73.58	74.60	0.09	85.26	0.76	4
KNN	79.09 %(±0.20)	68.00	63.19	64.22	0.13	81.49	0.66	6
RF	82.81 %(±0.16)	76.27	70.31	72.29	0.08	88.19	0.78	3
NB	81.52%(±0.15)	71.46	65.50	66.91	0.11	83.45	0.71	5
LR	87.42%(±0.17)	84.34	81.08	82.17	0.05	92.92	0.86	2
CNN	91.91 %(±0.13)	90.50	89.15	89.60	0.01	97.63	0.96	1

4.3.3 Classification of GBM subtype by integrating the methylation and transcriptome data.

There are several studies where only one type of "omics" data was used, such as either gene expression or methylation data to identify the biomarkers or classify the cancers [237,238]. However, DNA methylation and gene expression are the integrated processes that determine cellular fate [239]. The perturbation of gene expression in many human cancers is due to the change of methylation pattern [230]. Hence, integrating these strongly interlink cellular processes and subsequent analysis could facilitate finding a more effective diagnostic option [69]. The patients having both transcriptome and methylome data were selected for data integration. Next, the gene and methylation sites were screened based on z-score, i.e. $z>1$ and $z<-1$. Z-score greater than 1 or less than -1 indicates the expression and methylation is greater or less than the population mean. We identified common genes whose expression and methylation both are $z>1$ or $z<-1$ in each subtype. Next, all these genes (n=4231) were combined and used their gene expression level to find the most variable features (n=75) using LASSO. We observed that 75 feature genes form the distinct subtype-specific clusters with PCA (Figure 4.9). Compared to previous features from transcriptome and methylome data, the feature genes of integrated data significantly improved the clustering of the GBM subtype. Next, CNN was implemented using these feature genes and compared CNN performance with the other five ML algorithms (Figure 4.8).

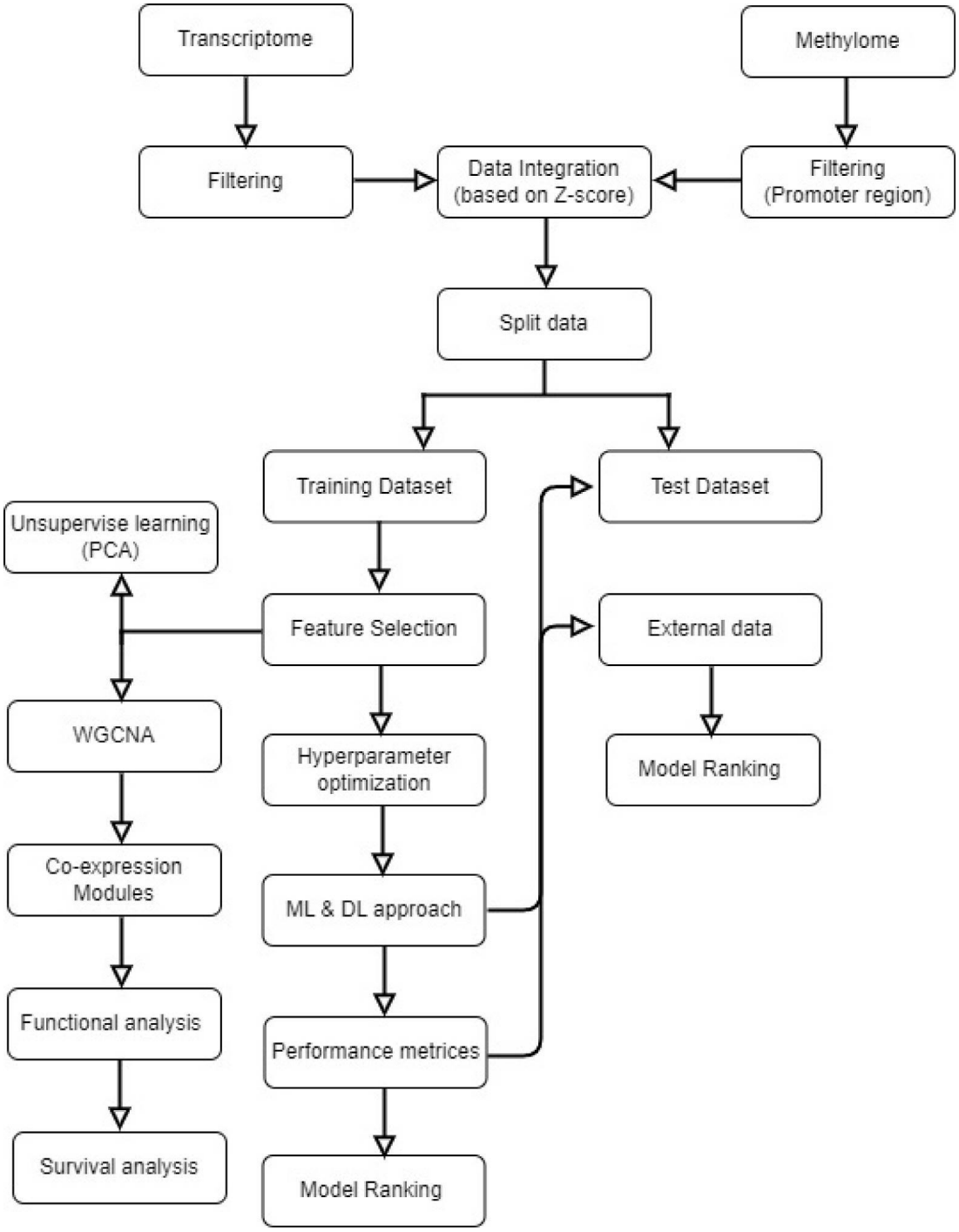

Figure 4.8: *Pipeline of GBM subtype classification using integrated data. The flow chart shows deep-learning and machine-learning pipelines using the integrated data of transcriptome and methylome to classify the subtypes.*

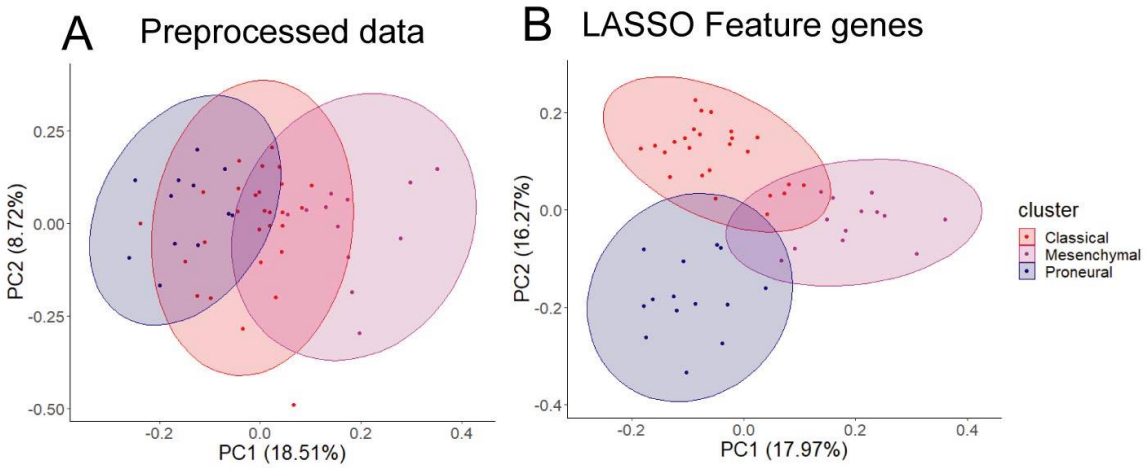

Figure 4.9: *In (A) and (B), PCA plots to visualize the subtype-specific clustering of patient from features gene.*

In this case, the CNN performance was also ranked on top (Table 4.10). Furthermore, we validated the model with 30% test data (Tables 4.11) and external data (Table 4.12). ROC plots generated using test data explain the decent performance of CNN (AUC=0.91 and accuracy=87.50%) (Figure 4.10A-F). The validation with external data showed that CNN was the top performer (accuracy=94.48%) for classification (Table 4.12). It can be concluded that in all three types of analysis, CNN efficiently classified the GBM subtypes. However, the features from integrated data specifically cluster the subtype of GBM with PCA. Moreover, the consistent all-around performance of CNN proves that CNN can be used as a computational tool for the clinical diagnosis GBM subtype.

Table 4.10: Models performance and ranking using integrated data

Method	**Performance measures (Average of 10-fold cross-validation)**							**MCDM Rank**
	Accuracy	**Recall**	**Precision**	**F1-score**	**FPR**	**GM**	**MCC**	
SVM	89.94 %(±0.10)	86.47	81.11	81.65	0.07	90.02	0.82	5
KNN	91.87%(±0.13)	88.35	82.68	84.57	0.06	91.81	0.84	3
RF	93.67%(±0.10)	88.70	84.63	86.06	0.04	93.52	0.89	2
NB	89.95%(±0.14)	83.16	77.12	79.14	0.08	89.43	0.79	6
LR	92.18%(±0.10)	87.10	81.38	83.43	0.06	91.77	0.85	4
CNN	98.20%(±0.05)	98.44	97.97	97.60	0.01	98.25	0.97	1

Table 4.11: Models performance and AUC using test data (integrated)

Method	**Performance measures (on test dataset)**							
	Accuracy	**Recall**	**Precision**	**F1-score**	**FPR**	**GM**	**MCC**	**AUC**
SVM	90.63	82.50	93.06	85.16	0.06	91.48	0.82	0.9
KNN	88.28	84.82	81.25	82.56	0.09	87.35	0.71	0.86
RF	87.50	80.54	80.54	80.54	0.09	87.50	0.71	0.85
NB	95.70	93.75	92.71	92.45	0.04	95.79	0.90	0.94
LR	95.70	93.75	92.71	92.45	0.04	95.79	0.90	0.94
CNN	87.50	78.75	84.31	79.01	0.10	87.50	0.72	0.91

ROC plots

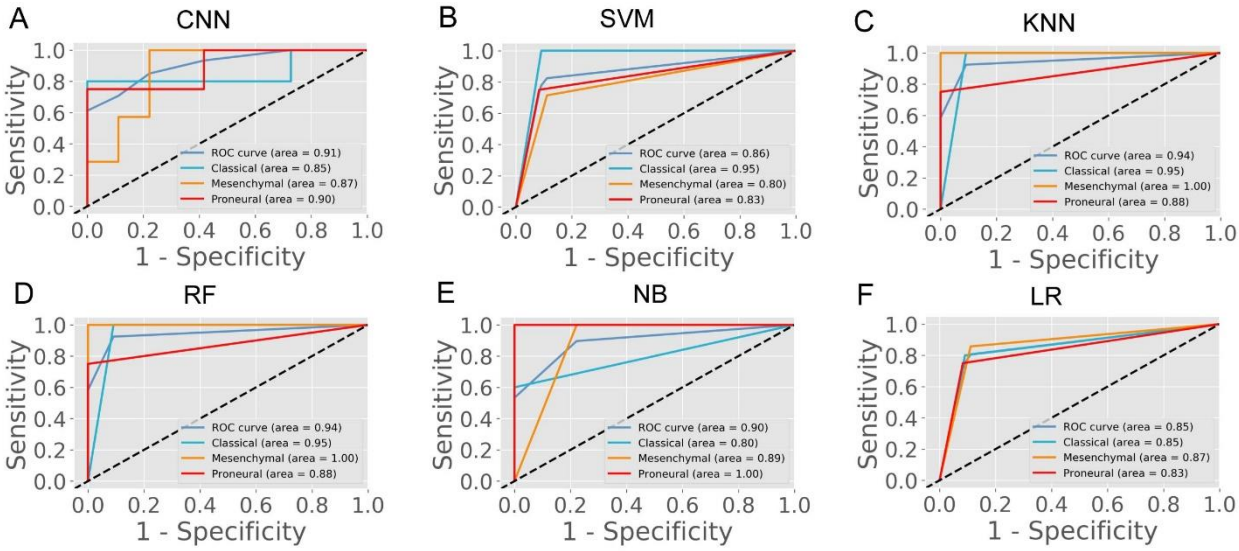

Figure 4.10: *In (A-F), ROC of various prediction models. ROC plots were generated using the test dataset.*

Table 4.12: Models performance and ranking for external data (transcriptome)

Method	Performance measures (Average of 10-fold cross-validation)							MCMC Rank
	Accuracy	Recall	Precision	F1-score	FPR	GM	MCC	
SVM	63.15%(±0.12)	46.43	35.70	37.89	0.22	68.38	0.38	6
KNN	67.08%(±0.17)	49.56	38.83	42.39	0.20	72.31	0.39	5
RF	80.00%(±0.19)	72.24	66.21	67.70	0.09	85.81	0.73	2
NB	66.14%(±0.17)	55.59	45.69	48.47	0.22	71.35	0.43	4
LR	70.74%(±0.10)	49.26	37.11	41.14	0.16	75.89	0.48	3
CNN	94.48%(±0.11)	94.48	94.48	94.48	0	1	1	1

4.3.4 The biological relevance of features and identification of biomarkers

In the preceding steps, features were extracted from large-scale transcriptome and methylome datasets to develop the predictive tool for subtype identification. It is observed that selected features from each type of data have excellent separability power, and therefore we achieved classification accuracy > 90% in every case. This indicates that any subset of these features is probably associated with a particular subtype (or phenotype). Therefore, further analysis of these features genes can link the genotype to phenotype. Weighted gene co-expression network analysis (WGCNA) was performed to understand genotype-to-phenotype relationships. WGCNA can find the module of highly correlated genes and their association with a specific subtype of GBM [230]. The co-expression module was constructed using the feature genes expression from transcriptome, methylome, and integrated data and examined their association with specific subtypes. To find the co-expression module of feature methylation sites, we mapped the methylation site to gene name and extracted the gene expression data to construct co-expression modules. To construct the co-expression modules, the soft threshold, β (β= 4, 6, and 5 for transcriptome, methylome, and integrated data, respectively) was determined based on scale independence and mean connectivity (Appendix Figure II. 1). We then merged modules with similarities above 0.6 for all three types of data. Finally, the dynamic tree cut showed a gene cluster dendrogram containing 3, 6, 5 co-expression models in the features of transcriptome, methylome, and integrated data, respectively (Figure 4.11.A, Figure 4.12.A, Figure 4.13.A). To understand the genotype-phenotype relationship, the

module-trait relationship plot was generated. We found distinct patterns of association between modules and subtypes (Figure 4.11.B, Figure 4.12.B, Figure 4.13.B). Results showed that the blue module (Figure 4.11 B) was significantly and positively associated with the proneural subtype (r = 0.53, p = 4E-11). In contrast, it was negatively associated with mesenchymal (r = -0.73, p = 2E-23), and weakly correlated with classical subtype (r = 0.25, p = 0.004). Similarly, we found a distinct pattern of association between other modules (i.e., brown and turquoise) and subtypes (Figure 4.11 B). We observed the same in the features from methylome and integrated data. In methylome (Figure 4.12 B), brown module significantly and positively associated with only proneural subtype (r = 0.33, p = 0.02). The green module is positively associated with classical (r = 0.32, p = 0.03) and negatively associated with proneural (r = -0.46, p = 9E-04). The Blue module is strongly and positively correlated with mesenchymal subtype (r = 0.55, p = 4E-05), whereas it was negatively associated with proneural (r = -0.6, p =5E-06). However, the feature from integrated data showed a more specific module-subtype association. At least one module was strongly and positively correlated with a specific subtype. The red (r = 0.64, p = 3E-07), turquoise (r = 0.66, p = 8E-08) and blue (r = 0.56, p =1E-05) were explicitly and positively associated with classical, mesenchymal, and proneural, respectively (Figure 4.13B). The module-trait relationship analysis indicated that integration of transcriptome and methylome resulted in subsets of features strongly correlated with a particular subtype of GBM. Probably, the integrated datasets are mechanistically more relevant as the methylation, and gene expression are integrated cellular processes. Next, the gene set enrichment analysis (GSEA) was performed, i.e., GO Biological Process (BP) and Molecular Function (MF), using Enrichr to understand the biological relevance of each data type's top three positively correlated modules [229]. It was observed that modules were significantly (adjusted $p < 0.05$) associated with several BP and MF that are linked to the oncogenesis. For example, the turquoise module from transcriptome data in the classical subtype is involved in the RIG-I signaling pathway that elicits RIG-I-like receptors' expression and activity (RLRs) (Figure 4.11 C). These receptors stimulate both innate and adaptive immune responses against tumor antigens and promote the apoptosis of cancer cells [240]. In contrast, the brown module associated with the mesenchymal subtype (leukocyte adhesion to vascular endothelial cell) may be linked to the GBM-associated with endothelial cell, that is resistant to cytotoxic drugs, and also less apoptotic than healthy cells [241] (Figure 4.11 C). Phosphatidylinositol 3 phosphate activity enriched in the turquoise module, solute proton symporter activity in the brown module, and syndecan binding in the blue module are associated with higher tumor grades and poor

prognosis in GBM [242] (Figure 4.11 D). Similarly, it is observed that the blue module in the mesenchymal and brown module in the proneural are linked to positive regulation of GTPase activity and positive regulation of phosphorylation in methylome data (Figure 4.12 C) These processes are signatures of GBM formation and progression [243]. Even molecular functions of several co-expression modules are involved in tumorigenesis, like phosphatidylinositol 3, 4, 5 triphosphate binding enriched in the green module deregulates many key signaling pathways involving growth, proliferation, survival, and apoptosis in GBM [244] (Figure 4.12 D). Furthermore, endopeptidase inhibitor activity, GABA receptor activity enriched in blue and brown modules, respectively, are predominant events in GBM [245,246] (Figure 4.12 D) The gene co-expressed modules in integrated data, i.e., and turquoise module (mesenchymal) involved with negative regulation of T cell activation and proliferation is one of the signatures of GBM [247]. The MF of the same module shows it is associated with gap junction channel activity involved in cell communication, which is also linked to GBM [248] (Figure 4.13 C, D).

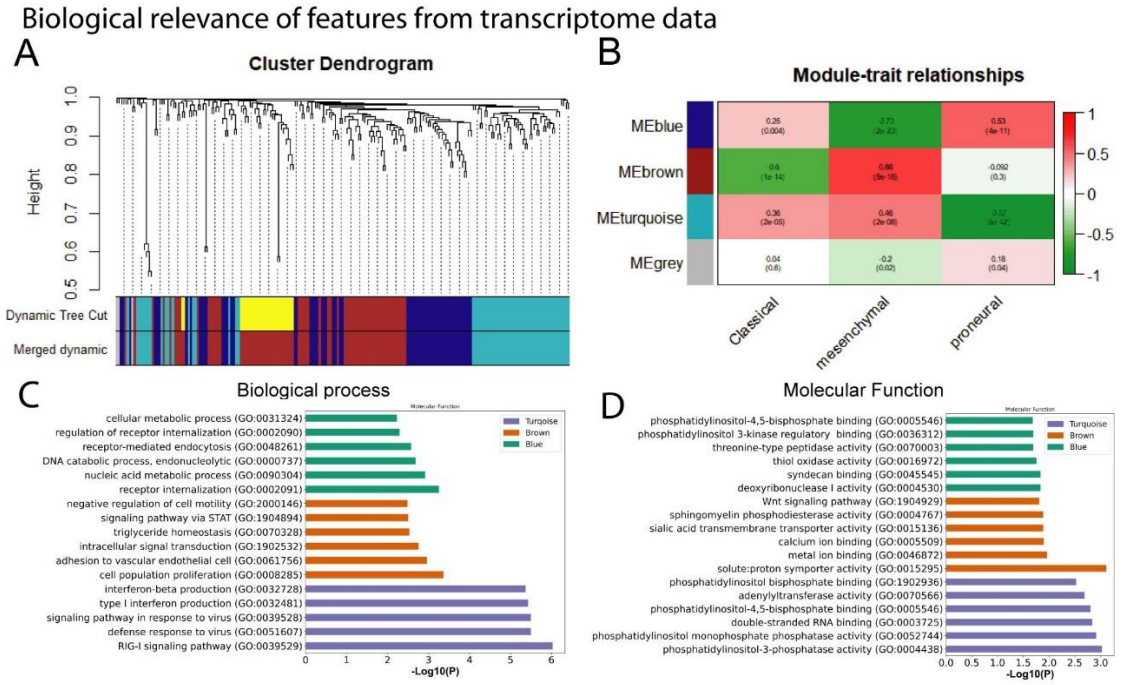

Figure 4.11: *Weighted gene co-expression network analysis and gene set enrichment of feature used for model building.* ***(A)*** *co-expression gene module,* ***(B)*** *module-trait relationship,* ***(C)*** *biological process, and* ***(D)*** *molecular function of feature from transcriptome data.*

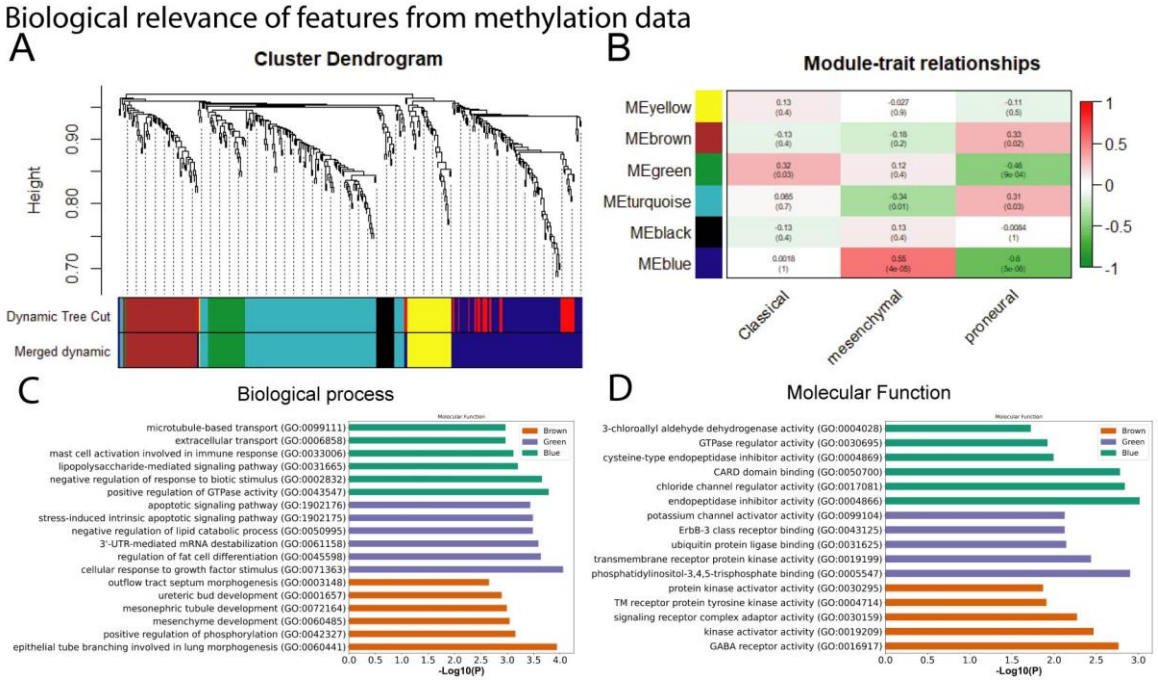

Figure 4.12: *Weighted gene co-expression network analysis and gene set enrichment of feature used for model building.* ***(A)*** *co-expression gene module,* ***(B)*** *module-trait relationship,* ***(C)*** *biological process, and* ***(D)*** *molecular function of feature from methylome data.*

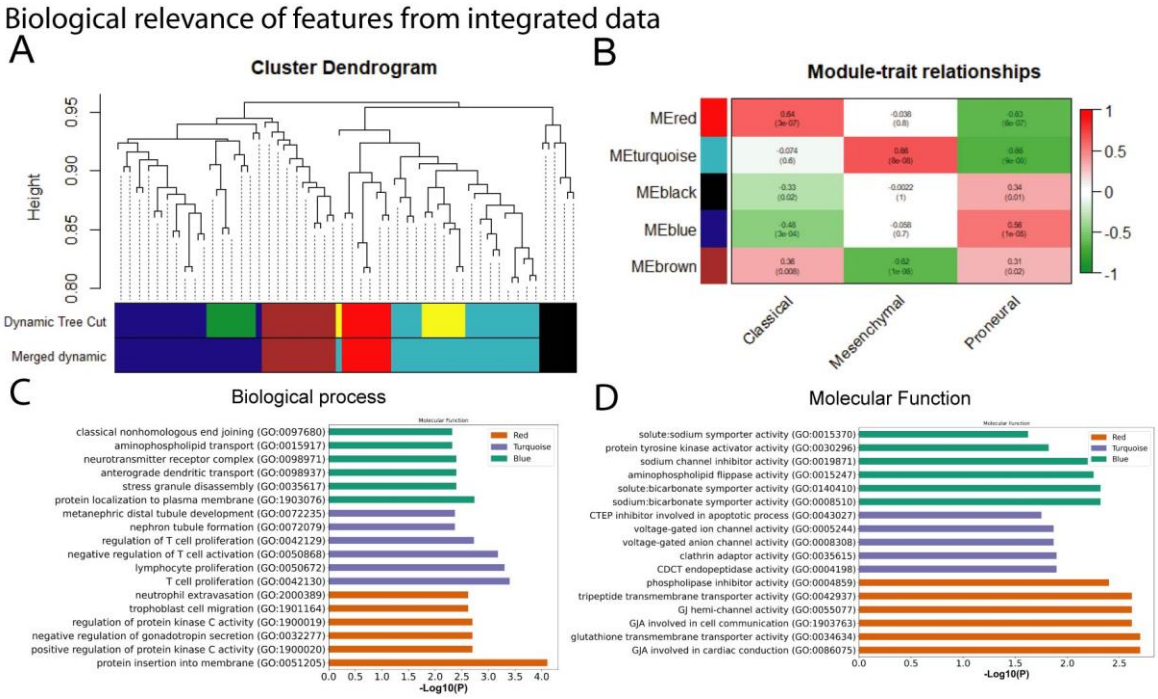

Figure 4.13: *Weighted gene co-expression network analysis and gene set enrichment of feature used for model building.* ***(A)*** *co-expression gene module,* ***(B)*** *module-trait relationship,* ***(C)*** *GO Biological Process term analysis, and* ***(D)*** *GO Molecular Function of feature from integrated data.*

Our results show that most of the positively correlated modules in GBM subtypes were involved in several BP and MF. Besides, many of these BP and MF are involved in oncogenic processes. This shows a possibility of identifying these modules' genes as cancer biomarkers for therapy or diagnosis. We performed survival analysis of positively correlated modules (Appendix Figure II.2). The turquoise module in the integrated feature is significantly (log-rank test, p=0.029) associated with the patient survival. Hence, we performed survival analysis of all genes separately present in these modules using GEPIA web tools (Figure 4.14 and appendix Figure II.3.) We found several genes that were present in the co-expression module and were also associated with the patient's survival (log-rank test, p<0.05). The higher expression of most of the genes was associated with worse survival of the patients, except DUOX1 (Figure 4.14O) and FOXN2 (Appendix Figure II.3). However, higher or lower expression of genes associated with worse survival can be considered biomarkers [249,250]. Furthermore, several experimental articles confirm the involvement of these genes in GBM formation and progression. For example, CCDC8, CLDN1, JMJD8, PTRF, RNF135, and SNX21 in classical [251–256] (Figure 4.14 A to F); GCNT1, RAB38, HLX, ZDHHC12, SRCRB4D (SSC4D), GNB2 and LETM2 in mesenchymal [257–263] (Figure4.14 G to M); and TOLLIP, DUOX1 [264,265] in proneural (Figure 4.14 N to O) are linked to GBM patients' survival. The association of genes from the modules with patient survival shows the possibility to identify them as subtype-specific prognostic biomarkers. We also observed that expression pattern of survival associated genes varied across the subtype (Appendix Figure II.4). Further, we illustrated with gene enrichment analysis that their biological process and molecular functions are also linked to oncogenic events. Therefore, these findings confirm the clinical validity of our models and can provide insight into the complex regulatory processes in different subtypes of GBM.

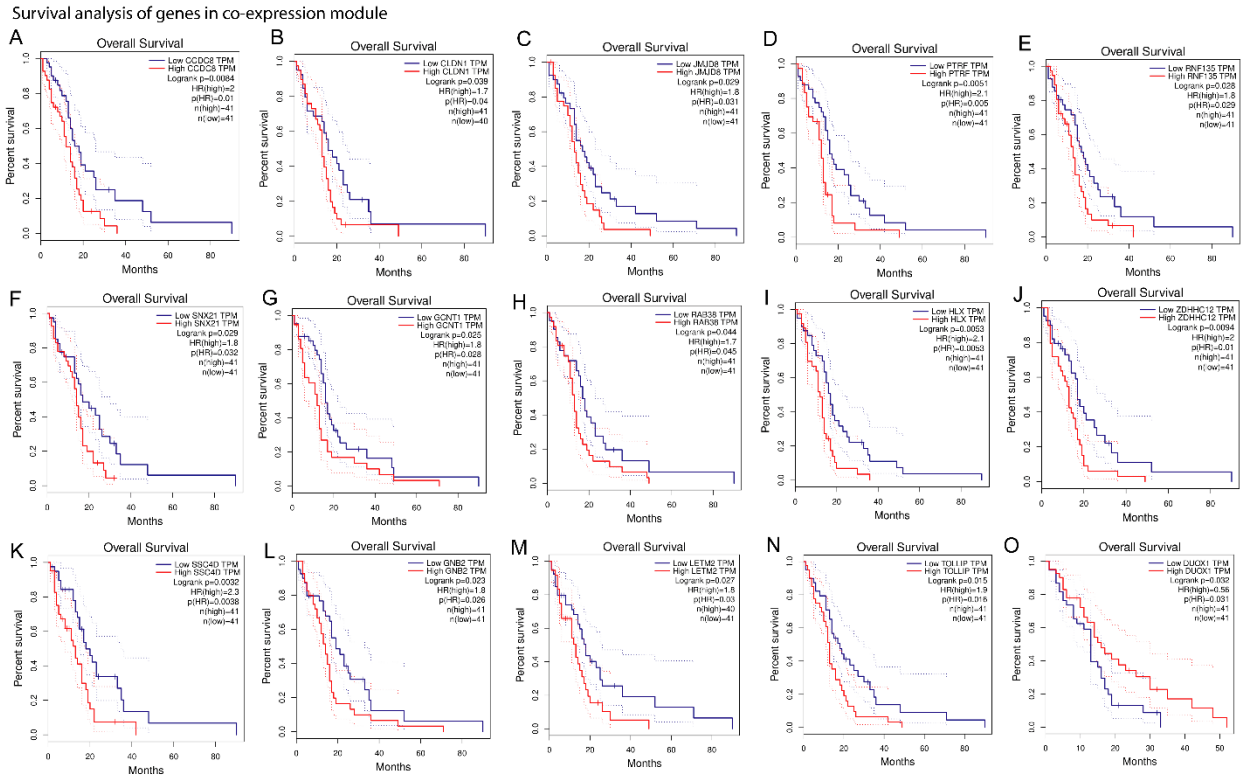

Figure 4.14: *Survival analysis of gene present in co-expression module.* ***(A-F)****, Kaplan-Meier plots of genes from positively associated modules with classical subtype.* ***(G-M)****, Kaplan-Meier plots of gene from positively associated modules with mesenchymal subtype.* ***(N-O)****, Kaplan-Meier plots of gene from positively associated modules with proneural subtype. Overall survival was analyzed based on the clinical information of the patients from TCGA and quartile method of 75 % cut-off of higher and 25% cut-off of lower limit (Extended version of this figure is provided in appendix figure II.3.).*

4.4 Discussion

This chapter indicates that DL and ML can be powerful tools for finding patterns in large-scale genetic and epigenetic data sets related to human cancer. In general, efficient DL and ML tools work like a ‘black -box’; researchers or clinicians may not be confident in diagnosing or classifying cancer patients using these approaches. However, if the basis of classification is biologically relevant and has higher accuracy, the diagnosis and patient management will be more assured and systematic. Here a biologically relevant DL and ML-based framework was presented to classify the subtype of GBM to increase accuracy in diagnosis; in turn, it can lead to better patient management. The previous studies tried to develop the subtype classification model for GBM using either imaging data or single type of omics data, however, these models

exhibited lower accuracy compared to our framework [266,267]. Additionally, models were mainly developed for binary classification to identify healthy and cancer patients. However, two types of high-throughput data were used, i.e., transcriptome and methylome; integrated forms of these data were explored to develop the classification framework. Most importantly, we have successfully separated three subtypes, classical, mesenchymal, and proneural of GBM. Although **we have dealt with multi-class classification problems, we still achieved classification accuracy >90%. DL and ML techniques were also compared to identify the most suitable method for interpreting the transcriptome, methylome, and integrated data. DL method, i.e., CNN outperforms other ML models.** Using CNN, we were able to classify the tumor into the correct subtype from the test and external cohort. **We observed that overall classification performance was higher using the transcriptome and integrated data than the methylome data.**

Another significant finding of this chapter is the biological relevance of features and the identification of subtype-specific prognostic biomarkers. To find the association of features genes with specific subtypes, WGCNA was performed. **The gene co-expression module-subtype relation analysis revealed how a subset of features is strongly and positively correlated with a particular subtype of GBM**. In addition to that, the gene set enrichment analysis revealed that all positively correlated modules are biologically relevant, even those that are linked to oncogenic processes. Among all data types, a strong module-trait relationship was observed in feature genes from integrated data. Furthermore, several genes present in these co-expressed modules were identified, which were linked to patient survival. **Our study explained how the features genes from the DL/ML framework could be used to find the subtype-specific biomarkers.** Good agreement was found when comparing prognostic markers from this work against published experimental data. The feature genes of this study and CNN can provide assured and clinically relevant deep learning-based diagnostic tools for the proper treatment of GBM patients. Furthermore, the results of this work will elucidate and shed light on the understanding of genotype-phenotype relationships of the GBM subtype.

CHAPTER 5

OBJECTIVE 3

Chapter 5: Objective 3
Implementation of a deep learning embedding system for multi-omics data integration for the subtyping of Glioma

5.1 Introduction

The majority of deep learning models are commonly perceived as black boxes, wherein they generate precise predictions without offering any accompanying explanations. One of the primary constraints associated with the utilisation of deep learning in the field of cancer research pertains to its lack of interpretability, which limit their application in biomedical settings. To increase clinical applicability, the classification model for glioma subtyping using machine learning (ML) and deep learning (DL) techniques should be biologically informed. Because biologically relevant approaches were employed to identify disease-specific biomarkers that exhibit associations with specific disease phenotype. Furthermore, it is a well-established that perturbations in various genomics layers can lead to cancer. The integration of these genomics layers with an AI-based model can aid in capturing the unique pattern necessary for developing accurate subtype classification models [268]. However, employing of data from different genomics layer can increase the dimensionality of the data, hence reducing of dimension is crucial step to develop the efficient model. Deep learning (DL) techniques such as autoencoder effectively integrate the genomics data and simultaneously reduce the dimension. The integration of multi-omics data across several levels can yields a more comprehensive understanding of disease specific alterations, and identification of cancer subtypes. Moreover, this approach elucidates the interconnections among diverse omics data modalities pertaining to a certain disease.

This chapter presents the development of deep-neural network-based framework, called DeepAutoGlioma for integrating the transcriptome and methylome and subsequently classified the subtypes of LGG and GBM. Transcriptome and methylome data of glioma patients were pre-processed and differentially expressed features from both datasets were identified. Subsequently, a cox regression analysis determined genes and CpGs associated with survival. Gene set enrichment analysis was carried out to examine the biological significance of the features. Further, CpG and gene pairs were identified by mapping them in the promoter region of corresponding genes. The methylation and gene expression levels of these CpGs and genes

were embedded in a lower dimensional space with an autoencoder. Next, ANN and CNN were used to classify subtypes using the latent features from embedding space. The framework is called DeepAutoGlioma. The present work introduces a new way for subtyping brain cancer, and it is believe that this research will shed light on the DL-based clinical support system for accurate disease prediction using multi-omic data.

5.2 Methodology

5.2.1 Data Collection and Preprocessing

The methylome (Illumina Infinium HumanMethylation450 platform), and transcriptome (RNA-seq) data of TCGA were retrieved from UCSC Xena (https://xena.ucsc.edu) [165]. log2 (RSEM + 1) values for gene expression and beta-values for methylation levels were considered for analysis. Here, RSEM stands for RNA-Seq by Expectation Maximization. Next, low-expressed genes were filtered out of the transcriptome data [log2 (RSEM +1) <0.1 in 90% sample]. Patients with both a transcriptome and methylome profile were considered for analysis. GBM patients (n = 52) were divided into three groups based on their clinical information: classical (n = 16), mesenchymal (n = 22), and proneural (n = 14). Similarly, the LGG patients (n = 281) were divided into three groups based on cancer subtype, i.e., astrocytoma (*n* = 96), oligoastrocytoma (*n* = 75), and oligodendroglioma (*n* = 110). The external data set was obtained from the Gene Expression Omnibus (GEO) repository. The subtyping of LGG was validated using the GSE74462, GSE43378 (gene expression data), and GSE129477 (DNA methylation data). The subtyping of GBM was validated using the gene expression data from GSE145645 and the DNA methylation data from GSE128654.

5.2.2 Identification of differentially expressed genes and differentially methylated regions

DEGs and DMRs were identified by z-score. The categorization of genes with high and low expression levels, as well as CpG sites with hyper- and hypo-methylation, was performed using the Z-score. This approach was used due to the unavailability of healthy patient data for both the transcriptome and methylome. The following formula was used to determine the Z-score for each gene or CpG site in a certain subtype:

$$Z - score = \frac{\bar{x} - \mu}{\sigma}$$

Here, $\bar{x}$ denotes the subtype-specific average gene expression or methylation level, whereas μ and σ stand for the population mean and population standard deviation, respectively. For each subtype of LGG and GBM, Z-score > 1 for higher expression and hypermethylation and Z-score < -1 for lower expression and hypomethylation were used. Then, considering that differential methylation in the promoter regions may affect the related gene's expression, the higher and lower-expressed genes whose promoter regions were differentially methylated were screened. Finally, genes with differential expression and methylated promoter regions were used for further analysis.

5.2.3 Construction of univariate Cox regression models and survival analysis

Univariate Cox regression analysis was implemented to build the prognostic risk-score model for a particular gene and CpG site [269]. Univariate Cox regression analysis was performed using the survminer and survival package in R. The *p*-value < 0.05 was considered the significant association of a gene or CpG site with patients' overall survival (OS).

$$h(t) = h_0(t) \times \exp\{b_1 x_1 + b_2 x_2 + \ldots\ldots + b_p x_p\}$$

Where t is survival time, h(t) is the hazard function determined by a set of covariates (x_1 , x_2, ……, x_p) for genes or methylation sites, b_1 , b_2, ……, b_p are the coefficients of regression, h_1 is baseline hazard.

5.2.4 Mapping and integration of methylation and gene expression data

CpG ids and genes were mapped through the promoter region. The TSS1500, TSS200, the first exon, and the 5′ UTR were considered promoters of a gene. If both gene expression and methylation levels at the promoter alter (i.e., DEGs and DMRs), then the CpG-gene pairings were subjected to screening. Next, the construction of methylation and gene expression matrices

was performed utilizing these CpGs and genes. These matrices were then utilized as input for an autoencoder, which consisted of two separate layers.

5.2.5 Biological processes and pathway enrichment analysis

The biological processes and pathway enrichment were analyzed using the Metascape tool (A detailed description of Metascape tool was provided in the chapter 3.2.9 in details) [178]. Enrichment analysis was performed using the following ontology sources: Gene Ontology (GO) Biological Processes, KEGG Pathway and Reactome Gene Sets, and the Kyoto Encyclopedia of Genes and Genomes (KEGG). If the adjusted *p*-value < 0.05, the biological process or pathway was considered significantly enriched.

5.2.6 Autoencoder Implementation

Autoencoders are feed-forward neural networks that aim to copy the input variable to the output variable with the minimum loss of information. It compresses the inputs into latent variables in the bottleneck layer's embedding space and then reconstructs the output from the embedding space. The autoencoder is composed of two parts: the encoder and the decoder. The encoder maps the high dimensional input data into latent variables in embedding space, and the decoder reconstructs the input data from the embedding. Here, one concatenated layer, one hidden layer and bottleneck layer were employed in the encoding part. A concatenated autoencoder to integrate the gene expression and methylation data were used. The concatenated autoencoder was implemented using the Keras library with TensorFlow [270] (Figure 5.1). To integrate the gene expression and metylation level of LGG, in the hidden layer of autoencoder, a rectified linear activation function (ReLU) was used. In bottleneck layer, uniform kernel initializer and linear activation function were implemented. Similarly in the decoding layer also one hidden layer and concatenated layer were used. ReLU activation function was applied to the decoder layer. Same architecture employed in the GBM dataset for integrating the gene expression and methylation data. In the GBM dataset Exponential Linear Unit (ELU) activation function was used in the hidden layers. Linear activation function and uniform kernel initializer were employed at bottleneck layer. Further, ELU activation function is applied to the decoder layer. Epoch size and batch size were 1500 and 16, respectively, in each dataset. The network design was implemented following the Fig. 5.1. A total of 1110 features from gene expression, and 3204 features from DNA methylation were selected, in the LGG dataset, while in the GBM

dataset, 268 features from gene expression, 447 features from DNA methylation were selected for the input layer. For the autoencoder, concatenate layer, hidden layers, and a bottleneck layer was set, respectively. The 400 and 100 features were obtained from the bottleneck in LGG and GBM datasets, respectively.

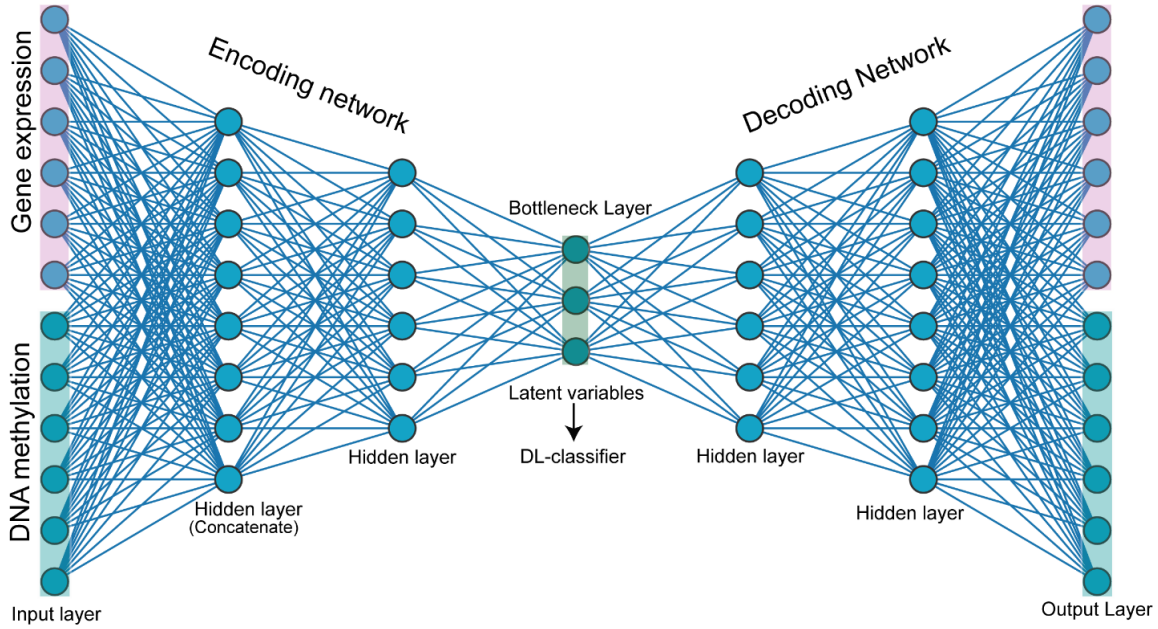

Figure 5.1.: *Architecture of autoencoder: The autoencoder used in DeepAutoGlioma consists of an encoder and a decoder made from 2 hidden layers and one bottleneck layer. The autoencoder has two input layers for DNA methylation and gene expression; in the first hidden layer, data is concatenated, and is passed to another hidden layer and finally compressed in the bottleneck layer. In the decoder part, the latent variables from the bottleneck layer are reconstructed to the initial ones.*

5.2.7 Deep learning classifier

ANNs, which imitate the human brain, are feed-forward neural networks. ANNs are represented by a weighted, directed graph connecting inputs to a series of interconnected "hidden" layers that are composed of multiple nodes called "neurons," that are in turn connected to an output layer [81]. ANNs are trained to recognize and categorize complex patterns. There are one input layer, one output layer and one hidden layer in the network. The hidden layers lies between the input and output layers. The number of output neurons varies depending on the

specific application, while the number of input neurons is equal to the number of attributes. Here latent variable obtained from bottleneck layer of autoencoder were used as input.

CNN is a type of deep learning method that directly learns from the data. CNN consists of three layers: convolutional, pooling and fully connected (FC) layers [82]. The convolutional layer is the first layer, while the FC layer is the last. In the first layer i.e., the convolutional layer, where filters are applied to raw data or feature maps in deep CNN, convolution is one linear operation utilized in place of generic matrix multiplication. The convolution operation (denoted by an asterisk) is defined by:

$$f(t) = (x * K)(t)$$

Where the function $x\,(t)$ is referred to as input, $K(t)$ is referred to as kernel, and the $f(t)$ is referred to as output. After convolutional layer, the genes is downscaled by the Pooling layer to save computation, and the final prediction is made by the fully connected layer. Since every node in a single layer is fully connected to every node in the subsequent layer, it represents a network that is fully connected. This paper uses the Keras library to build these two deep-learning classifiers on the Python platform. Furthermore, parameters were optimized with the grid search method using the GridSearchCV package in Python. After finding the best features, the 70% training dataset was employed using a stratified k-fold. In a stratified k-fold CV, the dataset is split into k different folds, of which k-1 was utilized to train the network, and the final fold was set aside for testing. This procedure is then repeated until all folds are used once as a test set. The final output is then computed by averaging the performance parameters obtained from each test set.

5.2.8 Performance evaluation

The performance of the DL model was evaluated based on the eight criteria: Accuracy, Sensitivity, Specificity, Precision, F1-score, FPR, Geometric mean, and MCC. All the matrices are described in chapter 3 in details.

5.2.9 Statistical analysis

Pairwise comparison was done using Mann-Whitney U test using Sigma Plot 11.0.

5.3 Results

5.3.1 Identification of biologically relevant features for classification of LGG and GBM subtypes

Deregulated gene expression and aberrant methylation are the hallmarks of human cancer [271]. Methylation status in the promoter region determines the level of gene expression. Therefore, linking the methylome and transcriptome is crucial in finding the genetic and epigenetic features that cause cancer, which is also important for making biologically relevant models. To connect the methylome and transcriptome, patients with transcriptome and methylome profiles were chosen to identify the upregulated, downregulated genes (DEGs) and hypomethylated and hypermethylated CpGs (DMRs). A z-score method was used to screen the DEGs and DMRs (see methodology). A z-score greater than 1 or less than -1 indicates the gene expression and methylation are greater or less than the population mean, respectively. The DEGs and DMRs for each subtype of LGG and GBM were identified. In LGG, a total of 3972, 4024, and 4088 DEGs (Figure 5.2A) and 177458, 181957, and 181163 DMRs were found (Figure 5.2B) in astrocytoma, oligoastrocytoma, and oligodendroglioma, respectively. In subtypes of GBM, a total of 3910, 3767, and 3745 DEGs (Figure 5.2C), and 211764, 208111, and 190743 DMRs were found (Figure 5.2D) in classical, mesenchymal, and proneural, respectively. It is also found that differences in average expression and methylation level between z > 1 and z <- 1 are statistically significant (p-value < 0.001) in all subtypes (Figure 5.2A-D). Next, a univariate cox regression analysis was performed to find the correlation between patient prognosis with DEGs and DMRs. The univariate predictive models for each differentially expressed gene (DEG) and differentially methylated region (DMR) were separately generated. Next, the survival-associated genes and CpG sites were screened based on the p-value < 0.05. Our results showed that, in LGG, a total of 2295 DEGs and 18068 DMRs, and in GBM total of 1055 DEGs and 5033 DMRs were linked to the patient's survival. It is found that a total of 50.83 % of DEGs and 20.35% of DMR in LGG; and 23.30% of DEGs and 5.41% of DMR in GBM were linked with patient survival. This indicates that a higher percentage of genes, or CpGs, are not linked with patient's survival. Therefore univariate cox analysis facilitates identifying the biologically important and cancer-associated features, which can lead to the development of a clinically relevant DL model while reducing the dimension of data to build better-fit prediction models. Subsequently, the survival-associated CpGs located in promoters, namely in regions (TSS1500, TSS200, the first exon, and the 5' UTR) were mapped and subsequently linked to their respective survival-associated genes. The linking of

these two layers of genomic data confirms that particular CpG and associated gene pairs are involved in cancer progression. It is found that in LGG, a total of 1110 genes (DEGs) and 3204 CpGs (DMRs) in the promoter, and in GBM, 268 genes (DEGs) and their 447 CpGs (DMRs) in the promoter are linked to the patient survival. If a gene is involved in patient survival, and if its methylation level in the promoter, which regulates its expression, is also linked to survival, indicating an additive impact of methylation and gene expression on patient prognosis. It is believed that integrating methylation level with gene expression data will be more biologically valid for diagnostic model development. It is also found that these genes (prognostic genes) are involved in biological processes and pathways that are linked to cancer (Figure 5.3 A and B) such as signaling by ALK in cancer [272], cell-cell adhesion [273], signaling by receptor tyrosine kinase [274], PID INTEGRIN A4B1 pathway [275], gliogenesis [276], positive regulation of cell adhesion [277] and VEGFA-VEGFR2 signaling pathways [278,279]. Therefore, these prognostic genes and CpGs were used for autoencoder-based data integration and model building.

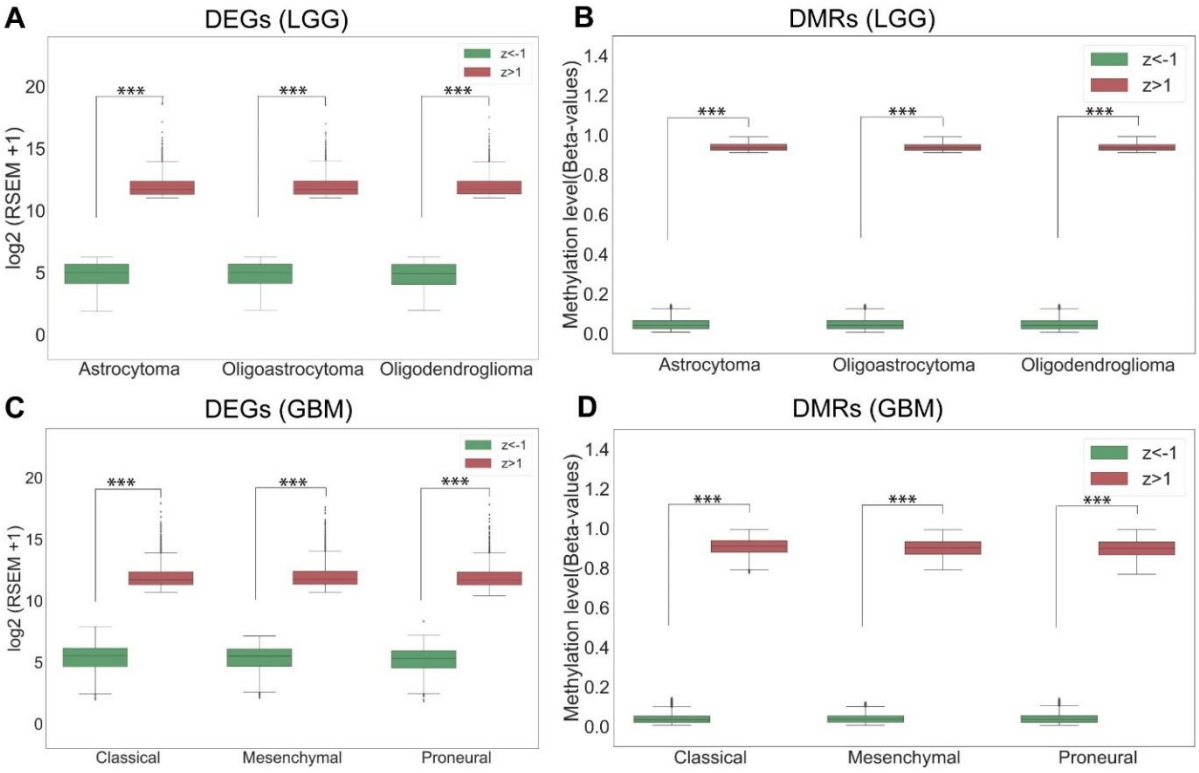

Figure 5.2: *Boxplots show the difference in gene expression and methylation level between $Z > 1$ and $Z < -1$. (A) DEGs and (B) DMRs in each LGG subtype; (C) DEGs and (D) DMRs in each GBM subtype (***p< 0.001). DEGs: differentially expressed genes, DMRs: differentially methylated CpGs.*

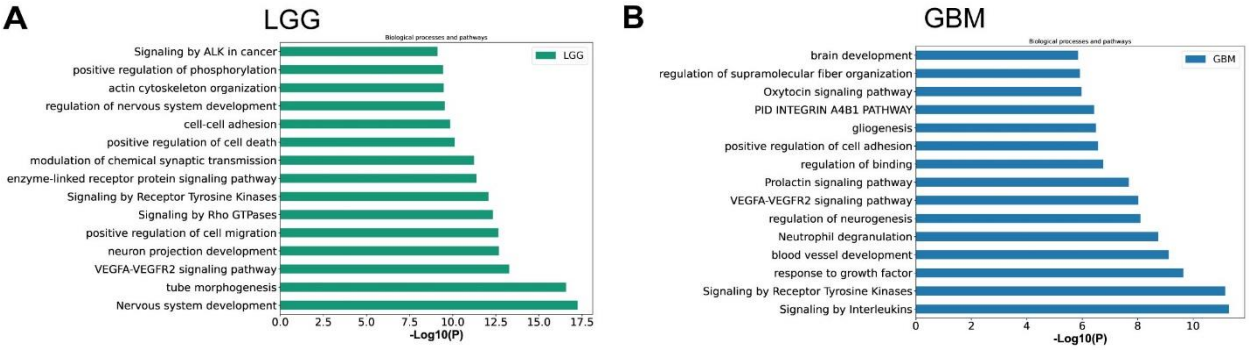

Figure 5.3: *(A) and (B) Bar plots represent significantly enriched Biological processes and pathways of genes used as input in the autoencoder (*p< 0.05).*

5.3.2 Integration of gene expression and its promoter methylation level by autoencoders shows superior accuracy in subtyping.

In the previous section, It is derived the list of genes and their CpG sites in promoters linked to patient survival using univariate cox regression analysis. Next, the gene expression and methylation matrix were extracted. These datasets into training (70%), and validation (30%) sets were divided. 70% of the data was utilized to optimize the model's parameters and evaluate the performance of each model, and the remaining 30% of data was employed as independent predictors. The gene expression and methylation matrices were fed into the autoencoder with concatenated inputs (CNC-AE). The methylation and gene expression levels are combined and compressed in the latent space or bottleneck layer learned by the autoencoder [280–284]. All the dimensions and parameters of the different layers in the autoencoder were optimized. The autoencoder consists of two parts, an encoder, and a decoder network. In the encoder network, gene expression and DNA methylation profiles of LGG and GBM are first encoded into two 4314 and 715-dimensional vectors separately through hidden layers, respectively. Next, the dimensions of the bottleneck layers at 400 and 100 for LGG and GBM were set. In the decoder network, the latent variables were again used to decode the original input data, and this was used to measure the reconstruction loss, which indicates the performance of the autoencoder. The network structure of the decoder is similar to the mirror image of the encoder network (Figure 5.1). If a latent variable captures the actual data pattern, i.e., intrinsic relationships between the variables, then the difference between the encoded and decoded vectors will be less. The reconstruction loss was measured by using Mean Squared Error (MSE). It is found

that MSE was significantly lower, i.e., 0.04 in LGG and 0.04 in GBM. This shows that the autoencoder efficiently learned the pattern in gene expression and methylation and encoded it in the latent space. Then these latent variables were used to develop the DL models for the classification of LGG and GBM subtypes.

Two DL algorithms, i.e., artificial neural network (ANN) and Convolutional neural network (CNN) were implemented, and compared their performance for subtype classification. During the model training step, the grid search method to find the best combination of hyperparameters were used (Table 5.1). Then, using these optimal hyperparameters, stratified k-fold cross-validation (k = 10) on the latent variables was performed and computed the average performance measures for each DL model (Table 5.2). Average accuracy, recall, precision, F1-score, False positive rate (FPR), Geometric mean (GM), and Matthew's Correlation Coefficient (MCC) were used to assess the model's performance (see materials and methods). It is found that CNN models had higher prediction accuracy in subtyping, i.e, 98.03 % and 94.07%, for LGG and GBM, respectively, than the ANN models. The standard deviation (SD) of accuracy from a 10-fold cross-validation was measured. The SD was between ±0.06 and ±0.10, indicating the stability of the CNN model in a wide range of patient samples. It is found that FPR (0.01 and 0.02) was minimal, and the MCC scores were high (0.96 and 0.93) in the case of CNN (Table 5.2). The higher MCC score represents a good correlation between the observed and predicted classes.

Table 5.1: Hyperparameters for ANN and CNN models

Parameters	**Datasets**			
	LGG (ANN)	**LGG (CNN)**	**GBM (ANN)**	**GBM (CNN)**
Activation	relu	relu	linear	elu
Batch_size	32	64	30	64
Dropout_rate	-	0.2	0.1	0.2
epochs	100	2000	50	2000
filters	-	1	-	1
Kernel_size	-	3	-	3
optimizer	adam	RMSprop	RMSprop	RMSprop

Table 5.2: Performance evaluation of LGG and GBM subtypes classification

	Methods	Performance measures (Average of 10 fold cross-validation)						
		Accuracy	**Precision**	**Recall**	**F1-score**	**FPR**	**Gmean**	**MCC**
LGG	ANN	95.40%(±0.09)	92.50	92.73	92.45	0.03	95.28	0.89
	CNN	98.03%(±0.06)	97.67	96.96	96.97	0.01	97.99	0.96
GBM	ANN	92.19%(±0.10)	88.05	89.77	87.75	0.03	94.76	0.90
	CNN	94.07%(±0.10)	90.40	91.18	90.25	0.02	96.51	0.93

Next, the classification using validation datasets were performed to check the reproducibility of the DL framework. It is found the accuracy of subtype classification (for LGG 95.23 % and GBM 90.26%) of CNN was superior, and the MCC score was 0.90 and 0.92 (Table 5.3). The accuracy of the current framework for subtyping LGG and GBM outperforms that of earlier machine learning (ML) and deep learning (DL) models [285,286]. This framework was named as DeepAutoGlioma (Figure 5.4). It is also observed the superior performance of DeepAutoGlioma using external GEO datasets (Table 5.4). The combination of feature genes and CpG sites in the model construction likely accounts for the impressive performance of DeepAutoGlioma. In most cases, feature selection approaches that rely on ML or DL ignore the biological relevance of features [287–289]. However, here the DEGs and DMRs in each subtype were screened, which was associated with LGG and GBM patients' survival. Also, the genes and methylation sites used as inputs into the autoencoder are linked through their genomic locations. Together, these approaches reduce the dimension of data, which significantly influences the model's performance. In our opinion, biologically relevant inputs to the autoencoder provided superior accuracy (95-98%) in the subtype classification achieved with CNN.

Table 5.3: Classification performance of deep learning algorithms on LGG and GBM subtypes for validation set

	Methods	Performance measures (Average of 10 fold cross-validation on test dataset)						
		Accuracy	**Precision**	**Recall**	**F1-score**	**FPR**	**Gmean**	**MCC**
LGG	ANN	90.18%(±0.13)	82.4	84.23	82.42	0.07	90.06	0.80
	CNN	95.23%(±0.09)	92.08	92.63	91.84	0.03	95.3	0.90
GBM	ANN	93.85%(±0.13)	93.85	93.85	93.85	0.00	100	1
	CNN	90.26%(±0.14)	85.38	87.69	86.15	0.02	95.26	0.92

Table 5.4: Classification performance of DeepAutoGlioma on external datasets

	Methods	Performance measures (Average of 10 fold cross-validation)						
		Accuracy	**Precision**	**Recall**	**F1-score**	**FPR**	**Gmean**	**MCC**
LGG	ANN	91.89%(±0.13)	91.20	88.00	86.90	0.06	92.13	0.83
	CNN	91.38%(±0.09)	91.38	91.38	91.38	0.00	100	1
GBM	ANN	84.1%(±0.23)	74.48	79.48	76.15	0.06	90.55	0.72
	CNN	86.41%(±0.24)	79.87	83.33	81.02	0.05	92.92	0.76

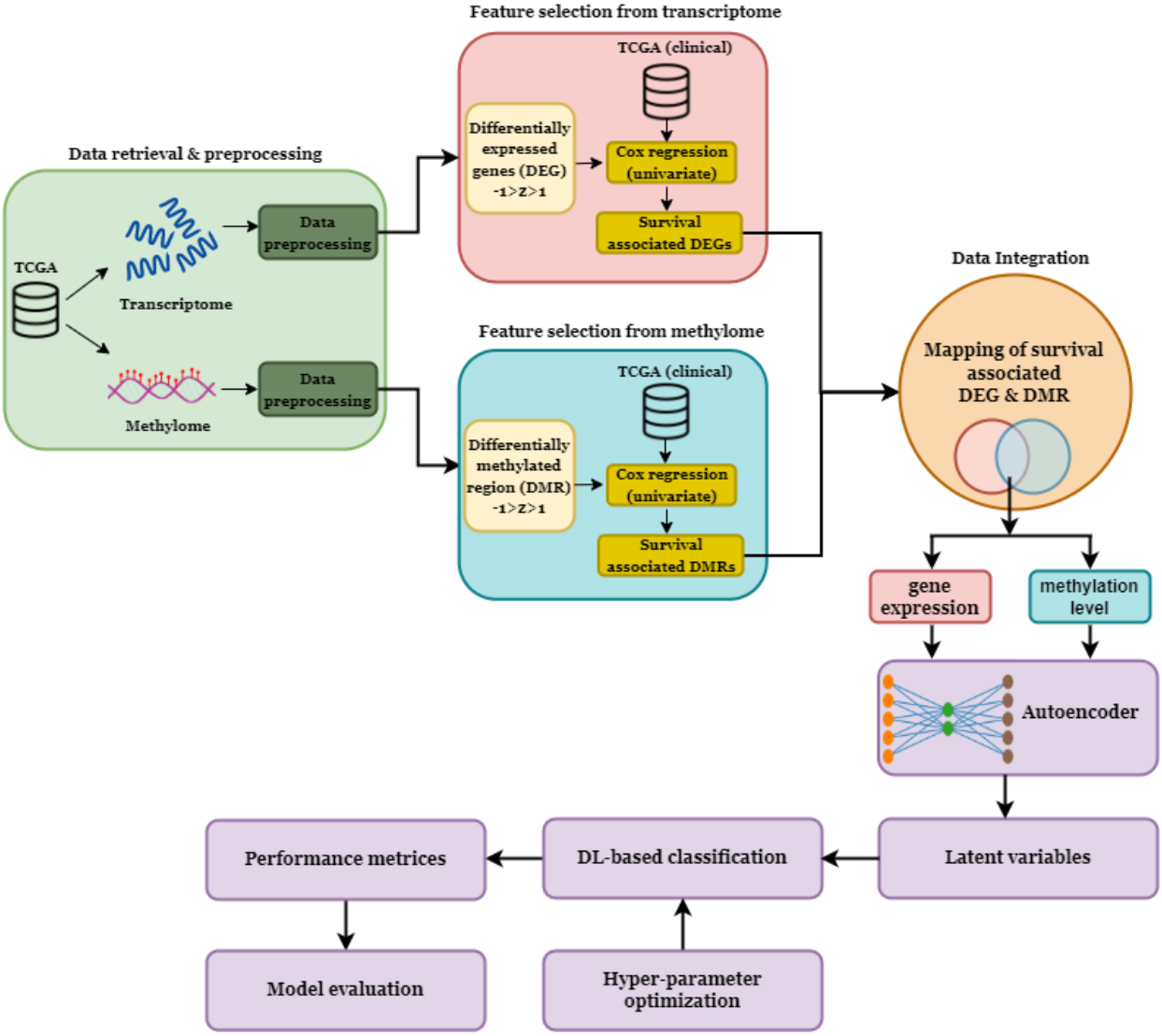

Figure 5.4: *Subtype classification framework of the DeepAutoGlioma. Methylome and transcriptome data are preprocessed, differentially expressed genes (DEGs) and differentially methylated regions (DMRs) are identified, and clinically significant features are extracted. Further, these features are mapped according to the genomic region to integrate the CpG-gene pair. Then, clinically relevant methylation (CpGs) and gene expression data are fed into the autoencoder, and latent variables are extracted to build deep learning models for subtyping brain cancer.*

5.3.3 DL-models with a random feature set, preprocessed data, and single omics data

To validate our findings and better understand the role of feature selection in model performance, the DL-based model by feeding different sets of inputs (features) were extended to the autoencoder and compared their performance to that of DeepAutoGlioma. First, the performance of mapped CpGs and gene expression with randomized CpG-genes pairs as input into the autoencoder were compared. The randomly the CpGs (n = 3204 for LGG and n = 447 for GBM) and genes (n = 1110 for LGG and n = 268 for GBM) from preprocessed data were selected. Then this unmapped randomly selected methylation and gene expression data were fed into the autoencoder. Then, ANN and CNN were used to classify the subtypes using the latent features from random datasets. This process was repeated ten times, and the accuracy varied from 60.68 - 71.43% in LGG, and 62.42 - 72.14% in GBM in all iterations (Table 5.5 and 5.6). And the average accuracy of all iterations in CNN are 66.12 and 66.59% in LGG and GBM, respectively. When compared to DeepAutoGlioma, the average accuracy of all ten iterations in CNN is significantly less (p-value < 0.001, Figure 5.5). Not only the accuracy but other parameters such as precision, MCC and FPR are very less compared to DeepAutoGlioma. This finding confirms that mapping the promoter methylation region to the gene has aided in predicting LGG and GBM subtypes with greater accuracy and precision.

Table 5.5: Model performance in LGG subtype classification using random features

	Methods	Performance measures (Average of 10 fold cross-validation)						
		Accuracy	**Precision**	**Recall**	**F1-score**	**FPR**	**Gmean**	**MCC**
Iteration1	ANN	64.02%(±0.08)	45.77	43.67	43.51	0.26	64.22	0.2
	CNN	64.09%(±0.05)	43.51	43.43	42.5	0.26	64.24	0.2
Iteration2	ANN	60.68%(±0.04)	23.49	33.52	22.72	0.28	58.33	0.08
	CNN	62.44%(±0.08)	31.62	37.27	28.9	0.27	60.23	0.11
Iteration3	ANN	69.06%(±0.08)	53.01	51.34	51.55	0.22	68.77	0.3
	CNN	70.62%(±0.08)	55.78	53.29	52.82	0.22	70.34	0.33
Iteration4	ANN	68.79%(±0.09)	51.89	51.21	51.05	0.23	68.92	0.3
	CNN	67.29%(±0.07)	47.83	48.09	47.03	0.24	67.52	0.27
Iteration5	ANN	65.69%(±0.10)	47.81	47.81	46.48	0.25	65.69	0.23
	CNN	68.24%(±0.07)	50.68	51.05	49.36	0.24	68.33	0.29
Iteration6	ANN	64.65%(±0.07)	45.16	44.47	43.76	0.26	64.35	0.2
	CNN	64.23%(±0.06)	44.92	44.12	43.53	0.26	63.75	0.18
Iteration7	ANN	69.14%(±0.04)	53.62	51.78	50.93	0.23	69.08	0.31
	CNN	65.96%(±0.07)	47.14	45.73	44.67	0.25	65.74	0.23
Iteration8	ANN	68.21%(±0.05)	51.45	49.82	48.8	0.23	68.22	0.28
	CNN	65.72%(±0.06)	47.76	47.07	45.88	0.25	65.75	0.24
Iteration9	ANN	64.9%(±0.11)	45.71	44.66	43.87	0.26	64.9	0.21
	CNN	65.74%(±0.08)	46.01	46.24	44.58	0.25	65.82	0.24
Iteration10	ANN	71.43%(±0.06)	55.35	54.88	53.7	0.21	71.56	0.36
	CNN	66.88%(±0.05)	50.12	48.96	48.47	0.24	66.88	0.26

Table 5.6: Model performance in GBM subtype classification using random features

	Methods	Performance measures (Average of 10 fold cross-validation)						
		Accuracy	**Precision**	**Recall**	**F1-score**	**FPR**	**Gmean**	**MCC**
Iteration1	ANN	64.59%(±0.11)	31.46	43.83	35.25	0.24	64.4	0.25
	CNN	62.42%(±0.16)	39.26	43.68	38.79	0.27	60.67	0.17
Iteration2	ANN	69.55%(±0.16)	46.55	57.47	47.91	0.22	68.81	0.39
	CNN	71.49%(±0.12)	50.88	56.94	49.59	0.21	70.36	0.38
Iteration3	ANN	69.68%(±0.18)	53.81	60.21	52.84	0.22	68.71	0.39
	CNN	70.31%(±0.16)	61.3	62.06	57.24	0.21	68.56	0.38
Iteration4	ANN	66.59%(±0.18)	46.9	48.83	45.35	0.26	61.95	0.27
	CNN	62.64%(±0.18)	35.44	43.66	36.27	0.28	61.94	0.18
Iteration5	ANN	65.64%(±0.12)	39.77	46.62	40.61	0.25	64.85	0.25
	CNN	69.71%(±0.13)	51.31	57.76	50.84	0.22	69.2	0.38
Iteration6	ANN	65.09%(±0.12)	38.56	48.02	39.36	0.26	64.46	0.25
	CNN	65.01%(±0.14)	33.05	48.04	37.55	0.26	63.7	0.24
Iteration7	ANN	68.64%(±0.17)	43.2	51.9	44.94	0.24	67.29	0.29
	CNN	68.47%(±0.07)	46.11	56.26	47.58	0.22	67.63	0.37
Iteration8	ANN	68.13%(±0.11)	37.23	45.75	39.30	0.22	65.76	0.26
	CNN	63.19%(±0.15)	32.73	41.66	34.25	0.26	62.49	0.23
Iteration9	ANN	72.14%(±0.17)	51.47	53.64	50.68	0.2	71.29	0.4
	CNN	67.36%(±0.08)	42.73	52.43	44.16	0.24	66.53	0.32
Iteration10	ANN	65.19%(±0.10)	33.65	48.04	37.07	0.24	64.09	0.26
	CNN	65.36%(±0.15)	39.78	50.04	41.71	0.26	63.02	0.26

To better understand the significance of biologically relevant features, such as DEGs and DMRs, as well as univariate Cox regression analysis for feature selection, the autoencoder is executed on preprocessed data and then classify using ANN and CNN. LGG and GBM gene expression and methylation data matrices contain 14517 and 14125 genes, respectively, as well as 139403 and 141672 CpGs. The autoencoder was then run on these preprocessed datasets, and the accuracy of prediction, as well as other model evaluation parameters, were measured (Table 5.7). When compared to DeepAutoGlioma, the prediction accuracy is significantly (*p* - value < 0.001) lower (Figure 5.5). The subtypes classification accuracy of LGG was 83.73% (±0.11) in CNN and 69.86% (±0.07) in ANN. Whereas in GBM classification, accuracy was 61.54% (±0.19) in CNN and 67.58 %(±0.15) in ANN. Furthermore, the results of other evaluation parameters were too low to be considered. This unequivocally demonstrates that cancer-associated features or features that are biologically relevant played a crucial role in achieving higher classification accuracy.

Table 5.7: Model performance in LGG and GBM subtyping using preprocessed data as a feature

	Methods	Performance measures (Average of 10 fold cross-validation)						
		Accuracy	**Precision**	**Recall**	**F1-score**	**FPR**	**Gmean**	**MCC**
LGG	ANN	69.86%(±0.07)	53.46	53.4	50.47	0.22	70.22	0.35
	CNN	83.73%(±0.11)	75.77	73.99	73.4	0.11	83.5	0.64
GBM	ANN	67.58%(±0.15)	39.24	48.83	41.05	0.24	65.9	0.27
	CNN	61.54%(±0.19)	34.77	44.05	36.26	0.29	61.35	0.18

Furthermore, the classification accuracy between di-omics and mono-omics data were compared. The mono-omics data, i.e., methylation or gene expression matrix, was used as input to the autoencoder. As previously stated, compressed features from latent space were extracted, used DL algorithms, and calculated average performance metrics for each DL model. It is observed that in the case of LGG, the single omics data showed good accuracy of prediction, i.e., 96.27% (±0.11) and 96.55 % (±0.10) using gene expression and methylation data, respectively (Table 5.8). But these accuracies are lower in comparison to the DeepAutoGlioma (98.03% ±0.06). However, the accuracy of prediction using test and external gene expression (66.07±0.08% and 63.81±0.13%) and methylation (66.51±0.09% and 74.79±0.12%) datasets is considerably less.

Whereas in the case of GBM subtype prediction accuracy using gene expression and methylation data were 91.54% (±0.11) and 43.89% (±0.07), respectively (Table 5.8). Although gene expression data showed an accuracy, however in the test and external datasets, the accuracy was 85.48% (±0.23) and 72.68% (±0.15). The good prediction accuracy in LGG and GBM was observed utilizing mono-omics data, particularly gene expression, models were unable to accurately predict subtypes using test and external datasets. This demonstrated that the individual omics data were inadequate for cancer subtype classification with superior accuracy. The models trained on multi-omics data outperformed those trained on single-omics data, owing to the fact that multi-omics data contains a wealth of information not found in a single type of omics data alone.

Table 5.8: Classification performance of deep learning algorithms on LGG and GBM subtyping using mono-omics data

		Methods	Performance measures (Average of 10 fold cross-validation)						
			Accuracy	Precision	Recall	F1-score	FPR	Gmean	MCC
LGG	only Gene expression	ANN	94.70%(±0.11)	91.22	91.41	91.29	0.03	94.64	0.87
		CNN	96.27%(±0.11)	94.25	94.3	94.15	0.02	96.3	0.91
	only DNA Methylation	ANN	92.61%(±0.13)	88.22	87.73	87.73	0.05	92.52	0.83
		CNN	96.55%(±0.10)	93.81	94.32	94.03	0.02	96.6	0.92
GBM	only Gene expression	ANN	85.92%(±0.10)	73.87	78.83	74.24	0.08	88.91	0.8
		CNN	91.54%(±0.11)	85.25	88.45	85.91	0.04	94.14	0.89
	only DNA Methylation	ANN	44.60%(±0.06)	42.67	11.88	18.52	0.2	43.13	0
		CNN	43.89%(±0.07)	42.67	13.05	19.77	0.19	44.89	0

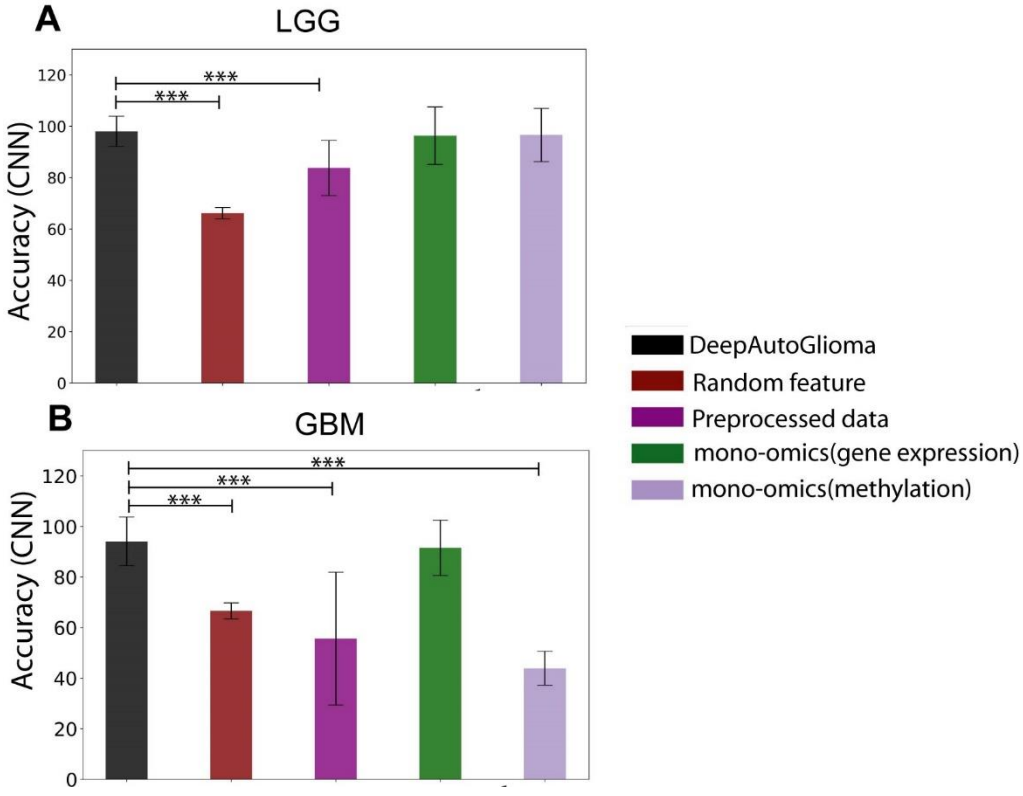

Figure 5.5: *Comparison of model performance using different sets of features to that of DeepAutoGlioma (***p< 0.001).*

5.4 Discussion

It is well established that molecular perturbations in different genomic layers cause cancer occurrence and progression. Therefore, it is crucial to perform integrative approaches that combine multi-omics data to comprehend the disease mechanism and develop novel diagnostic tools for brain cancer detection. The integration of high-throughput omics data from distinct genome layers can capture the interrelationships of biomolecules and facilitate interpreting their function in disease onset. Transcriptomics and epigenomics data are unpaired because they are usually measured in separate experiments, which demands effective and efficient in-silico multi-omics integration [283]. In the present study, **the deep autoencoder and deep learning (ANN & CNN) -based clinically relevant framework was designed for integrating the methylome and transcriptome to classify the glioma subtype with superior accuracy.** To strengthen the biological relevance, patient samples with transcriptome and methylome profiles were screened and measured the DEGs and DMRs in each subtype of LGG and GBM cancer. Further, a univariate cox regression analysis was performed to identify the DEGs and DMRs associated with the patient's survival. Univariate cox regression approach helps to determine clinically relevant feature genes and CpG sites based on the patient's overall survival information; further, it also decreases the data dimension. Next, we map the CpGs and genes based on the promoter regions. The linked CpGs and genes were used as input in the autoencoder. As a result, the input features in the autoencoder were biologically and clinically relevant in three ways, first, they are differentially regulated; second, they are linked to the patient's survival; and third, methylation in the promoter is linked to gene expression. **It is found that using latent variables learned by autoencoder as an input in deep learning models (ANN & CNN), we were able to predict the subtype of LGG and GBM with the accuracy of 98.03%(±0.06) and 94.07%(±0.10), respectively, using CNN.** Furthermore, the current framework classifies the GBM and LGG subtypes using the external datasets with 86.41% and 91.89% accuracy, respectively. On the other hand, **autoencoder-based deep learning with a single type of omics data, randomized CpG-gene pair, and preprocessed dataset did not perform well compared to DeepAutoGlioma. We believe that feature screening using various statistical methods and integration of di-omics data using autoencoders played an essential role in achieving higher subtyping accuracy.** The current study demonstrated how data integration could lead to the discovery of novel patterns in transcriptomics and epigenomics data and aid in developing efficient diagnostic tools.

CHAPTER 6

OBJECTIVE 4

Chapter 6: Objective 4
Identification of subtype-specific disease modules and development of drug response prediction models by combining network medicine and AI-based approaches

6.1 Introduction

Due to distinct molecular characteristics, the subtypes of glioma have different clinical outcomes and responses to treatment, highlighting the importance of personalized medicine for brain cancer treatment [20]. Hence, to address this issue, a framework by combining network medicine and AI-based approaches to systematically integrate omics data to identify subtype-specific disease modules for precision therapy and drug response prediction was developed. Cancer is developed through an evolutionary process in which healthy cells accumulate several genomic changes, including mutations and gene expression [290,291]. Some of these alterations provide a positive selection to cancerous cells, giving them an advantage in uncontrolled proliferation, which lead to the formation of tumors. Advances in sequencing techniques and genome-wide association studies have revealed that accumulated genetic variations associated with an increased risk for cancer are distributed throughout the genome. Further studies illustrate that disease genes are not distributed randomly in molecular networks. However, these genes work together in a biological pathway. Furthermore, genes associated with the same phenotype exhibit a tendency to interact with one another and form clusters within the same network neighborhood. As a result, a disease module forms, a subnetwork linked to a disease. Numerous genes that are known to be relevant to disease are found in disease modules. The disease modules, consisting of a known group of genes in cancers such as kidney, breast, sarcoma, colorectal, leukemia, and head and neck cancers, were found to be associated with cancer-specific biological processes [292]. Wu et al., showed that the active disease modules in breast and cervical cancer are associated with many cancer-related pathways [293]. These studies indicate that the identification of cancer-specific disease modules can help to identify novel biomarkers for therapeutic targets. Therefore, network medicine and rational drug-designing approaches recognize these modules as pharmacological targets as opposed to the individual genes or proteins in the network. Network medicine is the utilisation of network science to identify, prevent, and treat diseases. It provides a platform to comprehensively investigate the

molecular complexity of a specific disease, enabling the identification of disease modules and pathways. Network medicine have provided valuable insights into the connection between drugs targets and disease genes in disease modules. However, the therapeutic efficiency of drugs in cancer is highly context-dependent; often, drug resistance reduces the effectiveness of chemotherapy. Molecular heterogeneity is a major contributor to cancer drug resistance, as it can create subpopulations of cancer cells that may have different mutations or molecular characteristics that allow them to survive even in the presence of the drug. Therefore, the prediction of drug response, i.e., resistance or sensitivity, is essential for improving the efficacy of chemotherapy.

Here, algorithms relying on network medicine and artificial intelligence was deployed to design the framework for subtype-specific target identification and drug response prediction in glioma. The driver mutations that were differentially expressed in each subtype of lower-grade glioma and glioblastoma multiforme was identified that were linked to cancer-specific processes. Driver mutations that were differentially expressed were also subjected to subtype-specific disease module identification. The drugs from the drug bank database were retrieved to target these disease modules. However, the efficacy of anticancer drugs depends on the molecular profile of the cancer and varies among cancer patients due to intratumor heterogeneity. Hence, a deep-learning-based drug response prediction framework was developed using the experimental drug screening data. Models for 30 drugs that can target the disease module, were developed, where drug response measured by IC50 was considered a response; and gene expression and mutation data were considered predictor variables. The model construction consists of three steps, feature selection, data integration, and classification. The consistent performance of the models in training, test, and validation datasets was observed. We predicted drug responses for specific cell lines obtained from different subtypes of glioma. It is found that subtypes of gliomas respond differently to the drug, highlighting the importance of subtype-specific drug response prediction. Therefore, the development of personalized therapy by integrating network medicine and a DL-based approach can lead to the cancer-specific treatment and improved patient care.

6.2 Methodology

6.2.1 Driver gene identification

Brain cancer somatic mutation data was downloaded from the COSMIC database (https://cancer.sanger.ac.uk/cosmic/download) for each subtype [294]. Based on the clinical information, the patient's mutational data were stratified into different subtypes. There are a total of 281 and 123 samples of LGG and GBM, respectively. The LGG into astrocytoma (n = 96), oligoastrocytoma (n = 75), and oligodendroglioma (n = 110); and the GBM into classical (n = 39), mesenchymal (n = 48), and proneural (n = 36) was divided. OncodriveCLUSTL was used to find the driver mutation in subtypes [295]. OncodriveCLUSTL is an unsupervised clustering algorithm that can detect clusters of somatic mutations across a cohort of tumor samples. This algorithm OncodriveCLUSTL is a clustering method that utilizes nucleotide sequence data to identify cancer driver events inside genomic regions. Not all mutations are causative factors of cancer; rather, only certain mutations have the potential to aggregate and contribute to the development and progression of cancer (Figure 6.1). Based on the mutation frequency in each gene and statistical significance (number of mutations >2 and *p*-value <0.05), driver genes were selected in each subtype of glioma.

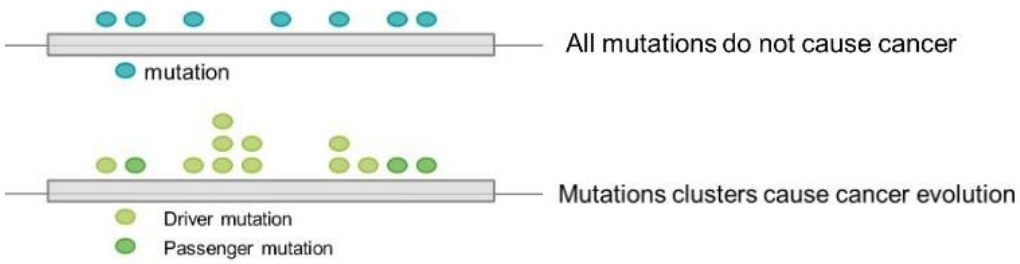

Figure 6.1: *Illustration of finding driver gene in mutation clusters by OncodriveCLUSTL.*

6.2.2 Identification of differentially expressed genes (DEGs)

For computing the DEGs, RNA sequencing data of LGG (n = 281) and GBM (n = 123) patients were obtained from UCSC Xena (https://xena.ucsc.edu/) [165]. Additionally, GTEx healthy brain gene expression data (n = 93) were obtained from the same database. Similarly, like in the previous step, patients were segregated into astrocytoma (n = 96), oligoastrocytoma (n = 75), oligodendroglioma (n = 110), classical (n = 39), mesenchymal (n = 48), and proneural

(n = 36). Next, the data was preprocessed, and low-expressed genes were removed. The cut-off of log2 (RSEM +1) < 0.1 (RSEM: RNA-Seq by Expectation Maximization) in 90% of the samples was used because they did not have any promising information. Finally, in LGG, there are 12,532 genes, and in GBM, 12,183 genes are expressed in cancer and healthy tissue. Next, the differentially expressed genes in each subtype of LGG and GBM was identified using the "limma" package in R. A Q-value (adjusted *p*-value) < 0.05 and a logFC ≥ 1 were used as the statistical threshold for screening DEGs.

6.2.3 Construction of subtype-specific disease module and network analysis

Human brain interactome data was retrieved from TissueNet v.2 database [296]. Brain interactome data contains 165,240 interactions. In TissueNet v.2, the RNA-sequencing raw counts were collected from the Genotype-Tissue Expression (GTEx) project, whereas the protein expression data were obtained from the Human Protein Atlas (HPA). The computation of tissue interactomes was performed for each RNA-sequencing data source, employing a threshold of 8 normalized counts. We performed this computation in order to eliminate protein-coding genes that were not consistently expressed in a brain tissue. Additionally, for the HPA protein, a threshold of low expression was utilized. TissueNet offers comprehensive insights into 16 major human tissues by integrating gene and protein expression profiles into a uniform dataset. It provides a comprehensive network of protein-protein interactions (PPI) partners specific to each tissue. TissueNet v.2 uses human PPIs and tissue-specific expression patterns to make PPIs that are specific to each tissue. The Disease Module Detection (DIAMOnD) algorithm was implemented to identify the disease modules in subtype [292]. DIAMOnD algorithm was used to identify the surrounding genes around a collection of known disease genes, helping to identify new biomarkers. In DIAMOnD algorithm, first connectivity significance was determined for all genes connected to the disease genes. Subsequently, the genes were ranked based on p-values. Those genes having the highest rank or lowest p-value was added to the set of seed node until the whole genes were added into the disease module network. TissueNet v.2 brain interactome and subtype-specific DEDGs are used as seed genes to identify the disease module. All the parameters in the DIAMOnD were kept as default. Cytoscape and the *igraph* package in R were used for network visualization and analysis.

6.2.4 The pipeline of DNN-based drug response prediction

Experimental data for cancer cell drug sensitivity were obtained from the Genomics of Drug Sensitivity in Cancer (GDSC) [297] project to develop drug response prediction models. This data set contains 1001 cancer cell lines and 288 drugs. The drug and target information were derived from the DrugBank database [298]. All the drugs for which the target genes are present in the disease modules were extracted. The 30 FDA-approved and investigational drugs were screened that can target the disease module and also have brain cancer-specific experimental data in GDSC. The drug response models were developed for these 30 drugs. Therefore, the IC50 values of these 30 drugs were downloaded for all cell lines along with the gene expression and mutational data. A total of 886 cell line data were used for model development. To develop the model, the following steps were performed: 1. Data preprocessing; 2. Feature selection; 3. Data integration; 4. Model development and evaluation; and 5. Model validation on external data.

Data preprocessing: The gene expression data was normalized using the log2 (TPM+1). The low expressed genes were removed using a cutoff (log2 (TPM +1) < 0.1 in 90% of samples. Genes possessing any mutation were assigned a value of 1; genes lacking mutations were assigned a value of 0.

Feature selection: A two-step feature selection method was employed to get more variable features from gene expression data. First, the genes were pre-selected based on a Pearson correlation coefficient r<0.5, and then LASSO was used [223] (A detailed description was provided in the chapter 4) to fine-select the predictor genes. For mutational data, the LASSO feature selection method was only used.

Data integration: After the feature selection step, gene expression and mutation data was integrated using a concatenated autoencoder (Figure 6.2) (A detailed description was provided in the chapter chapter 5). The Keras library with TensorFlow [270] was used to implement the concatenated autoencoder. To integrate the gene expression and mutation data, in the hidden layer of the autoencoder, a rectified linear activation function (ReLU) was used. In the bottleneck layer, uniform kernel initializer and linear activation function were implemented. ReLU activation function was applied to the decoder layer.

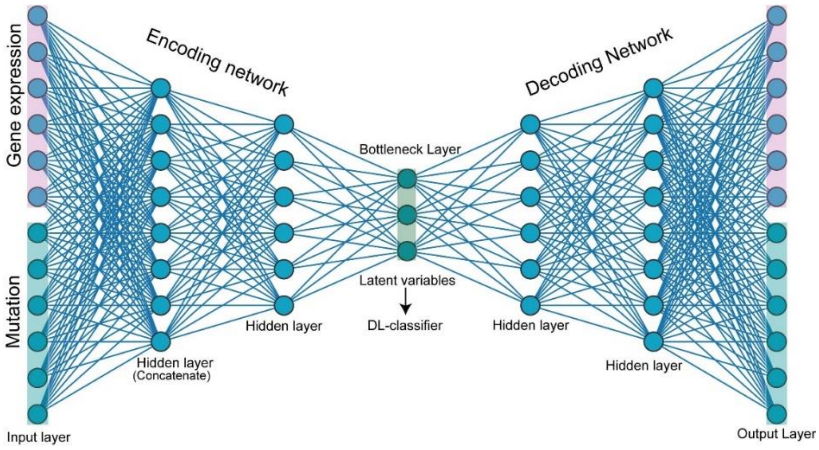

Figure 6.2: *Architecture of autoencoder used for integrating the gene expression and mutation profiles. It consists of an encoder and a decoder made from 2 hidden layers and one bottleneck layer. In the first hidden layer, data is concatenated, and is passed to another hidden layer and finally compressed in the bottleneck layer. In the decoder part, the latent variables from the bottleneck layer are reconstructed to the initial ones.*

Model development and evaluation:

A deep neural network (DNN) classification model was applied to predict the sensitive vs resistance. The input and output layers of a DNN are separated by a number of hidden layers. Complex non-linear relationships can be modelled using it. Deep neural networks handle data in intricate ways by using advanced mathematical modelling. DNNs are frequently used for their accuracy and adaptive nature in the research field of automatic classification tasks. For each drug, the model hyperparameters were optimized by the grid search method using the GridSearchCV package in Python. The DNN architecture consists of two hidden layers for all drugs: the ReLU activation function, adam optimizer, batch size 32, and epochs 2000. IC50 values were binarized to be sensitive and resistant. The model was trained on the 70% training dataset, and stratified k-fold (A detailed description was provided in the chapter 3) was used to compute the performance of the model.

The performance of the DL-model was evaluated based on eight criteria: accuracy, sensitivity, specificity, precision, F1-score, false positive rate, geometric mean, and Matthew's correlation coefficient. A true positive (TP) would indicate that the drug-sensitive cell was correctly identified, while a false positive (FP) indicates that a drug sensitive cell is identified as resistant. Conversely, true negatives (TN) and false negatives (FN) are also calculated (A detailed description was provided in the chapter 3).

Model validation on external data:

The model's performance was validated using the dataset from Cancer Cell Line Encyclopedia (CCLE) (https://portals.broadinstitute.org/ccle) [299]. The cell line gene expression profiles in CCLE and GDSC were generated using different platforms, and thus the data sets have significantly different magnitudes (Figure 6.3). To make these two datasets uniformly distributed, the batch effect was removed using the pyComBat package in Python [300,301]. Then the standardized gene expression profile (brain cancer cell lines) of CCLE was fed to the model built with GDSC datasets to validate the drug response.

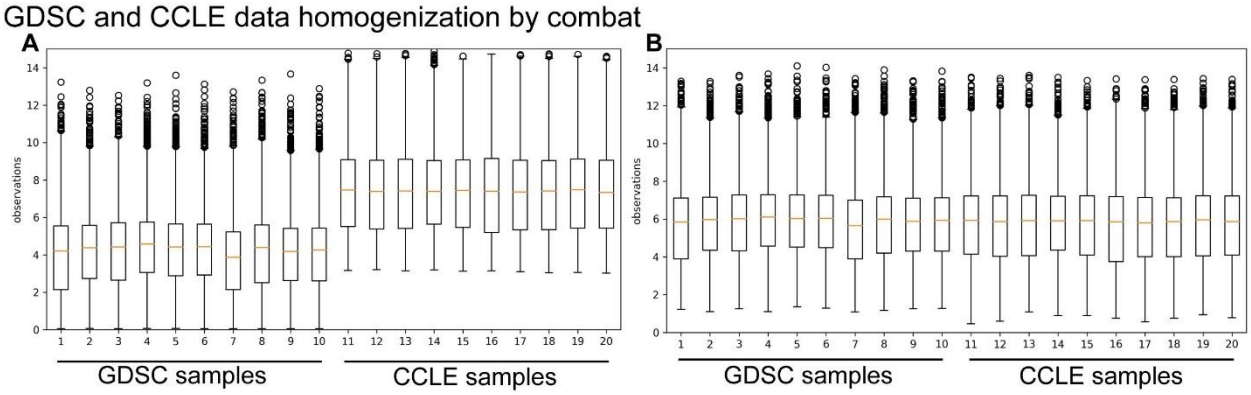

Figure 6.3: *Removal of Batch effect by ComBat. Boxplot showing gene expression distributions before* ***(A)*** *and after* ***(B)*** *ComBat for ten cell lines from GDSC and CCLE.*

6.2.5 Performance evaluation

The performance of the DNN model was evaluated based on the eight criteria: Accuracy, Sensitivity, Specificity, Precision, F1-score, FPR, Geometric mean, and MCC. All the matrices are described in chapter 3 in details.

6.3 Results

6.3.1 Genome-wide screening to identify the driver genes

Cancer mutation can be synonymous and non-synonymous. Synonymous mutations do not affect the amino acid sequence of proteins, whereas non-synonymous mutations cause a different amino acid to be included in the protein and have more immediate consequences for protein function. It is anticipated that nonsynonymous mutations will come under strong positive selection in order to drive oncogenesis. Owing to this fact, the non-synonymous driver mutations in each subtype of LGG and GBM was identified by implementing the OncodriveCLUSTL algorithm [295] and using somatic mutation data from the COSMIC (Table 6.1). It is observed that several driver mutations were associated with each subtype of glioma. Higher-grade tumors frequently have more aggressive features because they typically have more genetic mutations than lower-grade tumors. It is also found that all subtypes of GBM have a higher number of driver mutations than LGG subtypes (Table 6.1). This demonstrates why GBM is more aggressive than other varieties of brain cancer. It is also noticed that these mutations are scattered across the genome rather than being concentrated in a particular location (Figure 6.4 A - F). It is frequently observed that changes in coding sequence cause changes in the expression of driving genes. For instance, a mutation in an oncogene can result in it being overexpressed, promoting the development of cancer. Similar to this, a tumor suppressor gene's expression can be depleted as a result of a mutation, which reduces its growth inhibitory effect. Hence, the differentially expressed genes (DEGs) was identified in each subtype of cancer. The genes with log2Fold Change (FC) > 1 and < -1 and adjusted *p*-value <0.05 were considered DEGs (Figure 6.5 A-F and Table 6.1). The driver genes that are differentially expressed are named as differentially expressed driver genes (DEDGs) (Table 6.1). It is noticeable that a high percentage of the driver genes are differentially expressed, indicating that these genes, i.e., DEDGs, play a critical role in tumorigenesis (Table 6.1). The combined effect of mutations in cancer driver genes and changes in gene expression can enhance the oncogenic effects [302]. These genes may be involved in key pathways and processes involved in cancer development and progression. Therefore, DEDGs can be used to develop targeted therapies that can be used to selectively disrupt subtype-specific processes to regulate cancer.

Table 6.1: Differentially expressed driver genes (DEDGs) in subtypes of glioma

Grade	Histological Type	Number of Driver genes (p-value<0.05)	Number of differentially expressed genes	Number of differentially expressed driver genes (DEDGs)	Percentage of driver gene differentially expressed
LGG	Astrocytoma	1043	6825	460	44.1
	oligoastrocytoma	719	6738	321	44.64
	oligodendroglioma	994	6562	427	42.95
GBM	Classical	1114	6920	424	38.06
	Mesenchymal	1338	7154	553	41.33
	Proneural	1117	6771	426	38.13

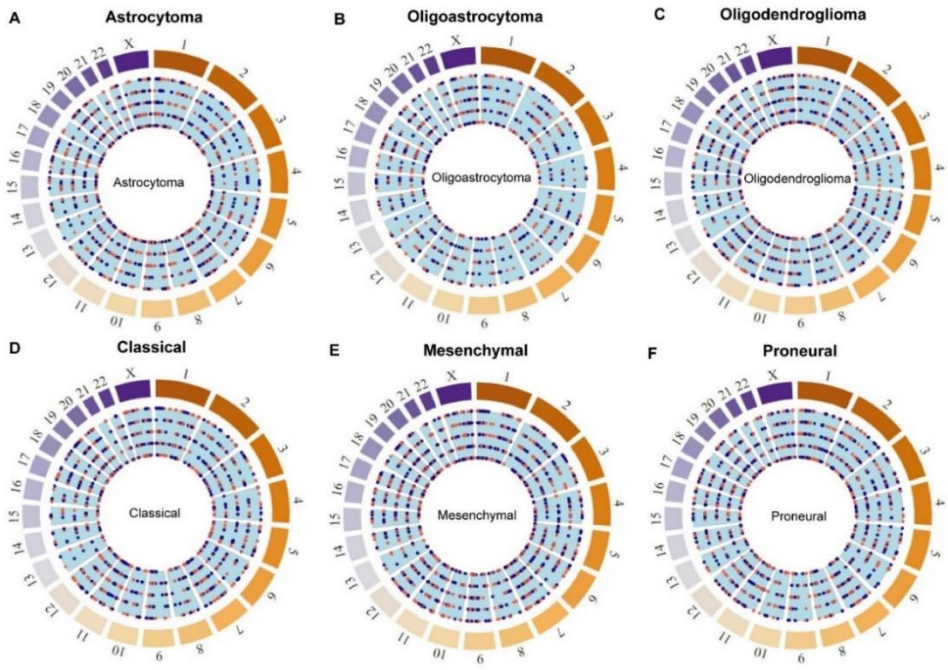

Figure 6.4: *Genome-wide distribution of driver genes in glioma subtypes. Circus plots show the driver genes in different subtypes of LGG and GBM (A-F). Blue and orange dots represent the chromosomal location of driver mutations in the circus plot and mutations are distributed throughout the genome in each subtypes.*

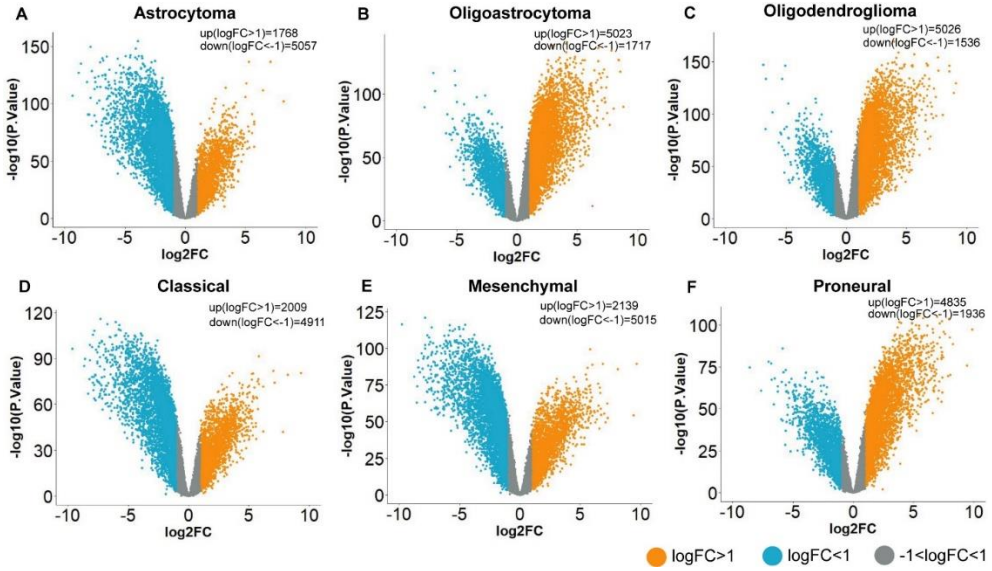

Figure 6.5: (*A-F), the volcano plots represent the differentially expressed genes (DEGs) in different subtypes of glioma. LogFC>1 (p-value< 0.05) is the upregulated gene (orange) and LogFC<-1 (p-value< 0.05) is the downregulated gene (Blue).*

6.3.2 Subtype-specific networks of driver genes (DEDGs) and identification of disease modules

In the previous section, it is observed that in all subtypes, driver genes are distributed across the genome. Many driver genes are also differentially expressed. This two-level perturbation in genes indicates their crucial role in cancer development because biological pathways and processes that involve these genes will likely be deregulated. The gene set enrichment analysis was conducted on DEDGs from each subtype to investigate the affected biological pathways and processes. It is found that cancer-associated processes and pathways were enriched in different subtypes of gliomas (Figure 6.6 A-F). Interestingly, it is found that processes and pathways are mostly distinct among the subtypes, such as in the astrocytoma NOTCH signaling pathway, ATM signaling pathway, and regulation of RNA splicing; in the oligoastrocytoma neovascularization process, EGFR signaling pathways, and FoxO signaling pathway; and in the oligodendroglioma PID ERBB1 internalization pathway and endocrine resistance, which were significantly ($p < 0.05$) enriched. In classical signal transduction by growth factor receptors and second messengers, negative regulation of cellular component organization and regulation of cellular response to stress are prevalent; in mesenchymal focal adhesion, proteoglycan in cancer and cytokine signaling in immune system, and in proneural glioblastoma signaling pathways

and MAPK signaling pathways are prevalent. These results demonstrate how the subtypes differ from one another in terms of their molecular function and biological processes.

Biological processes and pathways

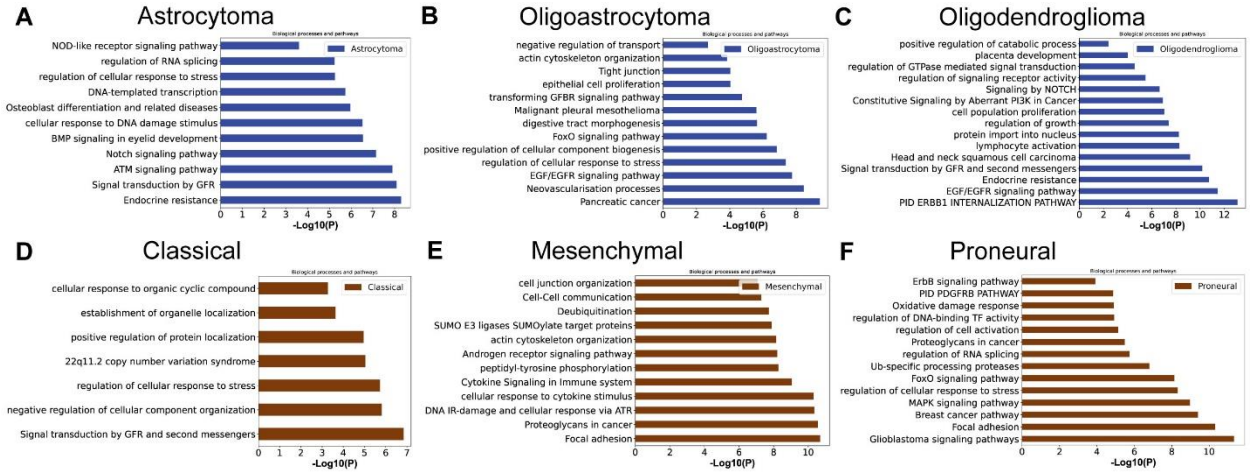

Figure 6.6: *The bar diagrams represent the biological process and pathway enrichment analysis of differentially expressed driver genes (DEDGs) in glioma subtypes (A-F). The highly significant (p-value<0.05) processes and pathways are shown in the figures.*

However, to be involved in biological processes and to drive cancer, these genes must interact. Network-based approaches to human disease demonstrate that abnormalities in a single effector gene product are infrequent causes of disease. Indeed, there is a higher likelihood that genes linked to the same disease will interact with one another [303]. Using the brain interactome data from the TissueNet v.2 database [296], the subtype-specific protein-protein interaction network of DEDGs was built, named the differentially expressed driver gene network (DEDGN), to analyze the interaction pattern. It is observed that a moderate portion of the DEDGs directly interact with each other. The size of the largest connected component (LCC) was calculated in each subtype. LCC refers to the largest subset of nodes in the network that are connected to each other, and often LCCs are involved in crucial signaling pathways that are essential to cellular function. Additionally, it can aid in the identification of prospective drug targets for therapeutic intervention. Figure 6.7 (A-F) shows the LCC in each subtype. It is observed that a lower percentage of DEDGs, i.e., 36.30% in astrocytoma, 34.26% in

oligoastrocytoma, 32.31% in oligodendroglioma, 23.11% in classical, 42.67% in mesenchymal, and 39.67% in proneural, form the LCC. The size of the LCC in reality may be larger than what we have depicted here because the human interactome is incomplete. These LCCs in each subtype, however, provided us with evidence that the development of a precision therapeutic strategy can be aided by the identification of subtype-specific disease modules.

Largest connected component (LCC) in DEDGN

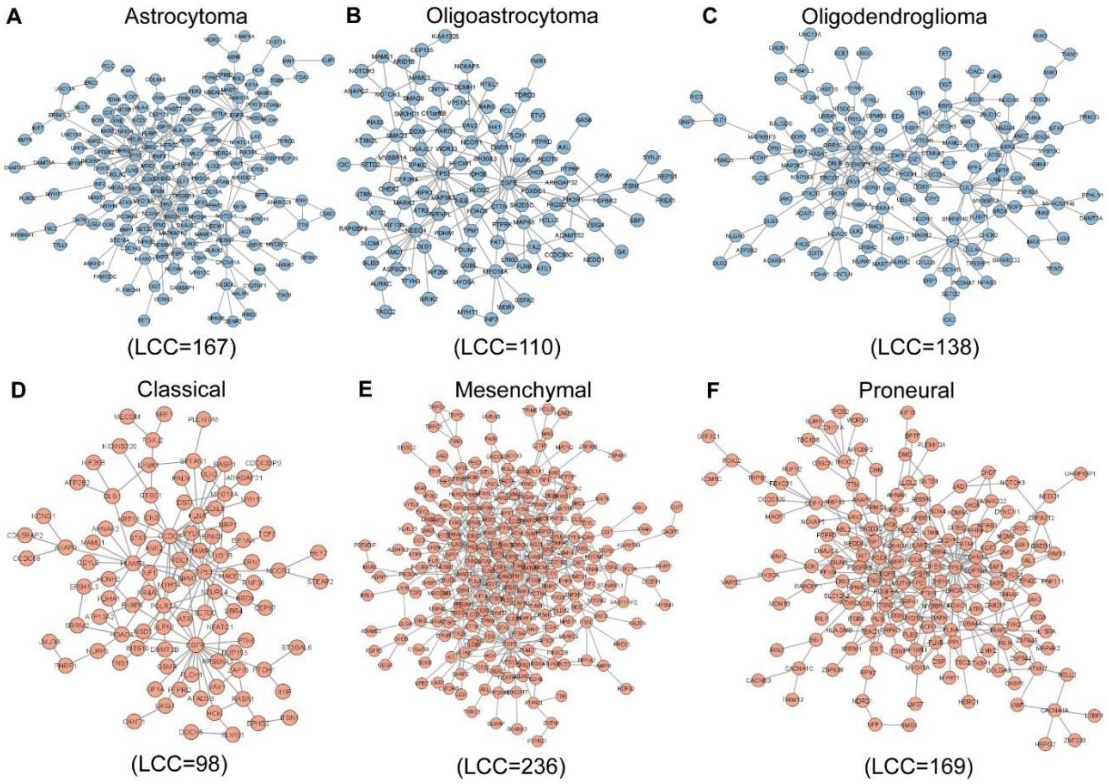

Figure 6.7: (*A-F), show the largest connected component (LCC) of the DEDGs networks in each subtype.*

Therefore, the process of disease module identification was initiated by applying the DIAMOnD algorithm. (Figure 6.8) DIAMOnD enables us to systematically examine the local network neighborhood surrounding a particular collection of known disease proteins in order to find new disease proteins. All DEDGs in each subtype to be known disease genes was considered and used them as seed genes in DIAMOnD. The number of DIAMOnD disease

module genes in each subtype of LGG and GBM was 607 in astrocytoma, 487in oligoastrocytoma, and 578 in oligodendroglioma; 572 in classical, 675 in mesenchymal, and 574 in proneural. Hence, a higher number of disease-associated genes were identified in each subtype by DIAMOnD. It is observed that the size of the LCC in each subtype provided by the DIAMOnD was much larger than the LCC DEDGN (Figure 6.9A). It should be noted that DIAMOnD LCCs contain DEDGs and relevant disease genes in the network neighbourhood. There is a higher percentage of genes, i.e., almost 72–80% of seed genes, present in the LCC. The clustering coefficients of the DIAMOnD LCCs are much higher than the LCC of DEDGN (Figure 6.9B). The higher clustering coefficient of genes in the LCC shows that each subtype has a local aggregation disease gene, and these genes interact with each other more frequently than would be predicted in a random network. This finding also implies that module genes work together in biological processes and pathways and aid in the development of disease. Therefore, these disease modules can be identified as targets for precision therapy of glioma subtypes.

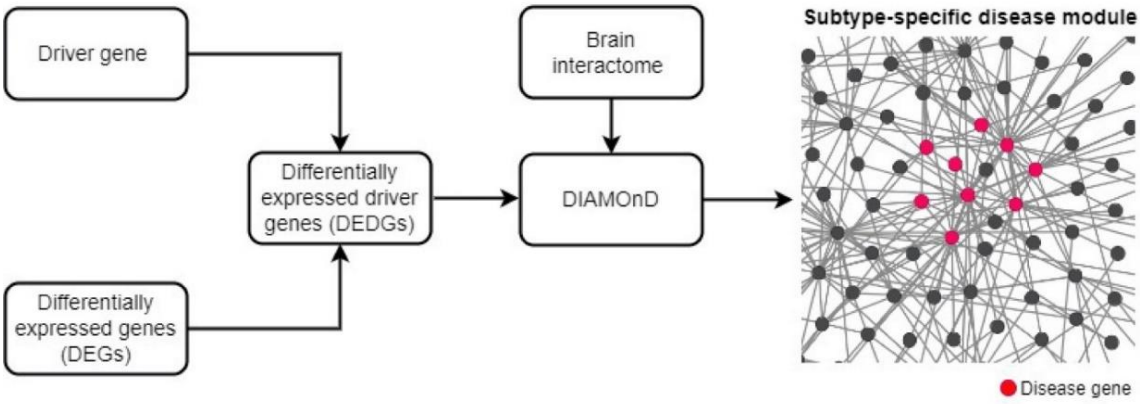

Figure 6.8: *Disease module in subtypes. The flow diagram shows the steps involved in disease module identification. DEDGs are screened from the list of driver genes and DEGs. DEDGs and brain interactome data are fed into DIAMOnD for disease module identification.*

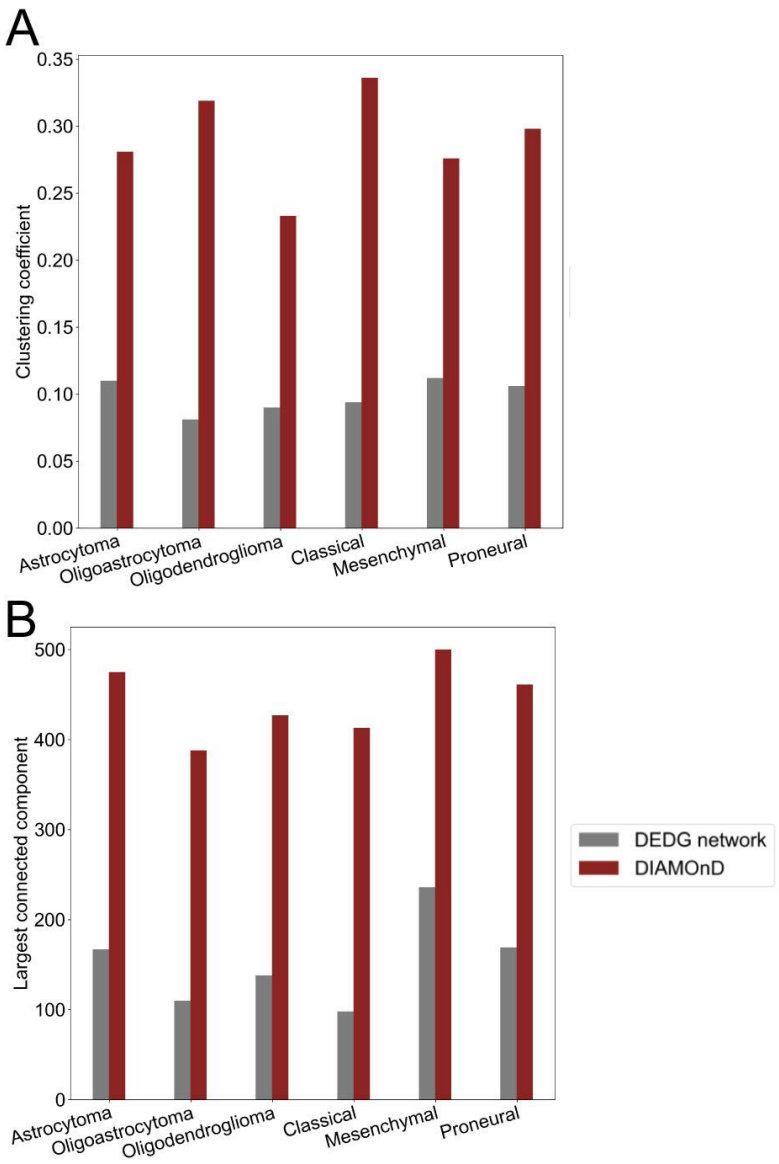

Figure. 6.9: *Disease module in subtypes.* ***(A),*** *The bar diagram compares the size of the LCC of DEDGs network (gray) and DIAMOnD disease module (brown).* ***(B),*** *The bar diagram compares clustering coefficient of LCC of DEDGs network (gray) and DIAMOnD disease module (brown).*

6.3.3 Targeting the disease module and developing the drug response prediction model

To target the disease module in glioma subtypes, the FDA-approved and investigational drugs were retrieved from the DrugBank database. The drugs for which the disease module has target genes were selected. It is observed that a total of 234, 187, 234, 178, 226, and 185 drugs

can be used to target the module genes in astrocytoma, oligoastrocytoma, oligodendroglioma, classical, mesenchymal, and proneural, respectively. Although there are targets in the modules, all these drugs may not be useful for anti-cancer therapy. Many times, drug resistance reduces the effectiveness of chemotherapy. The accurate prediction of cancer-specific drug responses is one of the significant challenges in precision medicine. Due to the genetic heterogeneity of cancers, patients' responses to cancer treatments vary depending on their distinctive genomic profiles. Due to this complexity, AI methods like ML and DL are becoming more efficient for predicting drug responses. Several large-scale drug screening programmes have made their data publicly available, such as GDSC and CCLE. These databases provide the IC50 (50% inhibitory concentration) of a particular drug on specific cancer cell lines along with cancer cell omics profiles. A lower IC50 value indicates a better sensitivity of the cell line to a given drug. Here the single-drug response classification model was developed using GDSC gene expression and driver mutation data to train and test the model, and CCLE data was used for external validation. Out of the 288 drugs that target disease modules, it is found that only 30 have experimental data on brain cancer cells. Hence, these 30 drugs were chosen to develop the drug response model. For a drug, the cell lines were classified as sensitive or resistant based on IC50 values. IC50 values at or below the 25th percentile were considered sensitive, and IC50 values at or above the 75th percentile were considered resistant for each drug. The GDSC dataset were randomly divided into 70:30 training and test sets. The model was developed on GDSC data, excluding brain cell lines. First, the gene expression and driver mutation data from GDSC were pre-processed, and feature selection was performed to reduce the multicollinearity and dimensionality of the data. The gene expression and mutation data were separately treated. Correlation-based feature selection was implemented to eliminate multicollinearity from gene expression data. The Pearson Correlation Coefficient (PCC) was computed, and genes with a $PCC > 0.5$ were dropped. The remaining 5233 genes were taken for dimensionality reduction using LASSO. LASSO feature selection was also employed on mutation data. After feature selection, both gene expression and mutation data were fed into the autoencoder with concatenated inputs (CNC-AE). Then, these two types of data were integrated and compressed in the bottleneck layer learned by the autoencoder [281,283]. All the parameters of the different layers in the autoencoder were optimized for individual drugs. However, the architecture of the autoencoder is almost the same for all drugs; for example, one hidden layer for data integration was used, and the dimension of the bottleneck layer was set to 64. An autoencoder consists of two parts: an encoder and a decoder network (Figure 6.10). The latent variables from the

bottleneck layer were employed in the decoder network to decode the original input data, and this was done in order to quantify the reconstruction loss, which represents the efficiency of the autoencoder. The mean squared error (MSE) was used to calculate the reconstruction loss. It is found that MSE was considerably lower in the range (0.02-0.19). This demonstrates that the autoencoder correctly learns to encode the pattern of gene expression and mutation in the latent space. Next, the DNN model was built to predict the drug response, i.e., whether it is sensitive or resistant, using the latent variables from the bottleneck layer of the autoencoder. In order to identify the optimized set of hyperparameters, the grid search method was employed. The average performance measures for each DNN model were then calculated using k-fold CV (k = 10). The model's performance was evaluated by computing the average accuracy, recall, specificity, precision, F1-score, FPR, GM, and MCC (see methodology). The performance matrix on training data and test data for 30 drugs is provided in the Table 6.2 and 6.3.

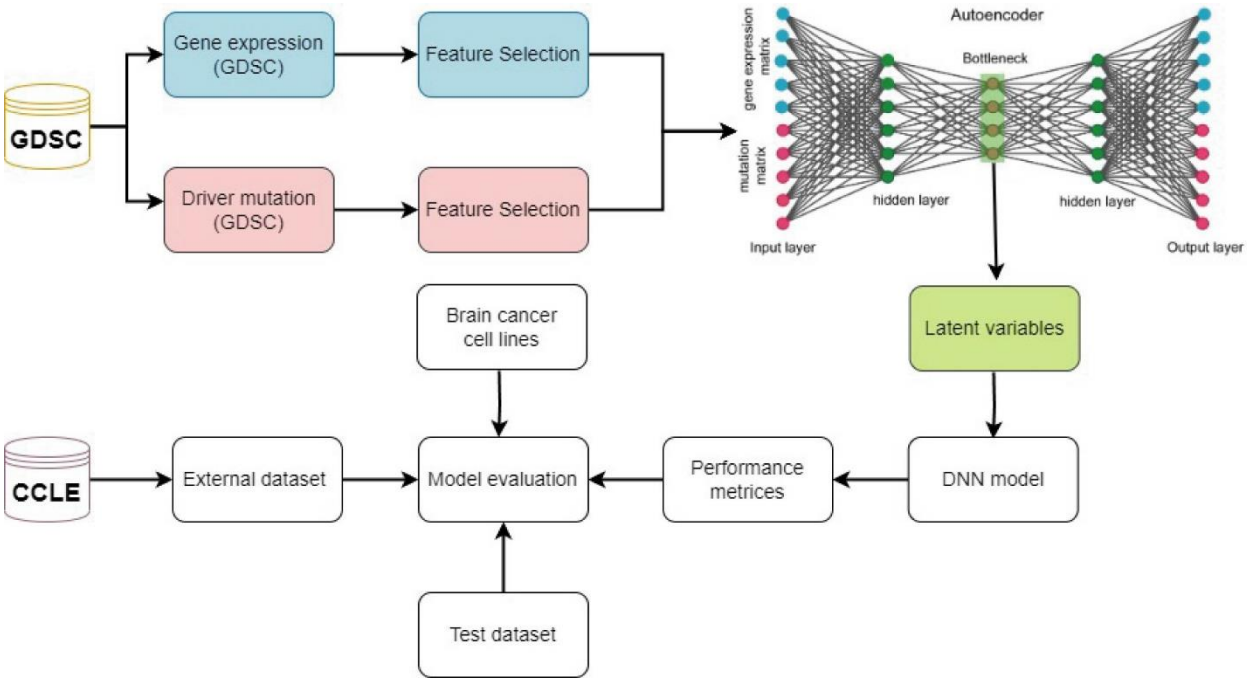

Figure 6.10: *The overall workflow of drug response model development. The gene expression and mutation data from GDSC are subjected to feature selection, and both data are integrated using an autoencoder. The latent variable from the bottleneck layer is used for developing the DNN model. The model validation was performed using test data, brain cancer data, and external data from CCLE.*

Table 6.2: Performance matrix of drug response on training data

Drugs	Accuracy	Recall	Specificity	Precision	F1-score	FPR	GM	MCC
Ruxolitinib	96.27%(±0.02)	96.26	96.15	96.55	96.23	0.03	97.49	0.92
Entinostat	95.78%(±0.01)	95.89	95.88	95.83	95.77	0.04	97.16	0.91
Lapatinib	95.62%(±0.02)	95.58	95.40	95.96	95.55	0.04	97.05	0.91
Vorinostat	94.94%(±0.02)	94.89	94.80	94.96	94.89	0.05	96.60	0.89
Tretinoin	94.50%(±0.07)	95.26	94.75	94.92	94.40	0.05	96.24	0.90
Olaparib	93.03%(±0.02)	93.22	93.05	93.03	92.99	0.06	95.29	0.86
Vinblastine	92.34%(±0.03)	92.30	92.27	92.45	92.31	0.07	94.82	0.84
Axitinib	91.80%(±0.03)	91.85	91.88	91.77	91.75	0.08	94.44	0.83
Crizotinib	91.69%(±0.01)	91.96	91.78	91.61	91.64	0.08	94.37	0.83
Trametinib	91.35%(±0.02)	91.56	91.52	91.44	91.29	0.08	94.14	0.82
Selumetinib	91.42%(±0.02)	91.37	91.31	91.48	91.38	0.08	94.19	0.82
Dasatinib	91.01%(±0.01)	91.15	91.00	91.28	90.96	0.08	93.91	0.82
Sorafenib	90.79%(±0.02)	90.69	90.53	90.76	90.68	0.09	93.75	0.81
Niraparib	90.36%(±0.02)	90.57	90.37	90.62	90.31	0.09	93.46	0.81
Rucaparib	89.77%(±0.03)	89.79	89.72	89.73	89.70	0.1	93.05	0.79
Dabrafenib	88.98%(±0.03)	88.98	88.95	88.91	88.88	0.11	92.49	0.78
Bicalutamide	87.50%(±0.05)	87.82	87.63	87.25	87.37	0.12	91.44	0.75
Bosutinib	87.16%(±0.03)	87.27	87.14	87.22	87.09	0.12	91.23	0.74
Erlotinib	86.25%(±0.03)	86.29	86.23	86.36	86.13	0.13	90.6	0.73
Nilotinib	85.80%(±0.03)	86.11	85.91	85.84	85.72	0.14	90.28	0.72
Vinorelbine	85.73%(±0.03)	85.87	85.64	85.92	85.64	0.14	90.23	0.72
Vincristine	85.52%(±0.04)	85.75	85.62	85.70	85.34	0.14	90.08	0.71
Ibrutinib	85.37%(±0.04)	85.40	85.37	85.46	85.32	0.14	89.97	0.71
Talazoparib	84.61%(±0.02)	84.64	84.55	84.57	84.52	0.15	89.45	0.69
Alpelisib	84.38%(±0.03)	84.53	84.38	84.42	84.28	0.15	89.28	0.69
Afatinib	84.16%(±0.04)	84.20	84.04	84.25	84.07	0.15	89.12	0.68
Osimertinib	84.04%(±0.02)	84.00	83.56	84.21	83.86	0.16	89.05	0.68
Gefitinib	83.37%(±0.03)	83.51	83.68	83.69	83.28	0.16	88.57	0.67
Tamoxifen	82.25%(±0.05)	82.12	81.93	82.18	82.04	0.18	87.74	0.64
Fulvestrant	70.90%(±0.03)	71.14	70.97	71.11	70.82	0.29	79.50	0.42

Table 6.3: Performance matrix of drug response on test data

Drugs	Accuracy	Recall	Specificity	Precision	F1-score	FPR	GM	MCC	AUC
Ruxolitinib	94.12	94.12	94.22	94.74	94.10	0.05	96.04	0.88	0.97
Entinostat	94.96	94.96	94.87	95.42	94.94	0.05	96.61	0.90	0.95
Lapatinib	96.09	96.09	96.09	96.38	96.09	0.03	97.38	0.92	0.98
Vorinostat	91.27	91.27	91.27	91.36	91.26	0.08	94.09	0.82	0.91
Tretinoin	82.76	82.76	81.53	87.07	82.12	0.18	88.15	0.69	0.82
Olaparib	93.75	93.75	93.75	93.92	93.74	0.06	95.79	0.87	0.95
Vinblastine	85.87	85.87	85.87	85.89	85.87	0.14	90.34	0.71	0.86
Axitinib	91.27	91.27	91.27	91.36	91.26	0.08	94.09	0.82	0.91
Crizotinib	94.53	94.53	94.53	94.63	94.53	0.05	96.32	0.89	0.95
Trametinib	89.84	89.84	89.84	89.85	89.84	0.10	93.11	0.79	0.90
Selumetinib	87.78	87.78	87.75	87.93	87.77	0.12	91.68	0.75	0.88
Dasatinib	92.91	92.91	92.90	92.92	92.91	0.07	95.22	0.85	0.96
Sorafenib	92.86	92.86	92.86	93.13	92.85	0.07	95.18	0.85	0.97
Niraparib	89.83	89.83	89.83	90.01	89.82	0.10	93.1	0.79	0.93
Rucaparib	88.8	88.80	88.77	88.84	88.8	0.11	92.39	0.77	0.90
Dabrafenib	85.71	85.71	85.71	85.75	85.71	0.14	90.23	0.71	0.86
Bicalutamide	86.21	86.21	85.70	86.85	86.11	0.14	90.58	0.72	0.86
Bosutinib	85.60	85.60	85.50	86.18	85.53	0.14	90.15	0.71	0.86
Erlotinib	82.54	82.54	82.54	82.57	82.54	0.17	87.99	0.65	0.83
Nilotinib	83.33	83.33	83.33	83.75	83.28	0.16	88.55	0.67	0.83
Vinorelbine	85.71	85.71	85.71	85.75	85.71	0.14	90.23	0.71	0.86
Vincristine	85.42	85.42	85.42	85.98	85.36	0.14	90.02	0.71	0.85
Ibrutinib	83.76	83.76	83.80	83.86	83.75	0.16	88.86	0.67	0.85
Talazoparib	84.13	84.13	84.13	84.13	84.13	0.15	89.12	0.68	0.84
Alpelisib	76.56	76.56	76.56	76.67	76.54	0.23	83.69	0.53	0.77
Afatinib	84.38	84.38	84.38	84.51	84.36	0.15	89.29	0.68	0.84
Osimertinib	84.13	84.13	84.13	84.27	84.11	0.15	89.12	0.68	0.85
Gefitinib	84.92	84.92	84.92	86.02	84.80	0.15	89.68	0.70	0.85
Tamoxifen	82.54	82.54	82.54	82.57	82.54	0.17	87.99	0.65	0.85
Fulvestrant	73.54	73.54	73.52	73.58	73.53	0.26	81.48	0.47	0.74

Based on the performance parameters, the Ruxolitinib drug had an accuracy to predict sensitivity or resistance was 96.26% (±0.02). The precision and specificity of the model were >0.90. Due to the superior performance of the Ruxolitinib model, further investigation was done and found that it is a potent inhibitor of the JAK/STAT signaling pathway and can inhibit the invasion and tumorigenesis of glioma cells [304]. This drug is also in clinical trials for glioma treatment (https://clinicaltrials.gov/). Further, It is found that the accuracy of prediction using

test data was 94.12% and that using only brain cancer cell lines was 84.28%. It is tempting to state that the model prediction was as per the independent observations made by other researchers. Figure 6.11 A and B show that all models for the top 10 drugs have higher accuracy of prediction using training (91.34–96.26%), test (82.76–96.09%) data.

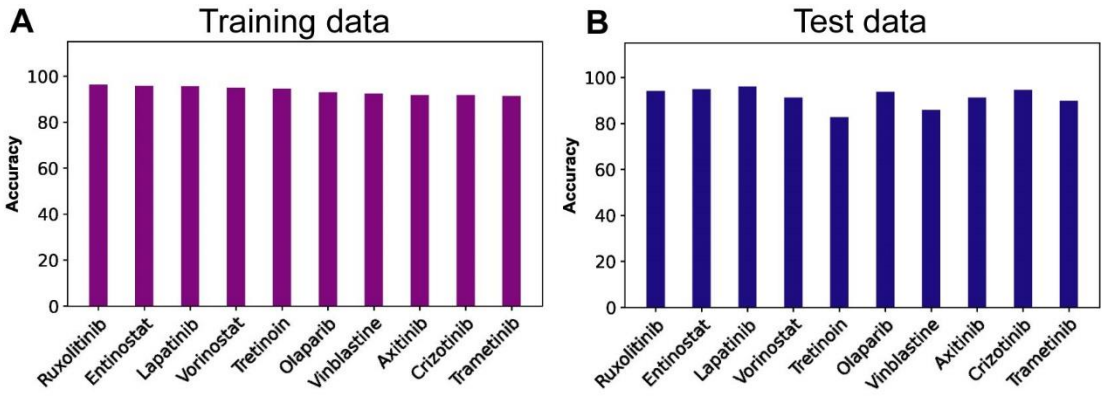

Figure 6.11: *Classification accuracy of DL models (**A),** training (**B),** test dataset of top 10 drugs.*

A highly sensitive and specific model for drug response prediction is always ideal. Therefore, the receiver operating characteristic (ROC) curve was used to illustrate the sensitivity and specificity of each model. For a range of different cutoff points, the ROC curve compares the probability of a true positive result, or the test's sensitivity, to the probability of a false positive result. Figure 6.12 (A-J) shows the area under the ROC curve (AUC) of the DNN models of the top 10 drugs. It is observed that the AUC values were high, i.e., 0.97 in Ruxolotinib, 0.95 in Entinostat, 0.98 in Lapatinib, 0.91 in Vorinostat, and 0.95 in Olaparib. Overall, all models show consistent prediction accuracy in training, testing, and brain cancer cell data. The performance matrix on brain cancer cell line data for 30 drugs is provided in the Table 6.4. The model accuracy >80% was obtained using only brain cancer cell line data (Figure 6.13).

ROC plots

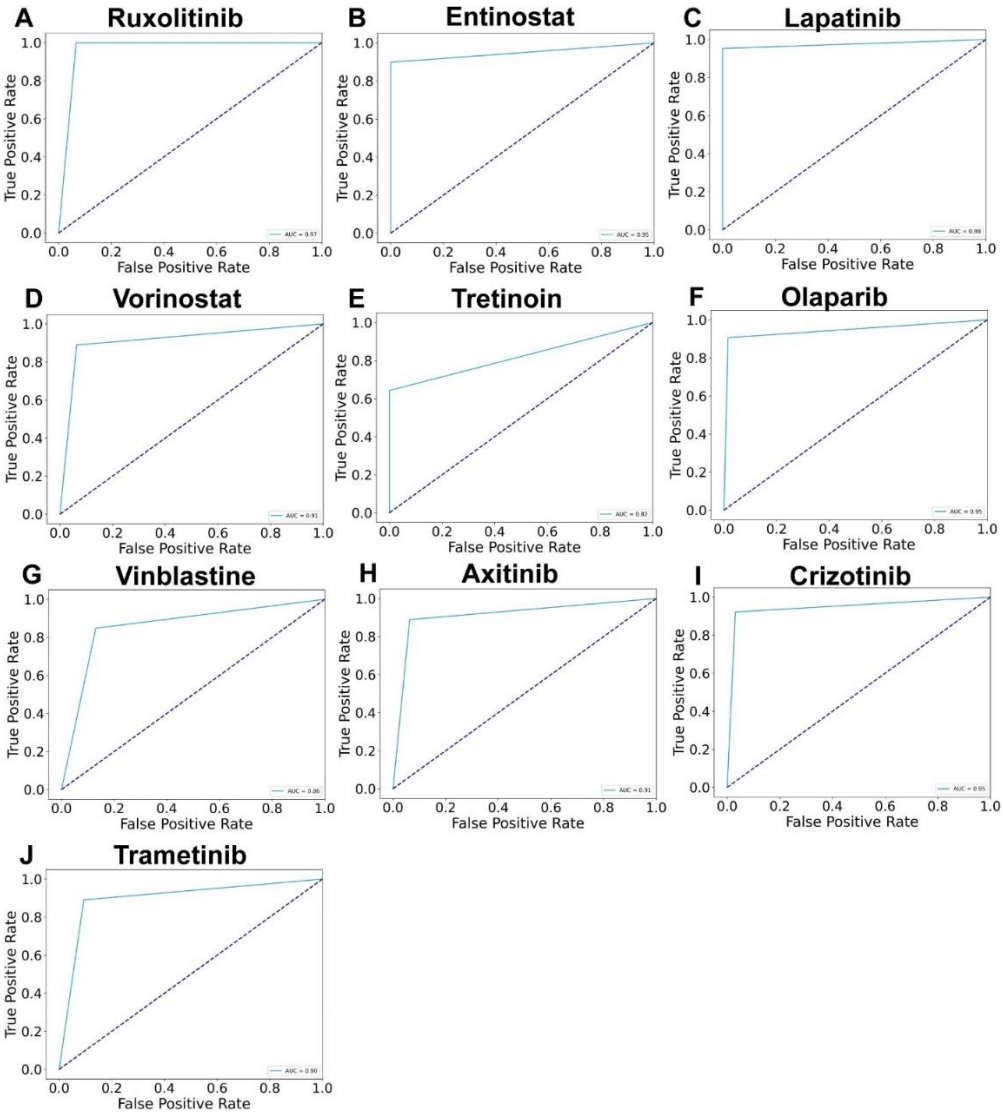

***Figure 6.12: (A-J)**, ROC plots of the top 10 drugs.*

Table 6.4: Performance matrix of drug response on brain cancer cell lines

Drugs	Accuracy	Recall	Specificity	Precision	F1-score	FPR	GM	MCC
Lapatinib	91.25%(±0.13)	93.38	93.63	92.69	91.23	0.06	93.86	0.86
Vinblastine	90.00%(±0.19)	90.08	89.09	90.55	87.66	0.1	92.66	0.79
Niraparib	90.00%(±0.11)	92.51	91.16	90.47	89.47	0.08	93.05	0.83
Rucaparib	88.75%(±0.12)	90.54	92.46	90.7	88.43	0.07	92.16	0.81
Ruxolitinib	84.29%(±0.10)	84.67	82.83	86.8	82.84	0.17	89.09	0.7
Entinostat	84.29%(±0.10)	85.4	86.93	85.6	83.77	0.13	89.09	0.7
Vorinostat	82.50%(±0.10)	85.46	84.04	84.89	81.5	0.15	87.82	0.7
Olaparib	81.25%(±0.13)	84.46	84.54	84.86	80.47	0.15	86.81	0.69
Sorafenib	80.00%(±0.12)	82	80.67	81.75	78.51	0.19	85.98	0.63
Dasatinib	80.00%(±0.06)	81.67	82	81.04	79.05	0.18	86.12	0.62
Selumetinib	79.17%(±0.09)	80.64	79.59	79.08	78.52	0.2	85.47	0.6
Talazoparib	77.50%(±0.11)	79.38	82.13	81.25	76.84	0.17	84.19	0.61
Trametinib	77.50%(±0.11)	80.17	79.83	77.84	76.82	0.2	84.19	0.58
Axitinib	76.25%(±0.13)	81.54	79.46	80.53	76.11	0.2	83.17	0.61
Osimertinib	76.25%(±0.13)	79.54	78.8	79.02	74.98	0.21	83.19	0.58
Ibrutinib	72.86%(±0.12)	74.06	77.61	77.13	71.88	0.22	80.73	0.5
Dabrafenib	72.50%(±0.19)	76.13	78.71	73.73	70.62	0.21	79.96	0.55
Tamoxifen	72.50%(±0.12)	74.63	75.21	75	71.63	0.24	80.44	0.49
Crizotinib	72.50%(±0.07)	75.5	73.83	75.28	71.49	0.26	80.61	0.49
Vinorelbine	68.75%(±0.16)	73	72.83	74.9	67.69	0.27	77.38	0.47
Nilotinib	68.75%(±0.08)	71.58	70.92	71.66	67.87	0.29	77.77	0.42
Afatinib	67.50%(±0.20)	72.54	75.29	72.68	65.77	0.24	76.22	0.48
Bosutinib	66.25%(±0.22)	70.54	70.79	68.32	64.63	0.29	75.04	0.43
Vincristine	66.00%(±0.18)	73.7	73.8	69.7	61.77	0.26	75.16	0.46
Bicalutamide	63.33%(±0.33)	62.33	62.35	62	60.8	0.32	70.14	0.45
Fulvestrant	61.67%(±0.15)	61.13	61.65	62.1	60.33	0.38	71.97	0.24
Erlotinib	60.00%(±0.18)	66.92	67.42	67.24	59.34	0.32	70.37	0.36
Alpelisib	58.75%(±0.26)	65.17	64.33	69.57	57.47	0.35	67.23	0.34
Gefitinib	57.50%(±0.15)	63.46	61.38	59.82	55.86	0.38	68.63	0.25
Tretinoin	56.67%(±0.38)	57.67	57.32	55	53.13	0.4	62.51	0.2

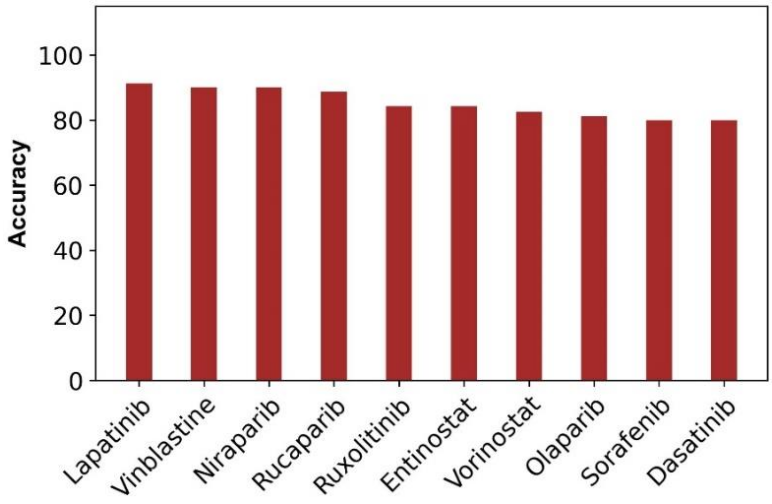

Figure 6.13: *Classification accuracy of DL models on test dataset of brain cancer cell lines of top 10 drugs.*

From earlier articles, it was also learned that several of the top 10 drugs in the training data exhibit promising anti-glioma cell activities, such as Entinostat, a histone deacetylase inhibitor, which can inhibit GBM growth [305]. Vorinostat, an FDA-approved drug, is already in use to treat cutaneous T-cell lymphoma, but it is now in a Phase II clinical trial to treat recurrent glioblastoma multiforme [306]. Vinblastine shows sensitivity for both LGG and GBM [307,308] and it is in clinical trials for the treatment of these cancers. Other drugs such as Olaparib [309], Crizotinib [310], Trametinib [311] have also shown encouraging results for the treatment of brain cancer. Our findings, along with those from the existing literature, suggest that the current approach may be used to aid in clinical decision-making for the treatment of gliomas. We forecast the drug sensitivity of 30 different drugs against 49 brain cancer cell lines using the saved models to assess their potential clinical utility. The features from gene expression and mutation data from particular cell line data were extracted, and integrated these two data sets, i.e., gene expression and mutation, using an autoencoder, and then fed this integrated data into the 30 different drug-specific DNN models. Lastly, the sensitivity or resistance of a drug against a particular brain cancer cell line were predicted. The drug sensitivity data for 10 drugs for 49 brain cancer cells is shown in Figure 14. The cell line's lineage from ATCC (https://www.atcc.org/) and cellosaurus (https://www.cellosaurus.org/) was acquired in order to demonstrate subtype-specific drug sensitivity. We were able to provide the drug sensitivity results for oligodendroglioma, astrocytoma, and GBM based on the data that was available. It

is found that the drug sensitivity of various cell types varied, and a major factor contributing to this variation is the cell line's genetic background, including both gene expression and mutation. Indeed, gene expression and mutations as features was used while developing the models. This provides a comprehensive view of the significance of subtype-specific drug response prediction utilizing genomic data in enhancing the clinical efficacy of the therapy.

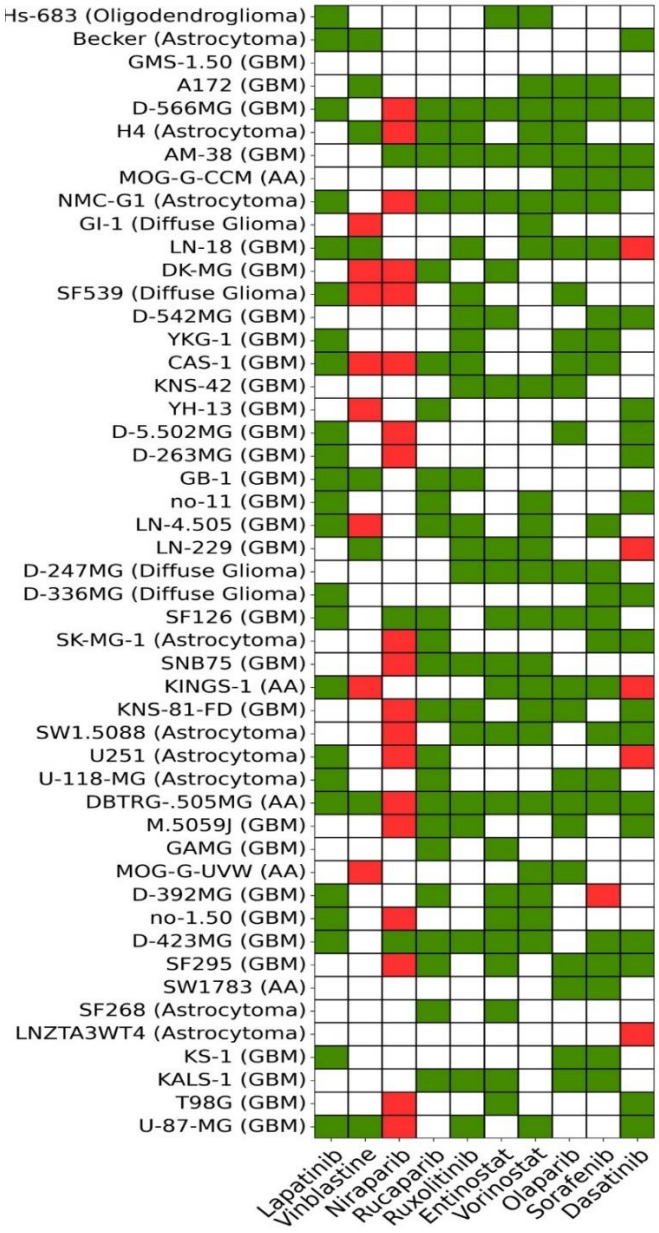

Figure 6.14: *Prediction of drug sensitivity in brain cancer cell lines. The heat map represents the drug sensitivity data for 49 brain cancer cell lines against 30 drugs. The red color indicates the resistant cell lines and the green color indicates the sensitive cell lines. The origin (or subtype) of each of the cell lines is mentioned in the figure.*

Lastly, this DL framework with external datasets was validated from CCLE, and the accuracy of prediction for drugs, Erlotinib, Lapatinib, Nilotinib, and Sorafenib was fairly accurate (Table 6.5). These results show that our models were able to consistently predict accurate drug responses. However, experimental investigations, i.e., in vitro and in vivo drug efficacy assays and sensitivity data across the many cancer cell types, need to be analyzed before the current framework is put into use in a clinical context.

Table 6.5: Model validation on external dataset from CCLE

Drugs	Accuracy	Recall	specificity	precision	F1	FPR	GM	MCC
Erlotinib	74.00 % (±0.21)	77.27	79.39	72.87	70.95	0.2	81.06	0.57
lapatinib	70.00 % (±0.16)	71.06	72.27	75.08	68.86	0.27	78.35	0.46
Nilotinib	62.50 % (±0.35)	63.43	64.36	64.06	61.5	0.35	69	0.45
Sorafenib	88.00 % (±0.10)	88.79	89.55	90.68	87.53	0.1	91.7	0.79

6.4 DISCUSSION

The clinical development of targeted and personalized brain cancer treatments continues to be a significant issue. There are many different types of brain cancer, and the fact that they each possess their own unique genetic abnormalities makes it challenging to design effective treatments. Finding a disease-specific biomarker for targeted therapy is a commonly used strategy. However, due to the molecular heterogeneity of cancer, targeted therapy is not always effective in treatment and frequently develops drug resistance. To address this, the current study combines network medicine-based techniques with DL-based drug response prediction to target glioma subtypes for precision therapy. Among all cancer-associated alterations, driver mutations and altered gene expression are majorly involved in oncogenic transformation [312]. Therefore, genome-wide screening of driver mutations was performed and identified the DEGs from transcriptome data in each subtype of LGG and GBM. **From the list of driver mutations and DEGs, the DEDGs were identified, which are further subjected to disease module identification.** Cancer is not a single gene disorder; rather, the interaction between many genes causes cancer. Hence, **the identification of disease modules using DEDGs can comprehensively represent the core structure of the subtype-specific network associated with the cancer phenotype.** The network medicine-based approach demonstrates that effective drugs must target the protein within or in the disease modules' immediate vicinity. Therefore,

drugs from the DrugBank database to target these disease modules were selected. Patients' responses to drugs, however, differ greatly from one another due to the diversity of molecular profiles. To address this further, a DL-based framework was developed to predict drug response using gene expression, mutation, and IC50 values from large-scale experimental data. The novel framework was designed by combining LASSO-based feature selection, autoencoder-based data integration, and then prediction using the DNN. It is noticed the consistent performance of the model in test data, brain cancer cell lines, and validation data. To examine the clinical application, we predict the drug response for each brain cancer cell line using a drug-specific model. Additionally, **it is shown that cancer cell lines from various subtypes of glioma exhibit varying degrees of drug sensitivity.** Earlier, several studies reported the drug response model for a particular cancer type, but in our study, **we have shown that models can be used to predict drug response for a specific subtype of cancer.**

CHAPTER 7

CONCLUSION AND FUTURE SCOPE

Chapter 7: Conclusion and future scope

In this thesis, discusses the ML and DL models that were utilized in the field of cancer genomics, and focusing on their respective architectures. ML & DL has the enormous potentiality to assist clinicians by reducing human error, helping in cancer diagnosis, and analyzing complex data. Therefore, it is proved as cutting-edge technology in cancer research. In this thesis, a comprehensive and precise ML-based approach was presented for cancer grading and subtyping. It is found an integrated approach consisting of correlation and SVF-RFE algorithm for feature gene selection, and then computation of SVM using those feature genes (n = 100) had shown superior performance (accuracy > 90%). It is found that the accuracy of subtype classification is always good using the gene expression data of a specific grade of cancer rather than a mixed grade. It is observed that other ML techniques produced repeatedly the same results. This gave us clues that cancer grading is essential to achieve higher accuracy for subtype prediction. It is also observed six-class classification for simultaneous grading and subtyping using the same ML framework and attained an overall accuracy of 91.0% (±0.02) and AUC=0.88. Therefore, the findings of this study strongly strengthen the fact that grading and subtyping are both required to achieve a higher accuracy of prediction. The correct set of feature genes and their discriminative ability play a crucial role in the superior performance of ML algorithms. In addition, the biological relevance of these features could lead to finding the mechanisms behind LGG formation and therapeutic targets. The subtype and grade-specific co-expressed feature genes associated with the oncogenesis was identified. Furthermore, survival analysis of these genes revealed several predictive biomarkers, which could be used as potential molecular indicators for diagnosis and treatment. Therefore, we conclude that gene expression data of a subtype of LGG without considering the grade is more heterogeneous than data of a specific grade. Further, the study of chapter 4 indicates that DL and ML can be powerful tools for finding patterns in large-scale genetic and epigenetic data sets related to human cancer. Here, a biologically relevant DL and ML-based framework was presented to classify the subtype of GBM to increase accuracy in diagnosis; in turn, it can lead to better patient management. Here, the successful separation of three subtypes of glioblastoma multiforme (GBM), namely classical, mesenchymal, and proneural, has been performed with a classification accuracy >90%. It is also compared DL and ML techniques to identify the most

suitable method for interpreting the transcriptome, methylome, and integrated data. DL method, i.e., CNN outperforms other ML models. It is observed that overall classification performance was higher using the transcriptome and integrated data than the methylome data. Another significant aspect of our findings is the biological relevance of features and the identification of subtype-specific prognostic biomarkers. To find the association of features genes with specific subtypes, we performed WGCNA. Furthermore, several genes present in these co-expressed modules was identified, which were linked to patient survival. Our study explained how the features genes from the DL/ML framework could be used to find the subtype-specific biomarkers. The feature genes of this study and CNN can provide assured and clinically relevant deep learning-based diagnostic tools for the proper treatment of GBM patients. These results indicate that DL is better than the ML algorithms. However, development of DL-based model with large scale multi-omics data can improve the overall precision and efficacy of diagnostic processes. However, clinical diagnosis still raises questions about the validity and interpretability of DL- or AI-based diagnostic models. In general, efficient DL and ML tools work like a ‘black -box’; researchers or clinicians may not be confident in diagnosing or classifying cancer patients using these approaches. However, if the basis of classification is biologically relevant and has higher accuracy, the diagnosis and patient management will be more assured and systematic. To promote the further development for building more accurate biological relevant models and identification of novel therapeutic marker multi-omics data analysis is essential, which has grown in popularity in cancer research in recent decades. Moreover, the integration of transcriptomic, mutational, and methylome data can reveal the intricate systemic dysregulation linked to the phenotype of glioma.

Therefore, it is essential to design a biologically and clinically relevant AI-based diagnostic model to increase the reliability of diagnosis. Hence, in the chapter5 the AI-based diagnostic tool was designed, i.e., DeepAutoGlioma, for subtyping the glioma. The transcriptome and methylome data of glioma patients were used to extract biologically and clinically relevant features for model development. The features from two levels of genomic layers were integrated to capture cancer-specific patterns for accurate subtyping. Integration of omics data enables us to achieve greater model performance because it provides a wealth of information from different genomic layers. The model developed based on multi-omics data can greatly support the clinician in personalizing treatment. Here, in chapter 6 the clinical development of targeted and personalized brain cancer treatments continues to be a significant issue. Finding a disease-specific biomarker for targeted therapy is a commonly used strategy. In this study, network

medicine-based techniques with DL-based drug response prediction was combined to target glioma subtypes for precision therapy. Therefore, gene expression and mutational profiles was integrated and performed genome-wide screening of driver mutations and identified the DEGs from transcriptome data in each subtype of LGG and GBM. From the list of driver mutations and DEGs, the DEDGs were identified, which are further subjected to disease module identification. Hence, the identification of disease modules using DEDGs can comprehensively represent the core structure of the subtype-specific network associated with the cancer phenotype. Therefore, drugs from the DrugBank database to target these disease modules were selected. Next, a DL-based framework was developed to predict drug response using gene expression, mutation, and IC50 values from large-scale experimental data. The novel framework by combining LASSO-based feature selection, autoencoder-based data integration was designed, and then prediction using the DNN was performed. It is noticed the consistent performance of the model in test data, brain cancer cell lines, and validation data. Additionally, it is shown that cancer cell lines from various subtypes of glioma exhibit varying degrees of drug sensitivity. Due to the limitations of the dataset and lack of information on cell lineage, we were unable to predict the drug response for all subtypes of LGG and GBM. But It is expected that this problem will be solved soon because the size of datasets is growing rapidly.

This thesis has enlightened with various aspects and use of ML and DL models from brain cancer diagnosis to the development of precision medicine. The Superior accuracies of ML and DL in each type of genomic data show the possibility to develop a robust AI model from heterogeneous data of cancer patients. The AI-models discussed in this thesis were developed using data from brain cancer tissue. AI developers and cancer biologists should focus on the data generated from liquid biopsies samples using non-invasive techniques, such as blood, saliva, serum, and urine. Data from liquid biopsies samples will facilitate the biomarker identification at an early stage of brain cancer. Furthermore, it will be less complicated for multiple time sample collection to evaluate the patient's response to the treatment. For complex diseases like cancer, combining the approaches of network medicine and DL-based drug response prediction presents enormous promise for the development of novel and efficient treatments. Network medicine can reveal the complex molecular interactions in the disease state, which can lead to the identification of novel drug targets, whereas DL can extract hidden patterns from large-scale omics data to develop a predictive model to determine the patient-specific therapeutic approach. It is believed that the present work can be extended to other types

of cancer to find subtype-specific targets and predict the drug response, and that it can contribute to developing personalized medicine and improving patient outcomes.

Appendix I

Subtyping and grading of lower-grade glioma (LGG)

Table I.1: Performance of various machine learning models for subtype classification using different sets of feature genes

		Mixed grade					Grade2					Grade3				
	Number of Feature	**20**	**50**	**100**	**200**	**500**	**20**	**50**	**100**	**200**	**500**	**20**	**50**	**100**	**200**	**500**
SVM	**Accuracy**	0.7	0.74	0.8	0.73	0.72	0.85	0.88	0.93	0.83	0.81	0.88	0.86	0.96	0.88	0.85
	F1 score	0.7	0.74	0.8	0.73	0.72	0.85	0.88	0.92	0.83	0.81	0.88	0.86	0.96	0.87	0.85
	Precision	0.71	0.74	0.81	0.72	0.72	0.86	0.88	0.93	0.84	0.82	0.88	0.86	0.96	0.87	0.86
	AUC	0.85	0.86	0.87	0.9	0.91	0.9	0.93	0.98	0.94	0.94	0.95	0.95	0.95	0.98	0.98
KNN	**Accuracy**	0.66	0.67	0.67	0.64	0.6	0.68	0.72	0.73	0.72	0.68	0.71	0.75	0.8	0.71	0.65
	F1 score	0.66	0.67	0.67	0.64	0.6	0.69	0.7	0.74	0.72	0.7	0.68	0.72	0.8	0.72	0.64
	Precision	0.67	0.68	0.67	0.64	0.6	0.69	0.71	0.75	0.72	0.72	0.72	0.73	0.81	0.73	0.64
	AUC	0.71	0.71	0.72	0.7	0.76	0.85	0.75	0.73	0.82	0.86	0.82	0.8	0.84	0.8	0.77
GaussianNB	**Accuracy**	0.67	0.72	0.72	0.66	0.65	0.7	0.75	0.83	0.78	0.73	0.73	0.75	0.93	0.78	0.75
	F1 score	0.66	0.72	0.71	0.66	0.65	0.7	0.73	0.82	0.77	0.72	0.71	0.74	0.93	0.77	0.74
	Precision	0.68	0.72	0.71	0.67	0.65	0.7	0.73	0.83	0.79	0.73	0.72	0.74	0.93	0.79	0.75
	AUC	0.78	0.8	0.8	0.77	0.76	0.84	0.89	0.91	0.92	0.9	0.83	0.88	0.89	0.87	0.87
Decision tree	**Accuracy**	0.6	0.62	0.62	0.6	0.57	0.68	0.73	0.75	0.68	0.6	0.63	0.66	0.68	0.61	0.6
	F1 score	0.59	0.6	0.62	0.6	0.57	0.68	0.72	0.75	0.68	0.6	0.63	0.63	0.67	0.61	0.58
	Precision	0.61	0.61	0.64	0.63	0.62	0.69	0.72	0.76	0.72	0.63	0.71	0.73	0.67	0.64	0.61
	AUC	0.72	0.72	0.65	0.61	0.71	0.71	0.71	0.71	0.75	0.79	0.77	0.7	0.73	0.77	0.73
Random forest	**Accuracy**	0.7	0.72	0.74	0.68	0.66	0.78	0.85	0.88	0.78	0.75	0.76	0.86	0.88	0.83	0.76
	F1 score	0.7	0.72	0.74	0.68	0.66	0.77	0.84	0.87	0.78	0.75	0.76	0.86	0.87	0.83	0.75
	Precision	0.7	0.73	0.74	0.69	0.66	0.79	0.83	0.88	0.81	0.78	0.76	0.86	0.9	0.83	0.77
	AUC	0.81	0.82	0.84	0.85	0.86	0.9	0.91	0.95	0.93	0.95	0.91	0.93	0.95	0.94	0.94

Table I.2: External dataset sample details

Dataset	Subtype	Original sample Number	Sample number after random sampling
GSE74462 (Grade2)	Astrocytoma	2	11
	Oligoastrocytoma	11	11
	Oligodendroglioma	1	11
GSE43378 (Grade3)	Astrocytoma	12	12
	Oligoastrocytoma	2	12
	Oligodendroglioma	4	12

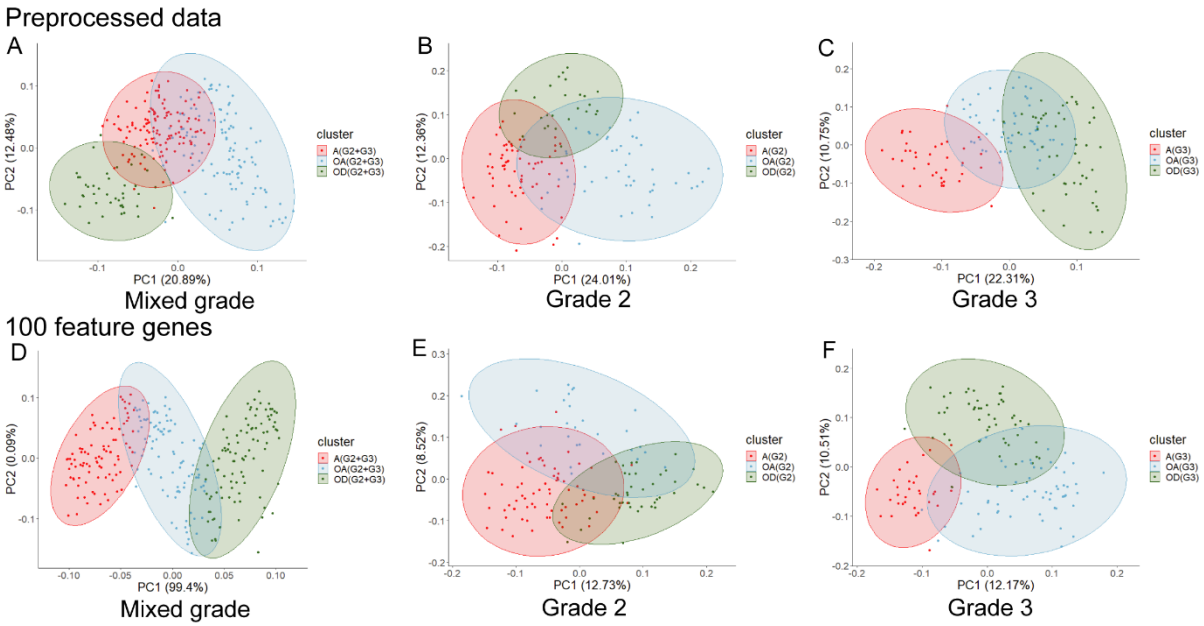

Figure I.1: PCA of preprocessed data and feature genes. (A, B, C) dot plots show the PCA using preprocessed gene expression data. (D, E, F) show the PCA using expression data of 100 feature genes. A: astrocytomas, OA: oligoastrocytomas, OD: oligodendrogliomas, N: healthy, G2: grade2, and G3: grade3, and G2+G3: mixed grade.

Table I.3: Performance of subtype classification using Boruta feature selection method

	Grade	Mixed grade	Grade2	Grade3
	Number of Feature	**132**	**209**	**170**
SVM	**Accuracy**	0.7172	0.8525	0.8333
	F1 score	0.7102	0.8405	0.8338
	AUC	0.69	0.91	0.89
	Precision	0.7213	0.8508	0.846
KNN	**Accuracy**	0.6465	0.7541	0.7333
	F1 score	0.6509	0.7307	0.7352
	AUC	0.76	0.86	0.78
	Precision	0.6846	0.7533	0.757
GaussianNB	**Accuracy**	0.6667	0.8033	0.75
	F1 score	0.6646	0.7974	0.7315
	AUC	0.78	0.89	0.83
	Precision	0.6677	0.8091	0.7519
Decision tree	**Accuracy**	0.6061	0.6885	0.7167
	F1 score	0.5941	0.689	0.709
	AUC	0.66	0.8	0.72
	Precision	0.6009	0.7343	0.742
Random forest	**Accuracy**	0.7071	0.8852	0.90
	F1 score	0.7035	0.8811	0.897
	AUC	0.83	0.95	0.91
	Precision	0.7109	0.8801	0.9002

Appendix II

Subtyping of glioblastoma multiforme (GBM)

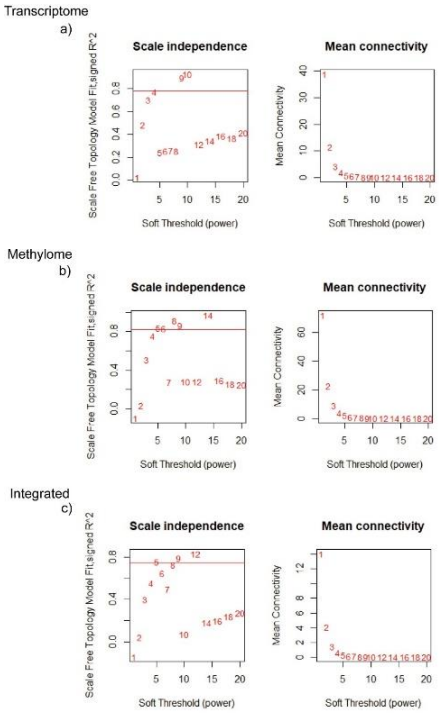

Figure II.1: Analysis of network topology for several soft thresholding power (β) in WGCNA

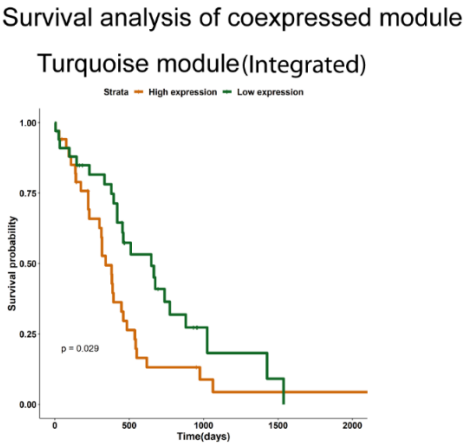

Figure II.2: Survival analysis of positively associated module. Overall survival was analyzed based on quartile method of 75 % cut-off of higher and 25% cut-off of lower limit.

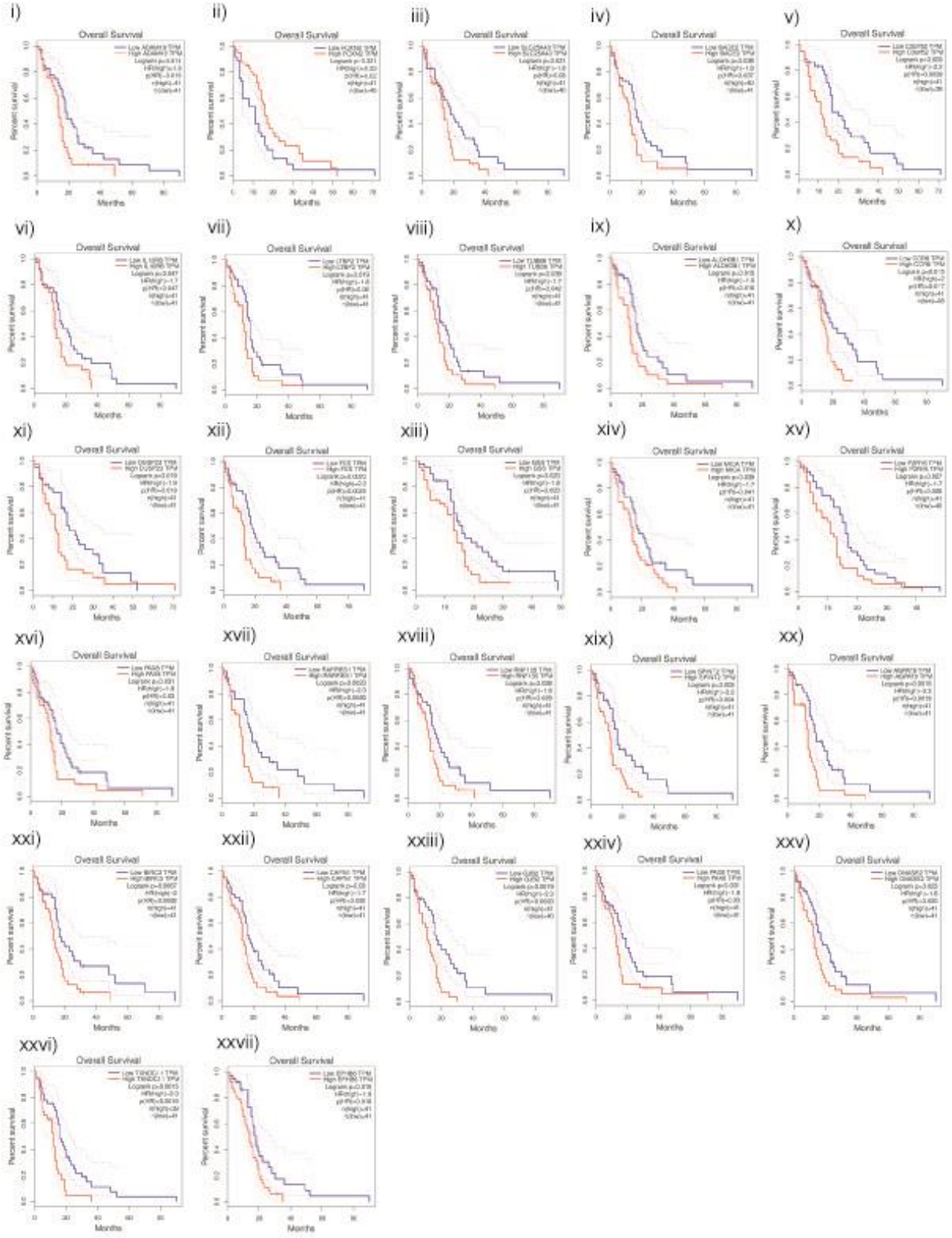

Figure II.3: (Extended figure of Fig.5) Survival analysis using GEPIA of gene present in coexpression module. (i to iii), Kaplan-Meier plots of genes from positively associated modules with classical subtype. (iv to xxiv), Kaplan-Meier plots of gene from positively associated modules with mesenchymal subtype. (xxv to xxvii), Kaplan-Meier plots of gene from positively

associated modules with proneural subtype. Overall survival was analyzed based on quartile method of 75 % cut-off of higher and 25% cutoff of lower limit.

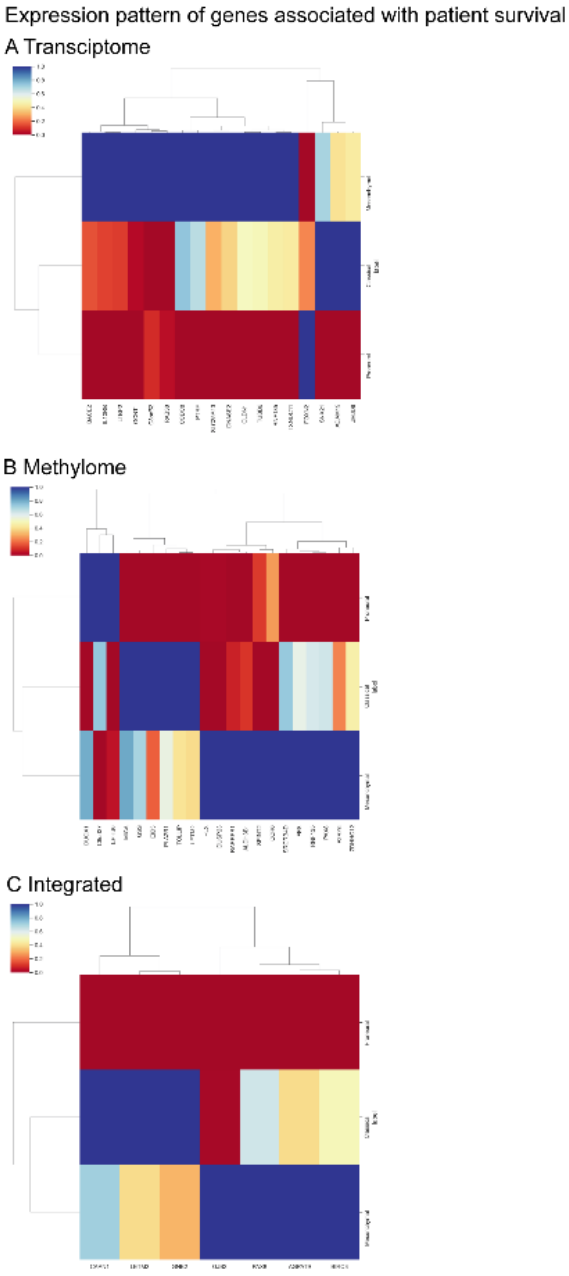

Figure II.4: Expression pattern of genes associated with patient survival. Heatmaps show the genes present in coexpressed modules of (A) transcriptome (B) methylome (C) integrated data.

REFERENCES

1. Gould, J. Breaking down the epidemiology of brain cancer. *Nature* **561**, S40–S41 (2018).

2. Louis, D. N. *et al.* The 2021 WHO Classification of Tumors of the Central Nervous System: a summary. *Neuro Oncol* **23**, 1231–1251 (2021).

3. Qazi, M. A., Bakhshinyan, D. & Singh, S. K. Deciphering brain tumor heterogeneity, one cell at a time. *Nat Med* **25**, 1474–1476 (2019).

4. Perry, A. & Wesseling, P. Histologic classification of gliomas. *Handb Clin Neurol* **134**, 71–95 (2016).

5. Maintz, D. *et al.* Molecular genetic evidence for subtypes of oligoastrocytomas. *J Neuropathol Exp Neurol* **56**, 1098–1104 (1997).

6. DJ, B. *et al.* Comprehensive, Integrative Genomic Analysis of Diffuse Lower-Grade Gliomas. *N Engl J Med* **372**, 2481–2498 (2015).

7. Ostrom, Q. T. *et al.* CBTRUS statistical report: Primary brain and central nervous system tumors diagnosed in the United States in 2006-2010. *Neuro Oncol* **15 Suppl 2**, (2013).

8. Wang, Q. *et al.* Tumor Evolution of Glioma-Intrinsic Gene Expression Subtypes Associates with Immunological Changes in the Microenvironment. *Cancer Cell* **32**, 42-56.e6 (2017).

9. Witthayanuwat, S. *et al.* Survival Analysis of Glioblastoma Multiforme. *Asian Pac J Cancer Prev* **19**, 2613–2617 (2018).

10. Zhang, P., Xia, Q., Liu, L., Li, S. & Dong, L. Current Opinion on Molecular Characterization for GBM Classification in Guiding Clinical Diagnosis, Prognosis, and Therapy. *Front Mol Biosci* **7**, (2020).

11. Mallik, S., Seth, S., Bhadra, T. & Zhao, Z. A Linear Regression and Deep Learning Approach for Detecting Reliable Genetic Alterations in Cancer Using DNA Methylation and Gene Expression Data. *Genes (Basel)* **11**, 1–15 (2020).

12. Crucitta, S. *et al.* Treatment-driven tumour heterogeneity and drug resistance: Lessons from solid tumours. *Cancer Treat Rev* **104**, (2022).

13. Lim, Z. F. & Ma, P. C. Emerging insights of tumor heterogeneity and drug resistance mechanisms in lung cancer targeted therapy. *J Hematol Oncol* **12**, (2019).

14. Sung, H. *et al.* Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin* **71**, 209–249 (2021).

15. Eoli, M. *et al.* Reclassification of oligoastrocytomas by loss of heterozygosity studies. *Int J Cancer* **119**, 84–90 (2006).

16. Kim, Y. H. *et al.* Molecular classification of low-grade diffuse gliomas. *Am J Pathol* **177**, 2708–2714 (2010).

17. Sahm, F. *et al.* Farewell to oligoastrocytoma: in situ molecular genetics favor classification as either oligodendroglioma or astrocytoma. *Acta Neuropathol* **128**, 551–559 (2014).

18. Van Den Bent, M. J. Interobserver variation of the histopathological diagnosis in clinical trials on glioma: a clinician's perspective. *Acta Neuropathol* **120**, 297–304 (2010).

19. Brennan, C. W. *et al.* The somatic genomic landscape of glioblastoma. *Cell* **155**, 462 (2013).

20. Zhang, P., Xia, Q., Liu, L., Li, S. & Dong, L. Current Opinion on Molecular Characterization for GBM Classification in Guiding Clinical Diagnosis, Prognosis, and Therapy. *Front Mol Biosci* **7**, (2020).

21. Das, A. B. Small-world networks of prognostic genes associated with lung adenocarcinoma development. *Genomics* **112**, 4078–4088 (2020).

22. Sumithra, B., Saxena, U. & Das, A. B. A comprehensive study on genome-wide coexpression network of KHDRBS1/Sam68 reveals its cancer and patient-specific association. *Sci Rep* **9**, 11083 (2019).

23. Ludwig, J. A. & Weinstein, J. N. Biomarkers in cancer staging, prognosis and treatment selection. *Nat Rev Cancer* **5**, 845–856 (2005).

24. Bild, A. H. *et al.* Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature* **439**, 353–357 (2006).

25. Jayanthi, V. S. P. K. S. A., Das, A. B. & Saxena, U. Grade-specific diagnostic and prognostic biomarkers in breast cancer. *Genomics* **112**, 388–396 (2020).

26. Di Carlo, A. *et al.* Epidermal growth factor receptor in human brain tumors. *J Endocrinol Invest* **15**, 31–37 (1992).

27. A, S. & K, F. Platelet-derived growth factor (PDGF) in primary brain tumours of neuroglial origin. *Histol Histopathol* **13**, 511–520 (1998).

28. Machein, M. R. & Plate, K. H. VEGF in brain tumors. *J Neurooncol* **50**, 109–120 (2000).

29. Wu, A., Aldape, K. & Lang, F. F. High rate of deletion of chromosomes 1p and 19q in insular oligodendroglial tumors. *J Neurooncol* **99**, 57–64 (2010).

30. Vastrad, B., Vastrad, C., Godavarthi, A. & Chandrashekar, R. Molecular mechanisms underlying gliomas and glioblastoma pathogenesis revealed by bioinformatics analysis of microarray data. *Med Oncol* **34**, (2017).

31. Zhong, S. *et al.* Identification of Driver Genes and Key Pathways of Glioblastoma Shows JNJ-7706621 as a Novel Antiglioblastoma Drug. *World Neurosurg* **109**, e329–e342 (2018).

32. D S Rickman *et al.* Distinctive molecular profiles of high-grade and low-grade gliomas based on oligonucleotide microarray analysis. *Cancer Res* **61**, 6885–68891 (2001).

33. Ahmadov, U. *et al.* The long non-coding RNA HOTAIRM1 promotes tumor aggressiveness and radiotherapy resistance in glioblastoma. *Cell Death Dis* **12**, (2021).

34. Cai, H. Q. *et al.* Overexpression of MCM6 predicts poor survival in patients with glioma. *Hum Pathol* **78**, 182–187 (2018).

35. Sugur, H. S. *et al.* IRX1 is a novel gene, overexpressed in high-grade IDH-mutant astrocytomas. *Pathol Res Pract* **245**, (2023).

36. Karsy, M., Guan, J. & Eric Huang, L. Prognostic role of mitochondrial pyruvate carrier in isocitrate dehydrogenase-mutant glioma. *J Neurosurg* **130**, 56–66 (2018).

37. Zou, Y. F. *et al.* Screening and authentication of molecular markers in malignant glioblastoma based on gene expression profiles. *Oncol Lett* **18**, 4593–4604 (2019).

38. Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458**, 719–724 (2009).

39. Cohen, A. L., Holmen, S. L. & Colman, H. IDH1 and IDH2 mutations in gliomas. *Curr Neurol Neurosci Rep* **13**, (2013).

40. Balss, J. *et al.* Analysis of the IDH1 codon 132 mutation in brain tumors. *Acta Neuropathol* **116**, 597–602 (2008).

41. Ichimura, K. *et al.* IDH1 mutations are present in the majority of common adult gliomas but rare in primary glioblastomas. *Neuro Oncol* **11**, 341–347 (2009).

42. Watanabe, T., Nobusawa, S., Kleihues, P. & Ohgaki, H. IDH1 mutations are early events in the development of astrocytomas and oligodendrogliomas. *Am J Pathol* **174**, 1149–1153 (2009).

43. Wijnenga, M. M. J. *et al.* Prognostic relevance of mutations and copy number alterations assessed with targeted next generation sequencing in IDH mutant grade II glioma. *J Neurooncol* **139**, 349–357 (2018).

44. Sanson, M. *et al.* Isocitrate dehydrogenase 1 codon 132 mutation is an important prognostic biomarker in gliomas. *J Clin Oncol* **27**, 4150–4154 (2009).

45. Bleeker, F. E. *et al.* IDH1 mutations at residue p.R132 (IDH1(R132)) occur frequently in high-grade gliomas but not in other solid tumors. *Hum Mutat* **30**, 7–11 (2009).

46. Parsons, D. W. *et al.* An integrated genomic analysis of human glioblastoma multiforme. *Science* **321**, 1807–1812 (2008).

47. McLendon, R. *et al.* Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061–1068 (2008).

48. Draaisma, K. *et al.* PI3 kinase mutations and mutational load as poor prognostic markers in diffuse glioma patients. *Acta Neuropathol Commun* **3**, 88 (2015).

49. Sahm, F. *et al.* CIC and FUBP1 mutations in oligodendrogliomas, oligoastrocytomas and astrocytomas. *Acta Neuropathol* **123**, 853–860 (2012).

50. Chatterjee, D., Radotra, B., Kumar, N., Vasishta, R. & Gupta, S. IDH1, ATRX, and BRAF V600E mutation in astrocytic tumors and their significance in patient outcome in north Indian population. *Surg Neurol Int* **9**, (2018).

51. Reis, G. F. *et al.* CDKN2A loss is associated with shortened overall survival in lower-grade (World Health Organization Grades II-III) astrocytomas. *J Neuropathol Exp Neurol* **74**, 442–452 (2015).

52. Huang, L. E. Impact of CDKN2A/B Homozygous Deletion on the Prognosis and Biology of IDH-Mutant Glioma. *Biomedicines* **10**, (2022).

53. Smith, J. S. *et al.* PTEN mutation, EGFR amplification, and outcome in patients with anaplastic astrocytoma and glioblastoma multiforme. *J Natl Cancer Inst* **93**, 1246–1256 (2001).

54. Abdulghani, M. M., Abbas, M. N. & Mohammed, W. R. Immunohistochemical Expression of Epidermal Growth Factor Receptor in Astrocytic Tumors in Iraqi Patients. *Open Access Maced J Med Sci* **7**, 3514–3520 (2019).

55. Karsy, M., Guan, J., Cohen, A. L., Jensen, R. L. & Colman, H. New Molecular Considerations for Glioma: IDH, ATRX, BRAF, TERT, H3 K27M. *Curr Neurol Neurosci Rep* **17**, (2017).

56. Ebrahimi, A. *et al.* ATRX immunostaining predicts IDH and H3F3A status in gliomas. *Acta Neuropathol Commun* **4**, 60 (2016).

57. Lee, Y. *et al.* The frequency and prognostic effect of TERT promoter mutation in diffuse gliomas. *Acta Neuropathol Commun* **5**, 62 (2017).

58. Weller, M. *et al.* Molecular predictors of progression-free and overall survival in patients with newly diagnosed glioblastoma: a prospective translational study of the German Glioma Network. *J Clin Oncol* **27**, 5743–5750 (2009).

59. Krex, D. *et al.* Long-term survival with glioblastoma multiforme. *Brain* **130**, 2596–2606 (2007).

60. Zhang, M., Yang, D. & Gold, B. Origin of mutations in genes associated with human glioblastoma multiform cancer: random polymerase errors versus deamination. *Heliyon* **5**, (2019).

61. Verhaak, R. G. W. *et al.* Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell* **17**, 98–110 (2010).

62. Mischel, P. S. & Cloughesy, T. F. Targeted molecular therapy of GBM. *Brain Pathol* **13**, 52–61 (2003).

63. Sanson, M. *et al.* Chromosome 7p11.2 (EGFR) variation influences glioma risk. *Hum Mol Genet* **20**, 2897–2904 (2011).

64. Shete, S. *et al.* Genome-wide association study identifies five susceptibility loci for glioma. *Nat Genet* **41**, 899–904 (2009).

65. Wrensch, M. *et al.* Variants in the CDKN2B and RTEL1 regions are associated with high-grade glioma susceptibility. *Nat Genet* **41**, 905–908 (2009).

66. Enciso-Mora, V. *et al.* Low penetrance susceptibility to glioma is caused by the TP53 variant rs78378222. *Br J Cancer* **108**, 2178–2185 (2013).

67. Walsh, K. M. *et al.* Variants near TERT and TERC influencing telomere length are associated with high-grade glioma risk. *Nat Genet* **46**, 731–735 (2014).

68. Kinnersley, B. *et al.* Genome-wide association study identifies multiple susceptibility loci for glioma. *Nat Commun* **6**, (2015).

69. Mallik, S., Qin, G., Jia, P. & Zhao, Z. Molecular signatures identified by integrating gene expression and methylation in non-seminoma and seminoma of testicular germ cell tumours. *Epigenetics* **16**, 1–15 (2021).

70. Mellai, M. *et al.* MGMT promoter hypermethylation and its associations with genetic alterations in a series of 350 brain tumors. *J Neurooncol* **107**, 617–631 (2012).

71. Wang, W. *et al.* A three-gene signature for prognosis in patients with MGMT promoter-methylated glioblastoma. *Oncotarget* **7**, 69991–69999 (2016).

72. Kanwal, R. & Gupta, S. Epigenetic modifications in cancer. *Clin Genet* **81**, 303–311 (2012).

73. Zacher, A. *et al.* Molecular Diagnostics of Gliomas Using Next Generation Sequencing of a Glioma-Tailored Gene Panel. *Brain Pathol* **27**, 146–159 (2017).

74. Afifi, S., Gholamhosseini, H. & Sinha, R. SVM classifier on chip for melanoma detection. *Annu Int Conf IEEE Eng Med Biol Soc* **2017**, 270–274 (2017).

75. Liu, Z., Bensmail, H. & Tan, M. Efficient feature selection and multiclass classification with integrated instance and model based learning. *Evol Bioinform Online* **8**, 197–205 (2012).

76. Yao, Z. & Ruzzo, W. L. A regression-based K nearest neighbor algorithm for gene function prediction from heterogeneous data. *BMC Bioinformatics* **7 Suppl 1**, (2006).

77. Kaviarasi, R. & Gandhi Raj, R. Accuracy Enhanced Lung Cancer Prognosis for Improving Patient Survivability Using Proposed Gaussian Classifier System. *J Med Syst* **43**, (2019).

78. Qiu, Y. L., Zheng, H. & Gevaert, O. Genomic data imputation with variational auto-encoders. *Gigascience* **9**, (2020).

79. Franco, E. F. *et al.* Performance Comparison of Deep Learning Autoencoders for Cancer Subtype Detection Using Multi-Omics Data. *Cancers (Basel)* **13**, (2021).

80. Way, G. P. & Greene, C. S. Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders. *Pacific Symposium on Biocomputing* **0**, 80–95 (2018).

81. Bukhari, M. M. *et al.* An Improved Artificial Neural Network Model for Effective Diabetes Prediction. *Complexity* **2021**, (2021).

82. Basha, S. H. S., Dubey, S. R., Pulabaigari, V. & Mukherjee, S. Impact of fully connected layers on performance of convolutional neural networks for image classification. *Neurocomputing* **378**, 112–119 (2020).

83. García-Díaz, P., Sánchez-Berriel, I., Martínez-Rojas, J. A. & Diez-Pascual, A. M. Unsupervised feature selection algorithm for multiclass cancer classification of gene expression RNA-Seq data. *Genomics* **112**, 1916–1925 (2020).

84. Awada, H. *et al.* Machine learning integrates genomic signatures for subclassification beyond primary and secondary acute myeloid leukemia. *Blood* **138**, 1885–1895 (2021).

85. Aruna, S. A Novel SVM based CSSFFS Feature Selection Algorithm for Detecting Breast Cancer. *Int J Comput Appl* **31**, 975–8887 (2011).

86. Golub, T. R. *et al.* Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531–527 (1999).

87. Kori, M. & Gov, E. Bioinformatics Prediction and Machine Learning on Gene Expression Data Identifies Novel Gene Candidates in Gastric Cancer. *Genes (Basel)* **13**, (2022).

88. Moler, E. J., Chow, M. L. & Mian, I. S. Analysis of molecular profile data using generative and discriminative methods. *Physiol Genomics* **4**, 109–126 (2000).

89. Liu, Y. Active learning with support vector machine applied to gene expression data for cancer classification. *J Chem Inf Comput Sci* **44**, 1936–1941 (2004).

90. Ayyad, S. M., Saleh, A. I. & Labib, L. M. Gene expression cancer classification using modified K-Nearest Neighbors technique. *Biosystems* **176**, 41–51 (2019).

91. C, L. *et al.* Novel Biomarker Prediction for Lung Cancer Using Random Forest Classifiers. *Cancer Inform* **22**, (2023).

92. Li, M. X. *et al.* Using a machine learning approach to identify key prognostic molecules for esophageal squamous cell carcinoma. *BMC Cancer* **21**, (2021).

93. Su, Y. *et al.* Colon cancer diagnosis and staging classification based on machine learning and bioinformatics analysis. *Comput Biol Med* **145**, (2022).

94. Maniruzzaman, M. *et al.* Statistical characterization and classification of colon microarray gene expression data using multiple machine learning paradigms. *Comput Methods Programs Biomed* **176**, 173–193 (2019).

95. Wu, Q., Ma, Z., Fan, J., Xu, G. & Shen, Y. A Feature Selection Method Based on Hybrid Improved Binary Quantum Particle Swarm Optimization. *IEEE Access* **7**, 80588–80601 (2019).

96. Salem, H., Attiya, G. & El-Fishawy, N. Classification of human cancer diseases by gene expression profiles. *Appl Soft Comput* **50**, 124–134 (2017).

97. Yuan, B., Yang, D., Rothberg, B. E. G., Chang, H. & Xu, T. Unsupervised and supervised learning with neural network for human transcriptome analysis and cancer diagnosis. *Sci Rep* **10**, (2020).

98. Shah, S. H., Iqbal, M. J., Ahmad, I., Khan, S. & Rodrigues, J. J. P. C. Optimized gene selection and classification of cancer from microarray gene expression data using deep learning. *Neural Comput Appl* 1–12 (2020) doi:10.1007/S00521-020-05367-8/TABLES/4.

99. Mohammed, M., Mwambi, H., Mboya, I. B., Elbashir, M. K. & Omolo, B. A stacking ensemble deep learning approach to cancer type classification based on TCGA data. *Sci Rep* **11**, (2021).

100. Rezaee, K., Jeon, G., Khosravi, M. R., Attar, H. H. & Sabzevari, A. Deep learning-based microarray cancer classification and ensemble gene selection approach. *IET Syst Biol* **16**, 120–131 (2022).

101. Almarzouki, H. Z. Deep-Learning-Based Cancer Profiles Classification Using Gene Expression Data Profile. *J Healthc Eng* **2022**, (2022).

102. Tasaki, S., Gaiteri, C., Mostafavi, S. & Wang, Y. Deep learning decodes the principles of differential gene expression. *Nat Mach Intell* **2**, 376–386 (2020).

103. Sekhon, A., Singh, R. & Qi, Y. DeepDiff: DEEP-learning for predicting DIFFerential gene expression from histone modifications. *Bioinformatics* **34**, i891–i900 (2018).

104. Singh, R., Lanchantin, J., Robins, G. & Qi, Y. DeepChrome: deep-learning for predicting gene expression from histone modifications. *Bioinformatics* **32**, i639–i648 (2016).

105. Suo, Y., Liu, T., Jia, X. & Yu, F. Application of Clustering Analysis in Brain Gene Data Based on Deep Learning. *IEEE Access* **7**, 2947–2956 (2019).

106. Eraslan, G., Avsec, Ž., Gagneur, J. & Theis, F. J. Deep learning: new computational modelling techniques for genomics. *Nat Rev Genet* **20**, 389–403 (2019).

107. Topol, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med* **25**, 44–56 (2019).

108. Tokheim, C. J., Papadopoulos, N., Kinzler, K. W., Vogelstein, B. & Karchin, R. Evaluating the evaluation of cancer driver genes. *Proc Natl Acad Sci U S A* **113**, 14330–14335 (2016).

109. Bailey, M. H. *et al.* Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **173**, 371-385.e18 (2018).

110. Cheng, F. *et al.* A genome-wide positioning systems network algorithm for in silico drug repurposing. *Nat Commun* **10**, (2019).

111. Azuaje, F. Artificial intelligence for precision oncology: beyond patient stratification. *NPJ Precis Oncol* **3**, (2019).

112. Vural, S., Wang, X. & Guda, C. Classification of breast cancer patients using somatic mutation profiles and machine learning approaches. *BMC Syst Biol* **10 Suppl 3**, (2016).

113. Amrane, M., Oukid, S., Gagaoua, I. & Ensari, T. Breast cancer classification using machine learning. *2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting, EBBT 2018* 1–4 (2018) doi:10.1109/EBBT.2018.8391453.

114. Li, Y. & Luo, Y. Performance-weighted-voting model: An ensemble machine learning method for cancer type classification using whole-exome sequencing mutation. *Quant Biol* **8**, 347–358 (2020).

115. Chen, Y., Sun, J., Huang, L. C., Xu, H. & Zhao, Z. Classification of Cancer Primary Sites Using Machine Learning and Somatic Mutations. *Biomed Res Int* **2015**, (2015).

116. Pandey, M., Anoosha, P., Yesudhas, D. & Gromiha, M. M. Identification of potential driver mutations in glioblastoma using machine learning. *Brief Bioinform* **23**, (2022).

117. Palazzo, M., Beauseroy, P. & Yankilevich, P. A pan-cancer somatic mutation embedding using autoencoders. *BMC Bioinformatics* **20**, (2019).

118. Maruf, F. A., Pratama, R. & Song, G. DNN-Boost: Somatic mutation identification of tumor-only whole-exome sequencing data using deep neural network and XGBoost. *J Bioinform Comput Biol* **19**, (2021).

119. Yuan, Y. *et al.* DeepGene: an advanced cancer type classifier based on deep learning and somatic point mutations. *BMC Bioinformatics* **17**, (2016).

120. Zeng, Z. *et al.* Deep learning for cancer type classification and driver gene identification. *BMC Bioinformatics* **22**, (2021).

121. Jin, B., Li, Y. & Robertson, K. D. DNA methylation: superior or subordinate in the epigenetic hierarchy? *Genes Cancer* **2**, 607–617 (2011).

122. Jurmeister, P. *et al.* DNA methylation-based machine learning classification distinguishes pleural mesothelioma from chronic pleuritis, pleural carcinosis, and pleomorphic lung carcinomas. *Lung Cancer* **170**, 105–113 (2022).

123. Tao, M. *et al.* Classifying Breast Cancer Subtypes Using Multiple Kernel Learning Based on Omics Data. *Genes (Basel)* **10**, (2019).

124. Leitheiser, M. *et al.* Machine learning models predict the primary sites of head and neck squamous cell carcinoma metastases based on DNA methylation. *J Pathol* **256**, 378–387 (2022).

125. Ren, J. *et al.* Identification of Methylation Signatures and Rules for Sarcoma Subtypes by Machine Learning Methods. *Biomed Res Int* **2022**, (2022).

126. Cai, Z. *et al.* Classification of lung cancer using ensemble-based feature selection and machine learning methods. *Mol Biosyst* **11**, 791–800 (2015).

127. Ma, B. *et al.* Diagnostic classification of cancers using DNA methylation of paracancerous tissues. *Sci Rep* **12**, (2022).

128. Bedon, L. *et al.* A Novel Epigenetic Machine Learning Model to Define Risk of Progression for Hepatocellular Carcinoma Patients. *Int J Mol Sci* **22**, 1–25 (2021).

129. Eissa, N. S., Khairuddin, U. & Yusof, R. A hybrid metaheuristic-deep learning technique for the pan-classification of cancer based on DNA methylation. *BMC Bioinformatics* **23**, (2022).

130. Levy, J. J. *et al.* MethylNet: an automated and modular deep learning approach for DNA methylation analysis. doi:10.1186/s12859-020-3443-8.

131. Tian, Q. *et al.* MRCNN: a deep learning model for regression of genome-wide DNA methylation. 14–16 (2019) doi:10.1186/s12864-019-5488-5.

132. Zheng, C. & Xu, R. Predicting cancer origins with a DNA methylation-based deep neural network model. (2020) doi:10.1371/journal.pone.0226461.

133. Khwaja, M., Kalofonou, M. & Toumazou, C. A Deep Autoencoder System for Differentiation of Cancer Types Based on DNA Methylation State. (2018).

134. Angermueller, C., Lee, H. J., Reik, W. & Stegle, O. DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. *Genome Biol* **18**, (2017).

135. Danielsson, A. *et al.* MethPed: a DNA methylation classifier tool for the identification of pediatric brain tumor subtypes. (2011) doi:10.1186/s13148-015-0103-3.

136. Choi, J. M., Park, C. & Chae, H. meth-SemiCancer: a cancer subtype classification framework via semi-supervised learning utilizing DNA methylation profiles. *BMC Bioinformatics* **24**, (2023).

137. Yuan, F., Lu, L. & Zou, Q. Analysis of gene expression profiles of lung cancer subtypes with machine learning algorithms. *Biochim Biophys Acta Mol Basis Dis* **1866**, (2020).

138. Ramos, B. *et al.* An Interpretable Approach for Lung Cancer Prediction and Subtype Classification using Gene Expression. *Annu Int Conf IEEE Eng Med Biol Soc* **2021**, 1707–1710 (2021).

139. Tao, M. *et al.* Classifying Breast Cancer Subtypes Using Multiple Kernel Learning Based on Omics Data. *Genes (Basel)* **10**, (2019).

140. Shen, J. *et al.* Deep learning approach for cancer subtype classification using high-dimensional gene expression data. *BMC Bioinformatics* **23**, (2022).

141. Xiao, Y., Bi, M., Guo, H. & Li, M. Multi-omics approaches for biomarker discovery in early ovarian cancer diagnosis. *EBioMedicine* **79**, (2022).

142. Yang, C., Wang, Y. T. & Zheng, C. H. A Random Walk Based Cluster Ensemble Approach for Data Integration and Cancer Subtyping. *Genes (Basel)* **10**, (2019).

143. Mohaiminul Islam, M. *et al.* An integrative deep learning framework for classifying molecular subtypes of breast cancer. *Comput Struct Biotechnol J* **18**, 2185–2199 (2020).

144. Vale-Silva, L. A. & Rohr, K. Long-term cancer survival prediction using multimodal deep learning. *Sci Rep* **11**, (2021).

145. Zhang, X., Xing, Y., Sun, K. & Guo, Y. OmiEmbed: A Unified Multi-Task Deep Learning Framework for Multi-Omics Data. *Cancers (Basel)* **13**, (2021).

146. Chaudhary, K., Poirion, O. B., Lu, L. & Garmire, L. X. Deep Learning-Based Multi-Omics Integration Robustly Predicts Survival in Liver Cancer. *Clin Cancer Res* **24**, 1248–1259 (2018).

147. Xu, J. *et al.* A hierarchical integration deep flexible neural forest framework for cancer subtype classification by integrating multi-omics data. *BMC Bioinformatics* **20**, (2019).

148. Dincer, A. B., Celik, S., Hiranuma, N. & Lee, S.-I. DeepProfile: Deep learning of cancer molecular profiles for precision medicine. *bioRxiv* 278739 (2018) doi:10.1101/278739.

149. Jia, P. *et al.* Deep generative neural network for accurate drug response imputation. *Nat Commun* **12**, (2021).

150. van Hilten, A. *et al.* GenNet framework: interpretable deep learning for predicting phenotypes from genetic data. *Commun Biol* **4**, (2021).

151. Nguyen, N. D., Jin, T. & Wang, D. Varmole: a biologically drop-connect deep neural network model for prioritizing disease risk variants and genes. *Bioinformatics* **37**, 1772–1775 (2021).

152. Argelaguet, R. *et al.* Multi-Omics Factor Analysis-a framework for unsupervised integration of multi-omics data sets. *Mol Syst Biol* **14**, (2018).

153. Chiu, Y. C. *et al.* Predicting drug response of tumors from integrated genomic profiles by deep neural networks. *BMC Med Genomics* **12**, (2019).

154. Wang, C., Lye, X., Kaalia, R., Kumar, P. & Rajapakse, J. C. Deep learning and multi-omics approach to predict drug responses in cancer. *BMC Bioinformatics* **22**, (2022).

155. Almutiri, T., Alomar, K. & Alganmi, N. Predicting Drug Response on Multi-Omics Data Using a Hybrid of Bayesian Ridge Regression with Deep Forest. *IJACSA) International Journal of Advanced Computer Science and Applications* **14**, (2023).

156. Malik, V., Kalakoti, Y. & Sundar, D. Deep learning assisted multi-omics integration for survival and drug-response prediction in breast cancer. *BMC Genomics* **22**, (2021).

157. Sharifi-Noghabi, H., Zolotareva, O., Collins, C. C. & Ester, M. MOLI: multi-omics late integration with deep neural networks for drug response prediction. *Bioinformatics* **35**, i501–i509 (2019).

158. Feng, R. *et al.* AGMI: Attention-Guided Multi-omics Integration for Drug Response Prediction with Graph Neural Networks. *Proceedings - 2021 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2021* 1295–1298 (2021) doi:10.1109/BIBM52615.2021.9669314.

159. Hiort, P. *et al.* DrDimont: explainable drug response prediction from differential analysis of multi-omics networks. *Bioinformatics* **38**, II113–II119 (2022).

160. Eckel-Passow, J. E. *et al.* Glioma Groups Based on 1p/19q, IDH, and TERT Promoter Mutations in Tumors. *N Engl J Med* **372**, 2499–2508 (2015).

161. Hartmann, C. *et al.* Type and frequency of IDH1 and IDH2 mutations are related to astrocytic and oligodendroglial differentiation and age: a study of 1,010 diffuse gliomas. *Acta Neuropathol* **118**, 469–474 (2009).

162. Ball, M. K. *et al.* Frequency of false-positive FISH 1p/19q codeletion in adult diffuse astrocytic gliomas. *Neurooncol Adv* **2**, (2020).

163. Forst, D. A., Nahed, B. V., Loeffler, J. S. & Batchelor, T. T. Low-grade gliomas. *Oncologist* **19**, 403–413 (2014).

164. Claus, E. B. *et al.* Survival and low-grade glioma: the emergence of genetic information. *Neurosurg Focus* **38**, (2015).

165. Goldman, M. J. *et al.* Visualizing and interpreting cancer genomics data via the Xena platform. *Nat Biotechnol* **38**, 675–678 (2020).

166. Wold, S., Esbensen, K. & Geladi, P. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems* **2**, 37–52 (1987).

167. Sanz, H., Valim, C., Vegas, E., Oller, J. M. & Reverter, F. SVM-RFE: selection and visualization of the most relevant features through non-linear kernels. *BMC Bioinformatics* **19**, (2018).

168. Pirooznia, M., Yang, J. Y., Qu, M. Q. & Deng, Y. A comparative study of different machine learning methods on microarray gene expression data. *BMC Genomics* **9 Suppl 1**, (2008).

169. Li, Z., Xie, W. & Liu, T. Efficient feature selection and classification for microarray data. *PLoS One* **13**, (2018).

170. Degenhardt, F., Seifert, S. & Szymczak, S. Evaluation of variable selection methods for random forests and omics data sets. *Brief Bioinform* **20**, 492–503 (2019).

171. Burges, C. J. C. A tutorial on support vector machines for pattern recognition. *Data Min Knowl Discov* **2**, 121–167 (1998).

172. Heckerman, D., Geiger, D. & Chickering, D. M. Learning Bayesian Networks: The Combination of Knowledge and Statistical Data. *Mach Learn* **20**, 197–243 (1995).

173. Kaviarasi, R. & Gandhi Raj, R. Accuracy Enhanced Lung Cancer Prognosis for Improving Patient Survivability Using Proposed Gaussian Classifier System. *J Med Syst* **43**, (2019).

174. Song, Y. Y. & Lu, Y. Decision tree methods: applications for classification and prediction. *Shanghai Arch Psychiatry* **27**, 130–135 (2015).

175. Breiman, L. Random forests. *Mach Learn* **45**, 5–32 (2001).

176. Metz, C. E. Basic principles of ROC analysis. *Semin Nucl Med* **8**, 283–298 (1978).

177. Triantaphyllou, E. Multi-Criteria Decision Making Methods. 5–21 (2000) doi:10.1007/978-1-4757-3157-6_2.

178. Zhou, Y. *et al.* Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat Commun* **10**, (2019).

179. Butte, A. The use and analysis of microarray data. *Nat Rev Drug Discov* **1**, 951–960 (2002).

180. Lenz, M., Muller, F. J., Zenke, M. & Schuppert, A. Principal components analysis and the reported low intrinsic dimensionality of gene expression microarray data. *Sci Rep* **6**, (2016).

181. Al-Rajab, M., Lu, J. & Xu, Q. Examining applying high performance genetic data feature selection and classification algorithms for colon cancer diagnosis. *Comput Methods Programs Biomed* **146**, 11–24 (2017).

182. Lever, J., Krzywinski, M. & Altman, N. Model selection and overfitting. *Nature Methods 2016 13:9* (2016).

183. Cawley, G. C. & Talbot, N. L. C. On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. *Journal of Machine Learning Research* **11**, 2079–2107 (2010).

184. Si, T., Miranda, P., Galdino, J. V. & Nascimento, A. Grammar-based automatic programming for medical data classification: an experimental study. *Artif Intell Rev* **54**, 4097–4135 (2021).

185. Wei, Q. & Dunbrack, R. L. The Role of Balanced Training and Testing Data Sets for Binary Classifiers in Bioinformatics. *PLoS One* **8**, e67863 (2013).

186. Wang, X. *et al.* Protein-protein interaction sites prediction by ensemble random forests with synthetic minority oversampling technique. *Bioinformatics* **35**, 2395–2402 (2019).

187. Yao, F., Zhang, C., Du, W., Liu, C. & Xu, Y. Identification of Gene-Expression Signatures and Protein Markers for Breast Cancer Grading and Staging. *PLoS One* **10**, (2015).

188. Zamecnik, J. The extracellular space and matrix of gliomas. *Acta Neuropathol* **110**, 435–442 (2005).

189. Wang, Q. W., Lin, W. W. & Zhu, Y. J. Comprehensive analysis of a TNF family based-signature in diffuse gliomas with regard to prognosis and immune significance. *Cell Commun Signal* **20**, (2022).

190. Colardo, M., Segatto, M. & Di Bartolomeo, S. Targeting RTK-PI3K-mTOR Axis in Gliomas: An Update. *Int J Mol Sci* **22**, (2021).

191. Jiang, Q. *et al.* Glioma malignancy is linked to interdependent and inverse AMOG and L1 adhesion molecule expression. *BMC Cancer* **19**, (2019).

192. Maklad, A., Sharma, A. & Azimi, I. Calcium Signaling in Brain Cancers: Roles and Therapeutic Targeting. *Cancers (Basel)* **11**, (2019).

193. Venkatesh, H. S. *et al.* Electrical and synaptic integration of glioma into neural circuits. *Nature* **573**, 539–545 (2019).

194. Atkinson, G. P., Nozell, S. E. & Benveniste, E. N. NF-kappaB and STAT3 signaling in glioma: targets for future therapies. *Expert Rev Neurother* **10**, 575–586 (2010).

195. Shangguan, W., Lv, X. & Tian, N. FoxD2-AS1 is a prognostic factor in glioma and promotes temozolomide resistance in a O 6-methylguanine-DNA methyltransferase-dependent manner. *Korean J Physiol Pharmacol* **23**, 475–482 (2019).

196. Zalenski, A., De, K. & Venere, M. Not just another biomarker: the role of integrin alpha 7 in glioblastoma. *Stem Cell Investig* **4**, (2017).

197. Ding, X. *et al.* Eps8 promotes cellular growth of human malignant gliomas. *Oncol Rep* **29**, 697–703 (2013).

198. Rammal, H. *et al.* Discoidin Domain Receptors: Potential Actors and Targets in Cancer. *Front Pharmacol* **7**, (2016).

199. Wastowski, I. J. *et al.* Human leukocyte antigen-G is frequently expressed in glioblastoma and may be induced in vitro by combined 5-aza-2'-deoxycytidine and interferon-γ treatments: results from a multicentric study. *Am J Pathol* **182**, 540–552 (2013).

200. Wiendl, H. *et al.* A functional role of HLA-G expression in human gliomas: an alternative strategy of immune escape. *J Immunol* **168**, 4772–4780 (2002).

201. Jesionek-Kupnicka, D. *et al.* TP53 promoter methylation in primary glioblastoma: relationship with TP53 mRNA and protein expression and mutation status. *DNA Cell Biol* **33**, 217–226 (2014).

202. Lee, Y. J. *et al.* Gene expression profiling of glioblastoma cell lines depending on TP53 status after tumor-treating fields (TTFields) treatment. *Sci Rep* **10**, (2020).

203. Liu, K. W., Hu, B. & Cheng, S. Y. Platelet-derived growth factor receptor alpha in glioma: a bad seed. *Chin J Cancer* **30**, 590–602 (2011).

204. Peng, G. *et al.* The HIF1α-PDGFD-PDGFRα axis controls glioblastoma growth at normoxia/mild-hypoxia and confers sensitivity to targeted therapy by echinomycin. *J Exp Clin Cancer Res* **40**, (2021).

205. Auvergne, R. M. *et al.* Transcriptional differences between normal and glioma-derived glial progenitor cells identify a core set of dysregulated genes. *Cell Rep* **3**, 2127–2141 (2013).

206. Weng, J. *et al.* PCDHGA9 acts as a tumor suppressor to induce tumor cell apoptosis and autophagy and inhibit the EMT process in human gastric cancer. *Cell Death Dis* **9**, (2018).

207. Bayin, N. S. *et al.* GPR133 (ADGRD1), an adhesion G-protein-coupled receptor, is necessary for glioblastoma growth. *Oncogenesis* **5**, (2016).

208. Wang, K. *et al.* Hedgehog/Gli1 signaling pathway regulates MGMT expression and chemoresistance to temozolomide in human glioblastoma. *Cancer Cell Int* **17**, (2017).

209. Dou, Y., Xu, H., Wu, X. & Liu, P. Tac2-N Promotes Glioma Proliferation and Indicates Poor Clinical Outcomes. *Tohoku J Exp Med* **255**, 247–256 (2021).

210. Azoitei, N. *et al.* Protein kinase D2 is a novel regulator of glioblastoma growth and tumor formation. *Neuro Oncol* **13**, 710–724 (2011).

211. Tritschler, I. *et al.* Modulation of TGF-beta activity by latent TGF-beta-binding protein 1 in human malignant glioma cells. *Int J Cancer* **125**, 530–540 (2009).

212. Yamaguchi, N. Multiple Roles of Vestigial-Like Family Members in Tumor Development. *Front Oncol* **10**, (2020).

213. Park, A. K., Kim, P., Ballester, L. Y., Esquenazi, Y. & Zhao, Z. Subtype-specific signaling pathways and genomic aberrations associated with prognosis of glioblastoma. *Neuro Oncol* **21**, 59–70 (2019).

214. Mardis, E. R. & Wilson, R. K. Cancer genome sequencing: a review. *Hum Mol Genet* **18**, (2009).

215. Campbell, P. J. *et al.* Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).

216. Bozdag, S. *et al.* Age-specific signatures of glioblastoma at the genomic, genetic, and epigenetic levels. *PLoS One* **8**, (2013).

217. Dong, Z. & Cui, H. Epigenetic modulation of metabolism in glioblastoma. *Semin Cancer Biol* **57**, 45–51 (2019).

218. Vinel, C. *et al.* Comparative epigenetic analysis of tumour initiating cells and syngeneic EPSC-derived neural stem cells in glioblastoma. *Nat Commun* **12**, (2021).

219. Moore, L. D., Le, T. & Fan, G. DNA methylation and its basic function. *Neuropsychopharmacology* **38**, 23–38 (2013).

220. Qin, G. *et al.* MicroRNA and transcription factor co-regulatory networks and subtype classification of seminoma and non-seminoma in testicular germ cell tumors. *Sci Rep* **10**, (2020).

221. Bandyopadhyay, S., Mallik, S. & Mukhopadhyay, A. A Survey and Comparative Study of Statistical Tests for Identifying Differential Expression from Microarray Data. *IEEE/ACM Trans Comput Biol Bioinform* **11**, 95–115 (2014).

222. Maegawa, S. *et al.* Widespread and tissue specific age-related DNA methylation changes in mice. *Genome Res* **20**, 332–340 (2010).

223. Muthukrishnan, R. & Rohini, R. LASSO: A feature selection technique in predictive modeling for machine learning. *2016 IEEE International Conference on Advances in Computer Applications, ICACA 2016* 18–20 (2017) doi:10.1109/ICACA.2016.7887916.

224. Hosmer, D. W., Lemeshow, S. & Sturdivant, R. X. Applied Logistic Regression: Third Edition. *Applied Logistic Regression: Third Edition* 1–510 (2013) doi:10.1002/9781118548387.

225. López-García, G., Jerez, J. M., Franco, L. & Veredas, F. J. Transfer learning with convolutional neural networks for cancer survival prediction using gene-expression data. *PLoS One* **15**, (2020).

226. Mostavi, M., Chiu, Y. C., Huang, Y. & Chen, Y. Convolutional neural network models for cancer type prediction based on gene expression. *BMC Med Genomics* **13**, (2020).

227. Chatterjee, S., Iyer, A., Avva, S., Kollara, A. & Sankarasubbu, M. Convolutional Neural Networks In Classifying Cancer Through DNA Methylation. (2018).

228. Xia, C., Xiao, Y., Wu, J., Zhao, X. & Li, H. A convolutional neural network based ensemble method for cancer prediction using DNA methylation data. *ACM International Conference Proceeding Series* **Part F148150**, 191–196 (2019).

229. Mallik, S. & Zhao, Z. Graph- and rule-based learning algorithms: a comprehensive review of their applications for cancer type classification and prognosis using genomic data. *Brief Bioinform* **21**, 221–247 (2020).

230. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, (2008).

231. Kuleshov, M. V. *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res* **44**, W90–W97 (2016).

232. Tang, Z. *et al.* GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. *Nucleic Acids Res* **45**, W98–W102 (2017).

233. Locke, W. J. *et al.* DNA Methylation Cancer Biomarkers: Translation to the Clinic. *Front Genet* **10**, (2019).

234. Sandoval, J. *et al.* Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. *Epigenetics* **6**, 692–702 (2011).

235. Yang, X. *et al.* Gene body methylation can alter gene expression and is a therapeutic target in cancer. *Cancer Cell* **26**, 577–590 (2014).

236. Dhar, G. A., Saha, S., Mitra, P. & Nag Chaudhuri, R. DNA methylation and regulation of gene expression: Guardian of our health. *Nucleus (Calcutta)* **64**, 259–270 (2021).

237. Díaz-Uriarte, R. & Alvarez de Andrés, S. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics* **7**, (2006).

238. Wang, X. *et al.* Comprehensive analysis of gene expression and DNA methylation data identifies potential biomarkers and functional epigenetic modules for lung adenocarcinoma. *Genet Mol Biol* **43**, (2020).

239. Basu, A. & Tiwari, V. K. Epigenetic reprogramming of cell identity: lessons from development for regenerative medicine. *Clin Epigenetics* **13**, (2021).

240. Bufalieri, F., Basili, I., Di Marcotullio, L. & Infante, P. Harnessing the Activation of RIG-I Like Receptors to Inhibit Glioblastoma Tumorigenesis. *Front Mol Neurosci* **14**, (2021).

241. Charalambous, C., Chen, T. C. & Hofman, F. M. Characteristics of tumor-associated endothelial cells derived from glioblastoma multiforme. *Neurosurg Focus* **20**, (2006).

242. Shi, S. *et al.* Syndecan-1 knockdown inhibits glioma cell proliferation and invasion by deregulating a c-src/FAK-associated signaling pathway. *Oncotarget* **8**, 40922–40934 (2017).

243. He, H. *et al.* The roles of GTPase-activating proteins in regulated cell death and tumor immunity. *J Hematol Oncol* **14**, (2021).

244. Mao, H., Lebrun, D. G., Yang, J., Zhu, V. F. & Li, M. Deregulated signaling pathways in glioblastoma multiforme: molecular mechanisms and therapeutic targets. *Cancer Invest* **30**, 48–56 (2012).

245. C, L., S, P., J, H. & H, K. Functional GABA(A) receptors on human glioma cells. *Eur J Neurosci* **10**, 231–238 (1998).

246. Lin, Y. *et al.* Role of Asparagine Endopeptidase in Mediating Wild-Type p53 Inactivation of Glioblastoma. *J Natl Cancer Inst* **112**, 343–355 (2020).

247. Woroniecka, K. I., Rhodin, K. E., Chongsathidkiet, P., Keith, K. A. & Fecci, P. E. T-cell Dysfunction in Glioblastoma: Applying a New Framework. *Clin Cancer Res* **24**, 3792–3802 (2018).

248. Aasen, T., Mesnil, M., Naus, C. C., Lampe, P. D. & Laird, D. W. Gap junctions and cancer: communicating for 50 years. *Nat Rev Cancer* **16**, 775–788 (2016).

249. Sun, J. *et al.* The survival analysis and oncogenic effects of CFP1 and 14-3-3 expression on gastric cancer. *Cancer Cell Int* **19**, (2019).

250. Liu, Z., Ru, L. & Ma, Z. Low Expression of ADCY4 Predicts Worse Survival of Lung Squamous Cell Carcinoma Based on Integrated Analysis and Immunohistochemical Verification. *Front Oncol* **11**, (2021).

251. Berezovsky, A. D. *et al.* Sox2 promotes malignancy in glioblastoma by regulating plasticity and astrocytic differentiation. *Neoplasia* **16**, 193-206.e25 (2014).

252. Karnati, H. *et al.* Down regulated expression of Claudin-1 and Claudin-5 and up regulation of β-catenin: association with human glioma progression. *CNS Neurol Disord Drug Targets* **13**, 1413–1426 (2014).

253. Pangeni, R. P. *et al.* The GALNT9, BNC1 and CCDC8 genes are frequently epigenetically dysregulated in breast tumours that metastasise to the brain. *Clin Epigenetics* **7**, (2015).

254. Yeo, K. S. *et al.* JMJD8 is a positive regulator of TNF-induced NF-κB signaling. *Sci Rep* **6**, (2016).

255. Huang, K. *et al.* The role of PTRF/Cavin1 as a biomarker in both glioma and serum exosomes. *Theranostics* **8**, 1540–1557 (2018).

256. Zhang, Y. *et al.* Downregulation of miR-485-3p promotes glioblastoma cell proliferation and migration via targeting RNF135. *Exp Ther Med* **18**, (2019).

257. Thaker, N. G. *et al.* Identification of survival genes in human glioblastoma cells by small interfering RNA screening. *Mol Pharmacol* **76**, 1246–1255 (2009).

258. Chen, X. *et al.* Protein Palmitoylation Regulates Cell Survival by Modulating XBP1 Activity in Glioblastoma Multiforme. *Mol Ther Oncolytics* **17**, 518–530 (2020).

259. Chen, Z., Gulzar, Z. G., St. Hill, C. A., Walcheck, B. & Brooks, J. D. Increased expression of GCNT1 is associated with altered O-glycosylation of PSA, PAP, and MUC1 in human prostate cancers. *Prostate* **74**, 1059–1067 (2014).

260. Toton, E. *et al.* Impact of PKCε downregulation on autophagy in glioblastoma cells. *BMC Cancer* **18**, (2018).

261. Bianchetti, E., Bates, S. J., Nguyen, T. T. T., Siegelin, M. D. & Roth, K. A. RAB38 Facilitates Energy Metabolism and Counteracts Cell Death in Glioblastoma Cells. *Cells* **10**, (2021).

262. Giambra, M. *et al.* Characterizing the Genomic Profile in High-Grade Gliomas: From Tumor Core to Peritumoral Brain Zone, Passing through Glioma-Derived Tumorspheres. *Biology (Basel)* **10**, (2021).

263. Katsushima, K. *et al.* The long noncoding RNA lnc-HLX-2-7 is oncogenic in Group 3 medulloblastomas. *Neuro Oncol* **23**, 572–585 (2021).

264. Humbert-Claude, M. *et al.* Tollip, an early regulator of the acute inflammatory response in the substantia nigra. *J Neuroinflammation* **13**, (2016).

265. Little, A. C. *et al.* DUOX1 silencing in lung cancer promotes EMT, cancer stem cell characteristics and invasive properties. *Oncogenesis* **5**, (2016).

266. Crisman, T. J. *et al.* Identification of an Efficient Gene Expression Panel for Glioblastoma Classification. *PLoS One* **11**, (2016).

267. Le, N. Q. K. *et al.* Radiomics-based machine learning model for efficiently classifying transcriptome subtypes in glioblastoma patients from MRI. *Comput Biol Med* **132**, (2021).

268. Zhao, L. *et al.* DeepOmix: A scalable and interpretable multi-omics deep learning framework and application in cancer survival analysis. *Comput Struct Biotechnol J* **19**, 2719–2725 (2021).

269. Stel, V. S., Dekker, F. W., Tripepi, G., Zoccali, C. & Jager, K. J. Survival analysis II: Cox regression. *Nephron Clin Pract* **119**, (2011).

270. Abadi, M. *et al.* TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. (2016) doi:10.48550/arxiv.1603.04467.

271. Darwiche, N. Epigenetic mechanisms and the hallmarks of cancer: an intimate affair. *Am J Cancer Res* **10**, 1954 (2020).

272. Azab, M. A. Expression of Anaplastic Lymphoma Kinase (ALK) in glioma and possible clinical correlations. A retrospective institutional study. *Cancer Treat Res Commun* **36**, 100703 (2023).

273. Jiang, Q. *et al.* Glioma malignancy is linked to interdependent and inverse AMOG and L1 adhesion molecule expression. *BMC Cancer* **19**, (2019).

274. Cheng, F. & Guo, D. MET in glioma: signaling pathways and targeted therapies. *J Exp Clin Cancer Res* **38**, (2019).

275. Ellert-Miklaszewska, A., Poleszak, K., Pasierbinska, M. & Kaminska, B. Integrin Signaling in Glioma Pathogenesis: From Biology to Therapy. *Int J Mol Sci* **21**, (2020).

276. Shafi, O. & Siddiqui, G. Tracing the origins of glioblastoma by investigating the role of gliogenic and related neurogenic genes/signaling pathways in GBM development: a systematic review. *World J Surg Oncol* **20**, (2022).

277. Mala, U., Baral, T. K. & Somasundaram, K. Integrative analysis of cell adhesion molecules in glioblastoma identified prostaglandin F2 receptor inhibitor (PTGFRN) as an essential gene. *BMC Cancer* **22**, (2022).

278. Xu, C., Wu, X. & Zhu, J. VEGF promotes proliferation of human glioblastoma multiforme stem-like cells through VEGF receptor 2. *ScientificWorldJournal* **2013**, (2013).

279. Michaelsen, S. R. *et al.* VEGF-C sustains VEGFR2 activation under bevacizumab therapy and promotes glioblastoma maintenance. *Neuro Oncol* **20**, 1462–1474 (2018).

280. Tan, K., Huang, W., Hu, J. & Dong, S. A multi-omics supervised autoencoder for pan-cancer clinical outcome endpoints prediction. *BMC Med Inform Decis Mak* **20**, (2020).

281. Kang, M., Ko, E. & Mersha, T. B. A roadmap for multi-omics data integration using deep learning. *Brief Bioinform* **23**, (2022).

282. Zhang, L. *et al.* Deep Learning-Based Multi-Omics Data Integration Reveals Two Prognostic Subtypes in High-Risk Neuroblastoma. *Front Genet* **9**, (2018).

283. Madhumita & Paul, S. Capturing the latent space of an Autoencoder for multi-omics integration and cancer subtyping. *Comput Biol Med* **148**, (2022).

284. Wu, X. & Fang, Q. Stacked Autoencoder Based Multi-Omics Data Integration for Cancer Survival Prediction. (2022) doi:10.48550/arxiv.2207.04878.

285. Munquad, S., Si, T., Mallik, S., Li, A. & Das, A. B. Subtyping and grading of lower-grade gliomas using integrated feature selection and support vector machine. *Brief Funct Genomics* **21**, 408–421 (2022).

286. Munquad, S., Si, T., Mallik, S., Das, A. B. & Zhao, Z. A Deep Learning-Based Framework for Supporting Clinical Diagnosis of Glioblastoma Subtypes. *Front Genet* **13**, (2022).

287. Dwivedi, A. K. Artificial neural network model for effective cancer classification using microarray gene expression data. *Neural Comput Appl* **29**, 1545–1554 (2018).

288. Yuvaraj, N. & Vivekanandan, P. An efficient SVM based tumor classification with symmetry Non-negative Matrix Factorization using gene expression data. *2013 International Conference on Information Communication and Embedded Systems (ICICES)* 761–768 (2013) doi:10.1109/ICICES.2013.6508193.

289. Nguyen, T., Khosravi, A., Creighton, D. & Nahavandi, S. Hidden Markov models for cancer classification using gene expression profiles. *Inf Sci (N Y)* **316**, 293–307 (2015).

290. Foo, J. *et al.* An Evolutionary Approach for Identifying Driver Mutations in Colorectal Cancer. *PLoS Comput Biol* **11**, (2015).

291. Ostroverkhova, D., Przytycka, T. M. & Panchenko, A. R. Cancer driver mutations: predictions and reality. *Trends Mol Med* (2023) doi:10.1016/J.MOLMED.2023.03.007.

292. Ghiassian, S. D., Menche, J. & Barabási, A. L. A DIseAse MOdule Detection (DIAMOnD) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome. *PLoS Comput Biol* **11**, (2015).

293. Wu, J., Zhang, Q. & Li, G. Identification of cancer-related module in protein-protein interaction network based on gene prioritization. *J Bioinform Comput Biol* **20**, (2022).

294. Forbes, S. A. *et al.* COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res* **43**, D805–D811 (2015).

295. Arnedo-Pac, C., Mularoni, L., Muiños, F., Gonzalez-Perez, A. & Lopez-Bigas, N. OncodriveCLUSTL: a sequence-based clustering method to identify cancer drivers. *Bioinformatics* **35**, 4788–4790 (2019).

296. Basha, O. *et al.* The TissueNet v.2 database: A quantitative view of protein-protein interactions across human tissues. *Nucleic Acids Res* **45**, D427–D431 (2017).

297. Yang, W. *et al.* Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res* **41**, (2013).

298. Wishart, D. S. *et al.* DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res* **46**, D1074–D1082 (2018).

299. Stransky, N. *et al.* Pharmacogenomic agreement between two cancer cell line data sets. *Nature* **528**, 84–87 (2015).

300. Behdenna, A., Haziza, J., Azencott, C.-A. & Nordor, A. pyComBat, a Python tool for batch effects correction in high-throughput molecular data using empirical Bayes methods. *bioRxiv* 2020.03.17.995431 (2021) doi:10.1101/2020.03.17.995431.

301. Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E. & Storey, J. D. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28**, 882–883 (2012).

302. Tegally, H. *et al.* Discovering novel driver mutations from pan-cancer analysis of mutational and gene expression profiles. *PLoS One* **15**, (2020).

303. Barabási, A. L., Gulbahce, N. & Loscalzo, J. Network medicine: a network-based approach to human disease. *Nat Rev Genet* **12**, 56–68 (2011).

304. Delen, E. & Doğanlar, O. The Dose Dependent Effects of Ruxolitinib on the Invasion and Tumorigenesis in Gliomas Cells via Inhibition of Interferon Gamma-Depended JAK/STAT Signaling Pathway. *J Korean Neurosurg Soc* **63**, 444–454 (2020).

305. Chen, R. *et al.* The application of histone deacetylases inhibitors in glioblastoma. *J Exp Clin Cancer Res* **39**, (2020).

306. Galanis, E. *et al.* Phase II trial of vorinostat in recurrent glioblastoma multiforme: a north central cancer treatment group study. *J Clin Oncol* **27**, 2052–2058 (2009).

307. Kipper, F. C. *et al.* Vinblastine and antihelmintic mebendazole potentiate temozolomide in resistant gliomas. *Invest New Drugs* **36**, 323–331 (2018).

308. Vairy, S. *et al.* Phase I study of vinblastine in combination with nilotinib in children, adolescents, and young adults with refractory or recurrent low-grade glioma. *Neurooncol Adv* **2**, (2020).

309. Schaff, L. R. *et al.* Combination Olaparib and Temozolomide for the Treatment of Glioma: A Retrospective Case Series. *Neurology* **99**, 750–755 (2022).

310. Junca, A. *et al.* Crizotinib targets in glioblastoma stem cells. *Cancer Med* **6**, 2625–2634 (2017).

311. Banasavadi-Siddegowda, Y. K. *et al.* Targeting protein arginine methyltransferase 5 sensitizes glioblastoma to trametinib. *Neurooncol Adv* **4**, (2022).

312. Chakravarthi, B. V. S. K., Nepal, S. & Varambally, S. Genomic and Epigenomic Alterations in Cancer. *Am J Pathol* **186**, 1724–1735 (2016).

Publications

1. Sana Munquad, Tapas Si, Saurav Mallik, Asim Bikas Das* and Zhongming Zhao*, 2022. A Deep Learning–Based Framework for Supporting Clinical Diagnosis of Glioblastoma Subtypes. Frontiers in genetics, 13:855420.

2. Sana Munquad, Tapas Si, Saurav Mallik, Aimin Li, Asim Bikas Das*, 2022. "Subtyping and Grading of Lower-grade Gliomas Using Integrated Feature Selection and Support Vector Machine" Briefings in Functional Genomics, 21(5), pp.408-421.

3. Sana Munquad, Asim Bikas Das*, 2023. DeepAutoGlioma: A Deep learning autoencoder-based multi-omics data integration and classification tools for glioma subtyping". BioData Mining, 16(1), p.32.

4. Sana Munquad, Asim Bikas Das*, 2023. Uncovering the subtype-specific disease module and the development of drug response prediction models for glioma. (Under-review).

5. Sana Munquad, Asim Bikas Das*, 2023. The expression pattern of genes encoding secretory proteins exhibits opportunities for improved clinical diagnosis and prognosis prediction in glioblastoma multiforme. (Submitted).

Conferences and Workshops

1. Sana Munquad, Asim Bikas Das (2019), Genome-wide screening to identify driver mutations leading to the development of Astrocytoma to glioblastoma, International conference on "World Congress on Biotechnology-2019, Current Research & Innovations in Biotechnology" held on 28th & 29th August 2019 organized by Indian Institute of Science, Bengaluru, India, 81-82

2. Sana Munquad, Asim Bikas Das (2020), Tree-based classifier to predict brain cancer using whole-genome expression profile, 5th IITM - Tokyo Tech Joint Symposium on "Current trends in Bioinformatics: Big data analysis, Machine Learning and Drug Design" held on 6th - 7th March 2020 at IIT Madras, India, 45-46, Machine Learning, Large scale data analysis.

3. Sana Munquad, Asim Bikas Das (2023), Development of deep-learning based diagnostic tools for glioma subtyping, Great Lakes Bioinformatics Conference

2023, the 15th conference, was hosted by the International Society for Computational Biology (ISCB) held on May 15 - 18, 2023 at Mcgill University, Canada.

4. Attended ‘Analysis of Genome scale Data from Bulk and Single-cell sequencing’ conducted by NIBMG-EMBL-EBI, Kalyani from 19th Nov to 23rd Nov 2018.

5. Attended ‘Research Methodology & Scholarly writing Skills (RMSWS-2019)’ organized by SC-ST Cell at NIT Warangal from 21st Jan to 25th Jan 2019.

6. Participated in “Recent Trends in Computer Simulations for Applications in Biotechnology: Teaching and Learning Strategies” organized by Department of Biotechnology at NIT Warangal from 17th to 21st August 2020.

7. Participated in virtual workshop on “Essential Bioinformatics for Life Science Researchers” organized by SynBiogenica Labs from 24th to 26th August 2020.

8. Participated in “Current Trends in Data Analytics through Hands-on Experience” conducted by Department of Computer Science and Engineering and Department of Biotechnology from 31-08-2020 to 04-09-2020.

9. Participated in the International Workshop on ‘Basic to Advanced Bioinformatics (Linux, Python, R, and NGS Data Analysis)’ organized by Nextgenhelper, New Delhi from September 17-20, 2020.

10. Participated in Industrial training on "Basics to Advanced data analysis for Bioinformatics, Genomics, Proteomics, and NGS datasets" organized by Nextgenhelper, New Delhi from January 23-24 & 30, 2020.

11. Participated in “A 5-day online FDP on “Advances in Biotechnology and Bioinformatics (ABB-2021)” conducted by the Department of Biotechnology from 22-03-2021 to 26-03-2021.

12. Participated in digital training on “Genome Informatics - Second Edition, 2021” conducted by Decode Life from 6th December to 20th December 2021.