Heliyon 10 (2024) e27190

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Uncovering the subtype-specific disease module and the development of drug response prediction models for glioma

Sana Munquad, Asim Bikas Das [*]

Department of Biotechnology, National Institute of Technology Warangal, Warangal, 506004, Telangana, India

ARTICLE INFO

Keywords:
Glioma
Subtype
Disease module
Deep learning
Drug response prediction

ABSTRACT

The poor prognosis of glioma patients brought attention to the need for effective therapeutic approaches for precision therapy. Here, we deployed algorithms relying on network medicine and artificial intelligence to design the framework for subtype-specific target identification and drug response prediction in glioma. We identified the driver mutations that were differentially expressed in each subtype of lower-grade glioma and glioblastoma multiforme and 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, we developed a deep-learning-based drug response prediction framework 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. We observed the consistent performance of the models in training, test, and validation datasets. Drug responses were predicted for particular cell lines derived from distinct subtypes of gliomas. We 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 deep learning-based approach can lead to cancer-specific treatment and improved patient care.

1. Introduction

Gliomas are a category of primary brain tumors that arise from glial cells, 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 [1]. The latest WHO documentation indicates that there are over 100 distinct forms of brain tumors [2]. The most common form of brain tumor is gliomas, which are classified according to how rapidly or slowly the cells divide. Slower-growing gliomas are known as lower-grade gliomas (LGG), whereas more aggressive 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 that has three subtypes: astrocytoma, oligoastrocytoma, and oligodendroglioma. Some of these LGGs turn into GBM, the grade IV

[*] Corresponding author.
E-mail address: asimbikas@nitw.ac.in (A.B. Das).

https://doi.org/10.1016/j.heliyon.2024.e27190
Received 1 August 2023; Received in revised form 24 February 2024; Accepted 26 February 2024
Available online 1 March 2024

tumor, but others stay in this stage for a long time [3,4]. Similarly, there are three subtypes of GBM, i.e., classical, proneural, and mesenchymal [5]. Each subtype has distinct molecular features, and they can be classified using genomics and epigenomics profiles. Despite the presence of intratumor molecular heterogeneity, recent research has shown that deep learning (DL) and machine learning (ML)-based methods may accurately identify glioma subtypes [6,7]. 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 [8]. Hence, to address this issue, we developed 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.

Cancer is developed through an evolutionary process in which healthy cells accumulate several genomic changes, including mutations and gene expression [9,10]. Certain alterations confer a favorable edge to cancer cells, promoting their growth and unregulated proliferation, ultimately resulting in the development of tumors. Advances in sequencing techniques and genome-wide have revealed that accumulated genetic variations associated with an increased risk for cancer are distributed throughout the genome. Additional research demonstrates that genes affected by genomic variations are not randomly distributed in molecular networks. Genes that are linked to the same disease are more prone to interacting with one another. Consequently, a disease module is created, which is a subnetwork connected to a certain disease. Numerous genes that are known to be relevant to disease are found in disease modules [11]. The disease modules, which consist of a known group of genes found in cancers like breast, sarcoma, colorectal, leukemia, and head and neck cancer, were linked to biological processes that are unique to cancer [12]. Wu et al. [13] showed that the active disease modules in breast and cervical cancer are associated with many cancer-related pathways. These studies indicate that the identification of cancer-specific disease modules can help identify novel biomarkers for therapeutic targets. Therefore, network medicine and rational drug-designing approaches recognize these modules as pharmacological targets as opposed to individual genes or proteins [14–16]. On the other hand, 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 [17]. Therefore, the prediction of drug response, i.e., resistance or sensitivity, is essential for improving the efficacy of chemotherapy.

In the present work, we performed genome-wide screening to identify subtype-specific non-synonymous mutations (driver mutations) [18] and analyzed transcriptome data to identify the differentially expressed genes. The genes that have driver mutations and are also differentially expressed (differentially expressed driver genes: DEDGs) were screened to build the glioma subtype-specific differentially expressed driver gene network (DEDGN). We found that these DEDGNs are unique to each subtype, indicating that DEDGNs may be good targets for subtype-specific therapeutic intervention. Therefore, we further experimented to find the disease module using these DEDGs. Next, we screened the drugs from the DrugBank database [19] to target the disease modules in each subtype of LGG and GBM. To evaluate the therapeutic efficacy of drugs, we developed deep neural network (DNN)-based drug response prediction models. Lastly, we identified a few effective drugs, such as Ruxolitinib, Alpelisib, Entinostat, Vinblastine, Vorinostat and Olaparib for glioma chemotherapy. In summary, we developed a novel approach that makes use of genome-level data to identify cancer-specific alterations, which led to the construction of disease modules specific to particular glioma subtypes and the building of drug response prediction models. We anticipate that our method will be useful in enhancing the therapeutic effectiveness of the treatment of gliomas.

2. Materials and methods

2.1. Driver gene identification

We downloaded brain cancer somatic mutation data from the COSMIC database (https://cancer.sanger.ac.uk/cosmic/download) for each subtype [20]. Based on the clinical information, we stratified the patient's mutational data into different subtypes. There are a total of 281 and 123 samples of LGG and GBM, respectively. We divided 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 ($n = 36$). We used OncodriveCLUSTL [21] to find the driver mutation in subtypes. OncodriveCLUSTL is an unsupervised clustering algorithm that can detect clusters of somatic mutations across a cohort of tumor samples. Based on the mutation frequency in each gene and statistical significance (number of mutations >2 and p-value <0.05), we selected driver genes in each subtype of glioma.

2.2. Identification of differentially expressed genes (DEGs)

For computing the DEGs, we obtained RNA sequencing data of LGG ($n = 281$) and GBM ($n = 123$) patients from UCSC Xena (https://xena.ucsc.edu/) [22]. Additionally, we obtained GTEx healthy brain gene expression data ($n = 93$) 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, we preprocessed the data, and low-expressed genes were removed. We used the cut-off of log2 (RSEM +1) < 0.1 (RSEM: RNA-Seq by Expectation Maximization) in 90% of the samples 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, we identified the differentially expressed genes in each subtype of LGG and GBM 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.

2.3. Construction of subtype-specific disease module and network analysis

Human brain interactome data was retrieved from the TissueNet v.2 database [23]. Brain interactome data contains 165,240 interactions [23]. TissueNet v.2 uses human PPIs (protein-protein interactions) and tissue-specific expression patterns to make PPIs that are specific to each tissue. We implemented the Disease Module Detection (DIAMOnD) algorithm to identify the disease modules [12]. We used DEDGs that are specific to a subtype as seed genes and the TissueNet v.2 brain interactome to find the disease module. All the parameters in the DIAMOnD were kept as defaults. Cytoscape and the *igraph* package in R were used for network visualization and analysis.

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) project to develop drug response prediction models [24]. This data set contains 1001 cancer cell lines and 288 drugs. The drug and target information were derived from the DrugBank database [19]. We extracted all the drugs for which the target genes are present in the disease modules. We screened the 30 FDA-approved and investigational drugs 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, we downloaded the IC50 values of these 30 drugs for all cell lines along with the gene expression and mutational data. We have implemented the threshold to classify the IC50 values into two classes sensitive or resistant [25,26]. For each of the 30 drugs, the cell lines were classified 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 to each drug. A total of 886 cell line data were used for model development. To develop the model, we performed the following steps: 1. data preprocessing; 2. feature selection; 3. data integration; 4. model development and evaluation; and 5. model validation on external data.

Data preprocessing: We normalized the gene expression data 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, we pre-selected genes using the Pearson correlation coefficient $r < 0.5$, and then we used LASSO [27,28] to fine-select the predictor genes. For mutational data, we only used the LASSO feature selection method.

Data integration: After the feature selection step, we integrated gene expression and mutation data using a concatenated autoencoder. We used the Keras library with TensorFlow to implement the concatenated autoencoder [29] 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, a uniform kernel initializer and a linear activation function were implemented. The ReLU activation function was applied to the decoder layer.

2.5. Model development and evaluation

We applied a deep neural network classification model to predict sensitivity vs. resistance. 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 was used to compute the performance of the model. In a stratified k-fold (10-fold) CV, the dataset is split into k separate 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 over the obtained performance parameters from each test set.

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 (MCC). 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. The metrics are defined by the following equations:

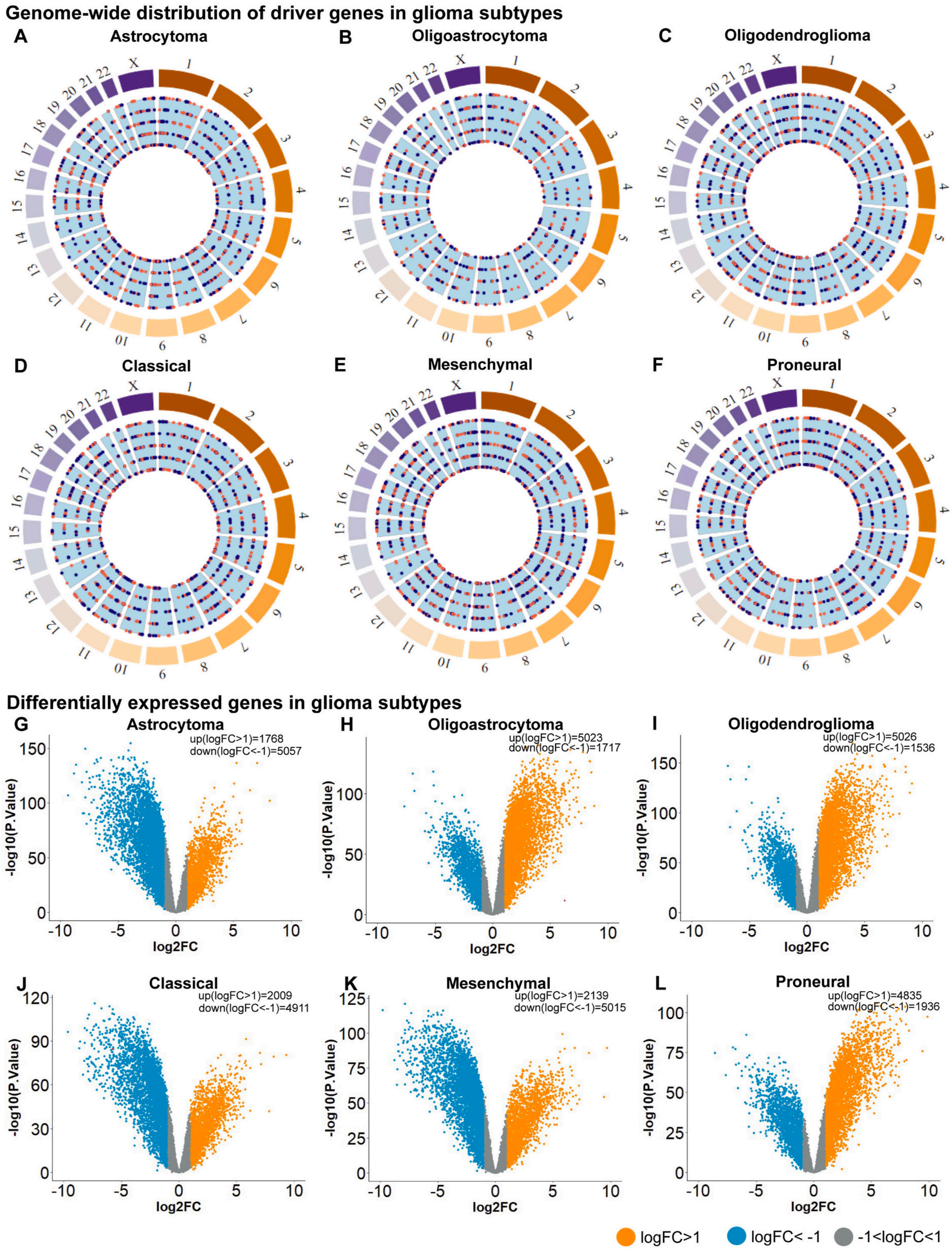
$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{2}$$

$$Specificity = \frac{TN}{TN + FP} \tag{3}$$

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$F1 - score = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{5}$$



(caption on next page)

Fig. 1. 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. G-L, 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). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

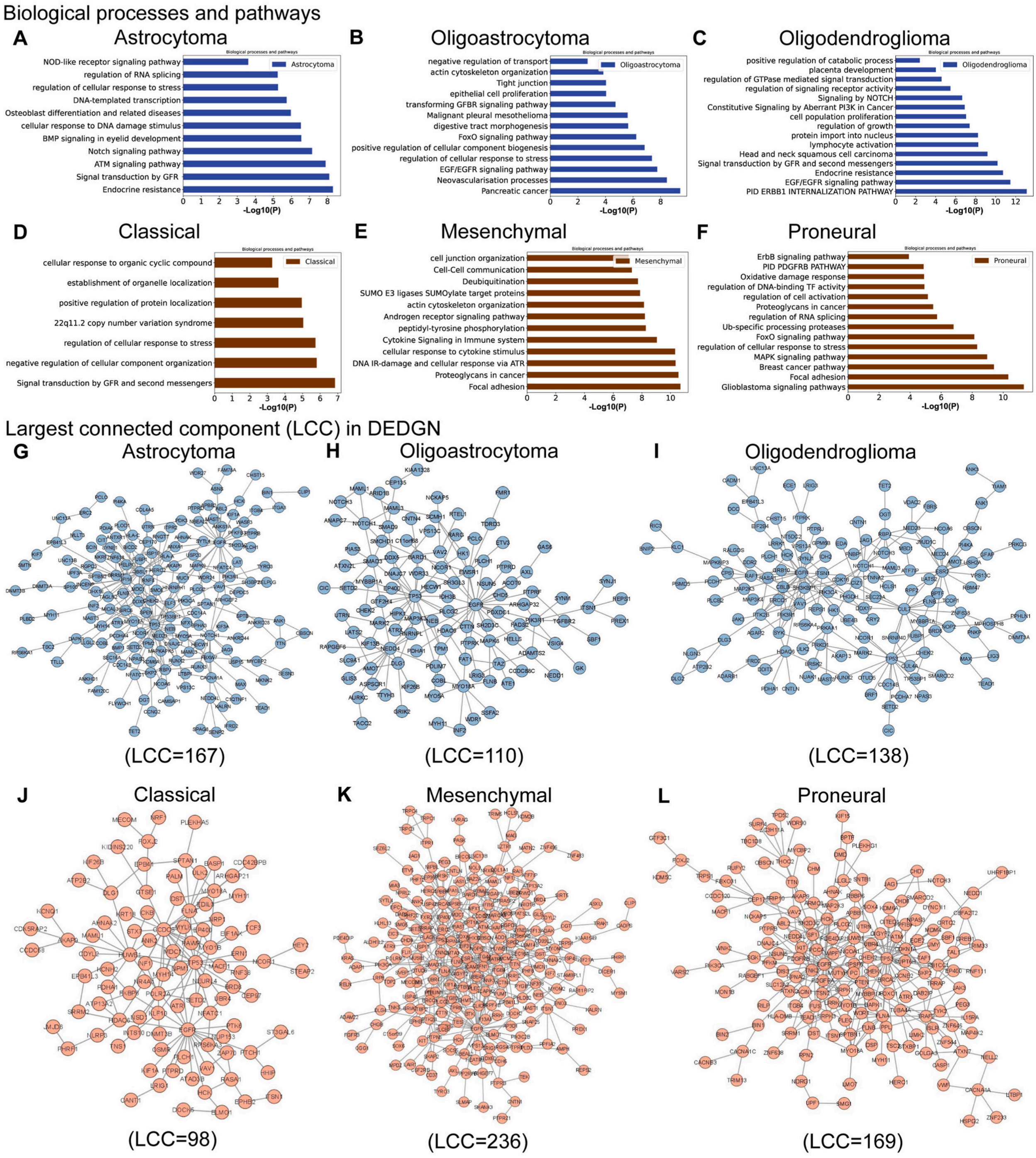


Fig. 2. The bar diagrams represent the biological process and pathway enrichment analysis of differentially expressed driver genes (DEDGs) in glioma subtypes (A–F). The highly significant (p-value<0.05) processes and pathways are shown in the figures. G–L, show the largest connected component (LCC) of the DEDGs networks in each subtype.

$$FPR = \frac{FP}{TN + FP} \tag{6}$$

$$Geometric\ mean = \sqrt[n]{(x_1 . x_2 ....x_n)} \tag{7}$$

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{8}$$

We used the sklearn.metrics library in Python to calculate the above score by importing functions such as confusion_matrix and classification_performance.

Model validation on external data:

We validated the model's performance using the dataset from the Cancer Cell Line Encyclopedia (CCLE) (https://portals.broadinstitute.org/ccle) [30]. The cell line gene expression profiles in CCLE and GDSC were generated using different platforms, and thus the data sets have significantly different magnitudes (Supplementary Fig. 1). To make these two datasets uniformly distributed, we removed the batch effect using the pyComBat package in Python [31,32]. Then the standardized gene expression profile (brain cancer cell lines) of CCLE was fed to the model for validation.

2.6. Ranking

Based on performance parameters such as accuracy, sensitivity, precision, G-mean, F-measure, FPR, and MCC, drugs were ranked. The technique for order of preference by similarity to ideal solution (TOPSIS), an established multi-criteria decision-making (MCDM) method, was used to rank the drugs.

3. Results

3.1. Genome-wide screening to identify the driver genes

Cancer mutations can be synonymous or 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. Therefore, genes with nonsynonymous mutations are identified as driver genes. dr Owing to this fact, we performed genome-wide screening to identify the non-synonymous driver mutations (p-value <0.05) in each subtype of LGG and GBM by implementing the OncodriveCLUSTL algorithm and using somatic mutation data from the COSMIC (Supplementary Table 1) [20,21]. We 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. We also found that all subtypes of GBM have a higher number of driver mutations than LGG subtypes (Supplementary Table 2). This demonstrates why GBM is more aggressive than other varieties of brain cancer. We also noticed that these mutations are scattered across the genome rather than being concentrated in a particular location (Fig. 1A–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, we identified the differentially expressed genes (DEGs) in each subtype of cancer. The genes with log2Fold Change (FC) $>$ 1 and $<$-1 and adjusted p-value <0.05 were considered DEGs (Fig. 1 G-L and Supplementary Table 3). The driver genes that are differentially expressed are named differentially expressed driver genes (DEDGs) (Supplementary Table 4). 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 (Supplementary Table 2). The combined effect of mutations in cancer driver genes and changes in gene expression can enhance the oncogenic effects [33]. 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 control cancer growth.

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

In the previous section, we observed that in all subtypes, driver genes are distributed across the genome. Many driver genes are also differentially expressed. The perturbations at two levels, i.e., differential expression and mutations, indicate their crucial role in cancer development because biological pathways and processes that involve these genes will likely be deregulated. We carried out the gene set enrichment analysis of DEDGs from each subtype to investigate the affected biological pathways and processes. We found that cancer-associated processes and pathways were enriched in different subtypes of gliomas (Fig. 2A–F). Interestingly, we 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 is prevalent; in mesenchymal focal adhesion, proteoglycan in cancer and cytokine signaling in the immune system; and in proneural glioblastoma signaling pathways and MAPK signaling pathways are prevalent. These results unequivocally show how the subtypes differ from one another in terms of their molecular characteristics.

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 [34]. Using the brain interactome from the TissueNet v.2 database [23], we built the subtype-specific protein-protein interaction network of DEDGs, named the differentially expressed driver gene network (DEDGN), to analyze the interaction pattern (Supplementary Table 5). We observed that a moderate portion of the DEDGs directly interact with each other. We calculated the size of the largest connected component (LCC) in each subtype. LCC refers to the largest subset of nodes in the network that are connected, 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. Fig. 2 G-L shows the LCC in each subtype. We 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. Therefore, we stepped into identifying the disease modules using the DIAMOnD algorithm (Fig. 3A) [12]. DIAMOnD enables us to systematically examine the local network neighborhood surrounding a particular collection of known disease proteins to find new disease proteins. We considered all DEDGs in each subtype to be known disease genes and used them as seed genes in DIAMOnD. The number of genes in the DIAMOnD disease module of each subtype of LGG and GBM was 607 in astrocytoma, 487 in oligoastrocytoma, and 578 in oligodendroglioma; 572 in classical, 675 in mesenchymal, and 574 in proneural (Supplementary Table 6). Hence, we have a higher number of disease-associated genes identified in each subtype by DIAMOnD. We observed that the size of the LCC (p-value <0.05) in each subtype provided by the DIAMOnD was much larger than the LCC in DEDGN (Fig. 3B). It should be noted that DIAMOnD LCCs contain DEDGs and relevant disease genes in the network neighborhood. 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 (Fig. 3C). 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

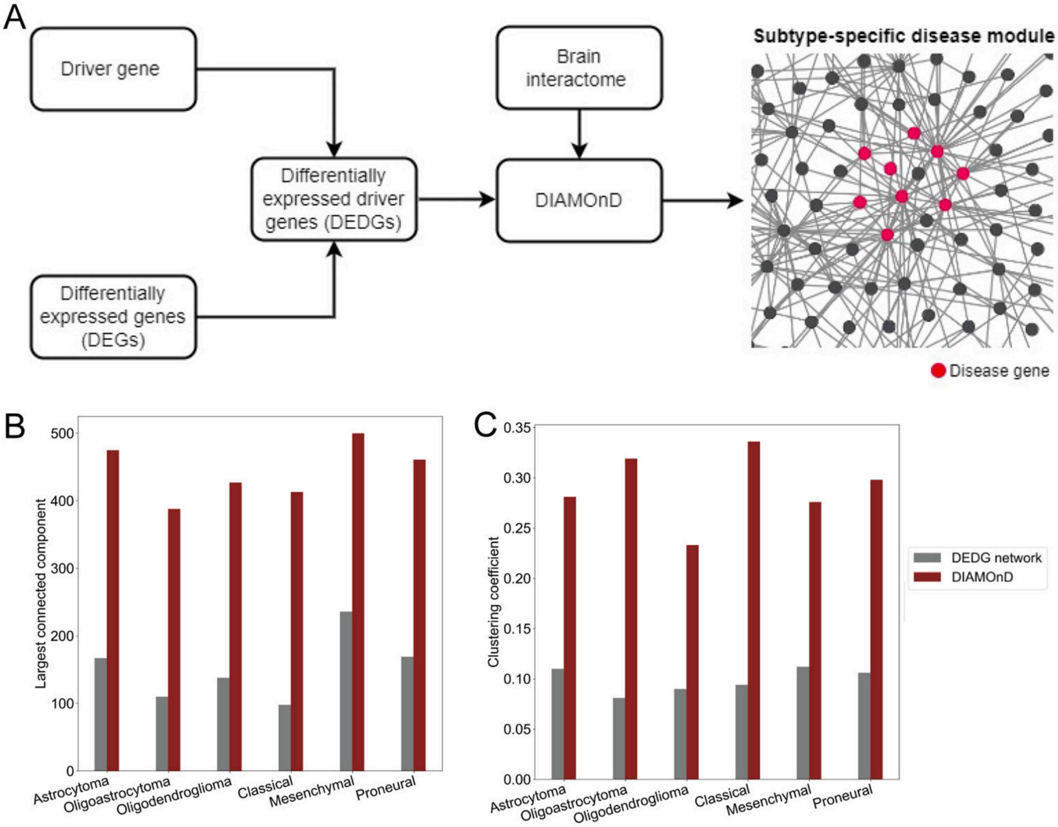


Fig. 3. Disease module in subtypes. A, the flow diagram shows the steps involved in disease module identification. DEDGs are screened from the list of driver genes and DEGs. DEDGs and brain interactome data are fed into DIAMOnD for disease module identification. B, The bar diagram compares the size of the LCC of DEDGs network (gray) and DIAMOnD disease module (brown). C, The bar diagram compares clustering coefficient of LCC of DEDGs network (gray) and DIAMOnD disease module (brown). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

pathways and aid in the development of disease. Therefore, these disease modules can be identified as targets for precision therapy of glioma subtypes.

Targeting the disease module and developing the drug response prediction model

To target the disease module in glioma subtypes, we retrieve the FDA-approved and investigational drugs from the DrugBank database. We selected the drugs for which the disease module has target genes (Supplementary Table 7). We 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 programs 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 we developed the drug response classification model 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, we found that only 30 have experimental data on brain cancer cells. Hence, we chose these 30 drugs 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 to each drug. The sample size of sensitive or resistant cell line for each drug is provided in the Supplementary Table 8. We randomly divided the GDSC dataset 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. We separately treated the gene expression and mutation data. We implemented correlation-based feature selection to eliminate multicollinearity from gene expression data. We computed the Pearson Correlation Coefficient (PCC), and genes with a PCC > 0.5 were dropped. The remaining 5233 genes were taken for dimensionality reduction using LASSO. We also employed LASSO feature selection 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 [35,36]. 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, we used one hidden layer for data integration, and the dimension of the bottleneck layer was set to 64. An autoencoder consists of two parts: an encoder and a decoder network (Fig. 4). The latent variables extracted from the bottleneck layer were utilized in the decoder network to reconstruct the original input data. This process was carried out to measure the reconstruction loss, which serves as an indicator of the autoencoder's efficacy. The reconstruction loss was calculated

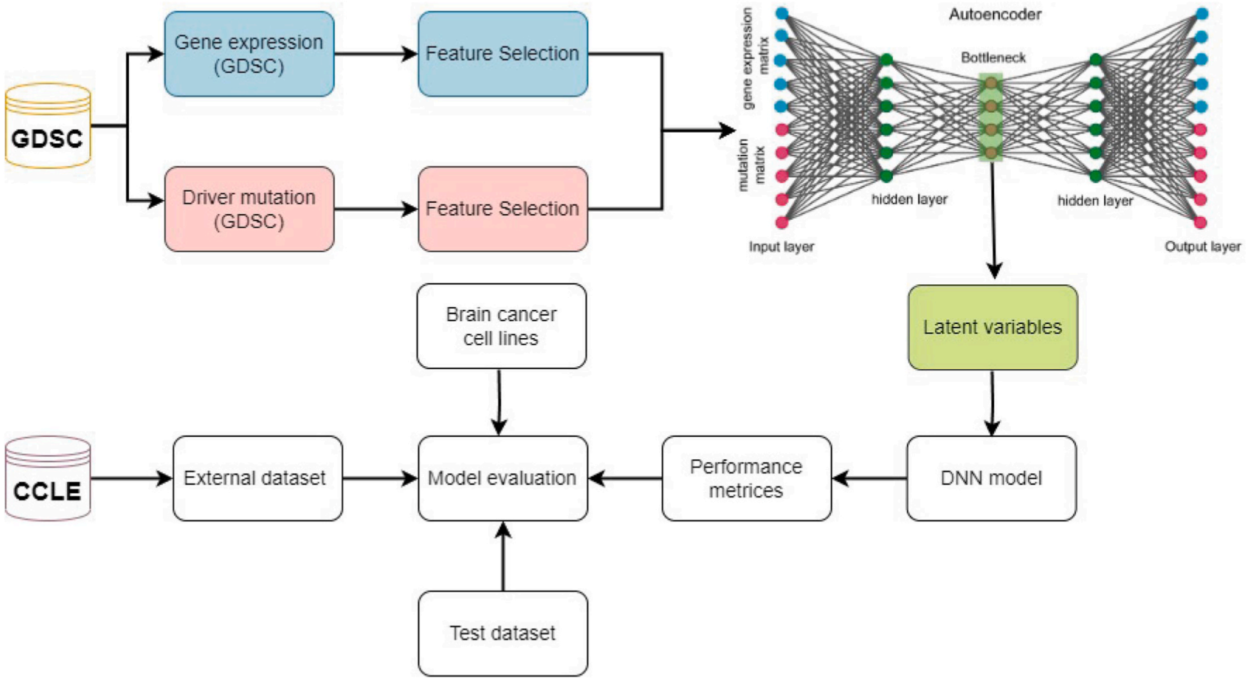


Fig. 4. The overall workflow of drug response model development. The gene expression and mutation data from GDSC are subjected to feature selection, and both data are integrated using an autoencoder. The latent variable from the bottleneck layer is used for developing the DNN model. The model validation was performed using test data, brain cancer data, and external data from CCLE.

using the mean squared error (MSE). We 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. Subsequently, we construct the DNN model to predict the drug's response, specifically determining its sensitivity or resistance, by utilizing the latent variables derived from the bottleneck layer of the autoencoder. In order to identify the optimized set of hyperparameters, we employed the grid search method. 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 materials and methods, eq. [1] to eq. [8]). The performance matrix for all 30 drugs is provided in the Supplementary Table 9. Based on the performance parameters, the drugs were ranked using the MCDM tool TOPSIS. Ruxolitinib topped the list, and the accuracy of the model was 96.26% (±0.02). The precision and specificity of the model were >0.90. Upon further investigation, we learned that Ruxolitinib is a highly effective JAK/STAT signaling pathway inhibitor. It can suppress the invasion and formation of tumors in glioma cells [37]. This drug is also in clinical trials for glioma treatment (https://clinicaltrials.gov/). Further, we 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. Fig. 5 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. Except for tretinoin, the model accuracy varies from 72.25 to 91.25 using only brain cancer cell line data (Fig. 5C). We noticed that, due to a smaller sample size, the accuracy was not correctly predicted for tretinoin. A highly sensitive and specific model for drug response prediction is always ideal. Therefore, we used the receiver operating characteristic (ROC) curve 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. Fig. 5D-M shows the area under the ROC curve (AUC) of the DNN models of the top 10 drugs. We 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. All models show consistent prediction accuracy in training, testing, and brain cancer cell data. Previous publications have indicated that some of the top 10 drugs have shown the potential to inhibit the growth of glioma cells. For example, Entinostat, a histone deacetylase inhibitor, has demonstrated the ability to reduce the growth of GBM [38]. Vorinostat, an FDA-approved drug, is already used to treat cutaneous T-cell lymphoma, but it is now in a Phase II clinical trial to treat recurrent glioblastoma multiforme [39]. Vinblastine shows sensitivity for both LGG and GBM [40,41] and it is in clinical trials for the treatment of these cancers. Other drugs such as Olaparib [42], Crizotinib [43], and Trametinib [44] 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. Next, we predict the subtype-specific drug sensitivity of 30 different drugs against 49 brain cancer cell lines from different tissues, using the saved models to assess their potential clinical utility. We extracted the features from gene expression and mutation data from particular cell line data, integrated these two data sets, i.e., gene expression and mutation, using an autoencoder,

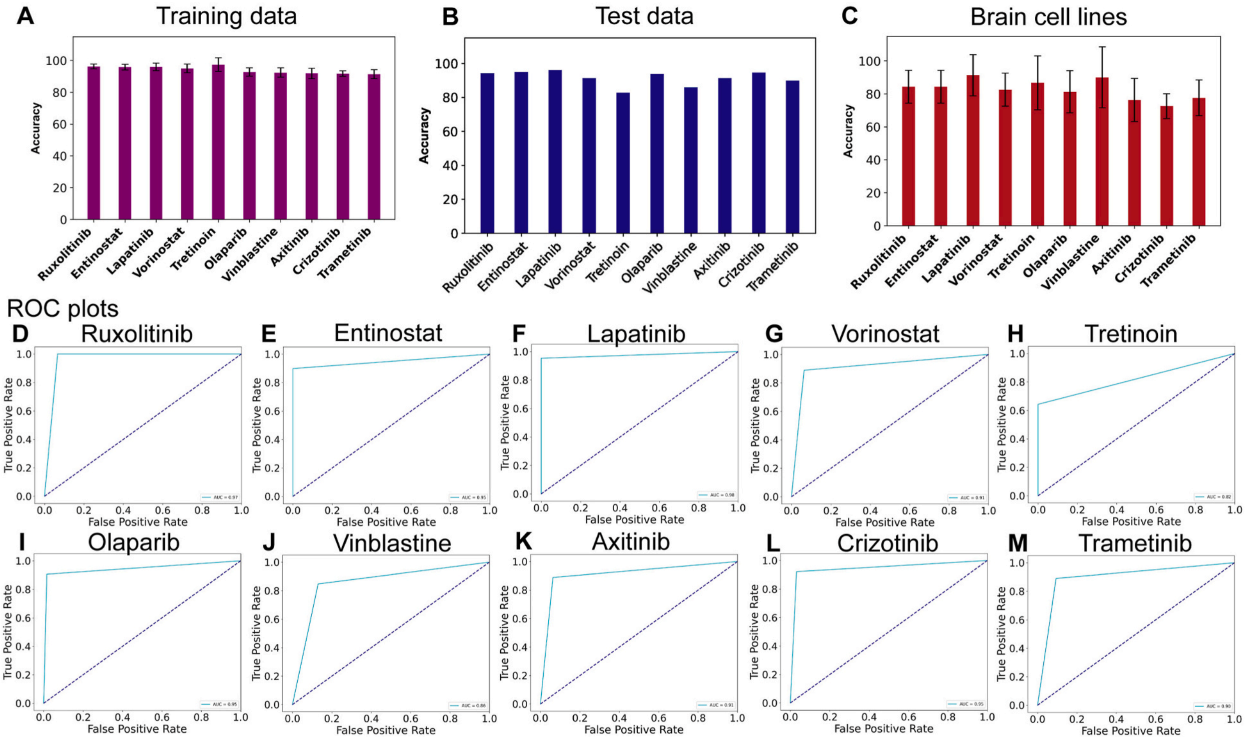


Fig. 5. Classification accuracy of DL models A, training B, test and C, brain cancer cell lines of top 10 drugs. The complete performance matrix is provided in the Supplementary Table 9. D–M, ROC plots of the top 10 drugs.

and then fed this integrated data into the 30 different drug-specific DNN models. Lastly, we predicted the sensitivity or resistance of a drug against a particular brain cancer cell line. The drug sensitivity data for all 30 drugs for 49 brain cancer cells is shown in Fig. 6. We acquired the cell line's lineage from ATCC (https://www.atcc.org/) and cellosaurus (https://www.cellosaurus.org/) 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. We 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, we used gene expression and mutations as features while developing the models. This provides a thorough understanding of the importance of using genomics data to predict drug responses specific to subtypes to improve therapeutic efficacy. Lastly, we validated this DL framework with external datasets from CCLE, and the accuracy of prediction for drugs Erlotinib, Lapatinib, Nilotinib, and Sorafenib was fairly accurate (Supplementary Table 10). These results show that our models were able to consistently predict accurate drug responses. But, before the current framework is used in a clinical setting, it needs to be examined in terms of *in vitro* and *in vivo* drug efficacy assays and sensitivity data across the many types of cancer.

4. Discussion

The clinical development of targeted and personalized brain cancer treatments continues to be a significant issue. Brain cancer encompasses a wide range of distinct forms, and the presence of unique genetic aberrations in each type is a significant obstacle in developing effective therapies. The identification of a disease-specific biomarker for targeted therapy is a widely employed approach in

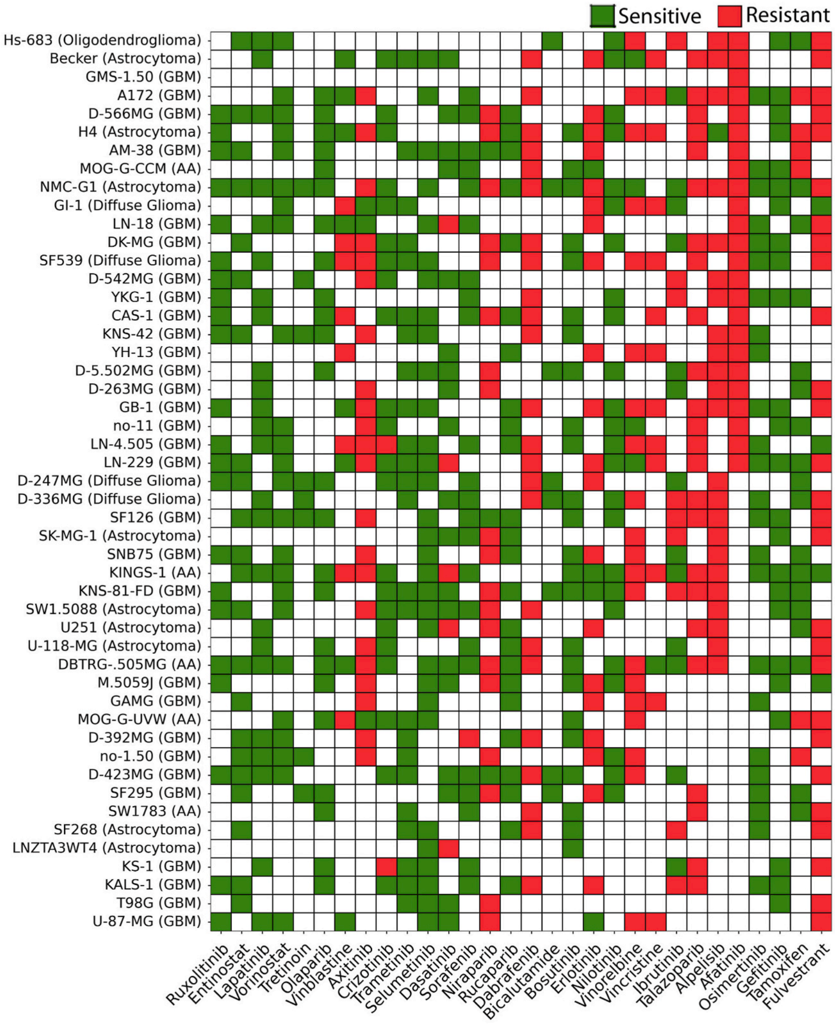


Fig. 6. Prediction of drug sensitivity in brain cancer cell lines. The heat map represents the drug sensitivity data for 49 brain cancer cell lines against 30 drugs. The red color indicates the resistant cell lines and the green color indicates the sensitive cell lines. The origin (or subtype) of each of the cell lines is mentioned in the figure. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

the development of anti-cancer drugs [45,46]. Nevertheless, the presence of molecular heterogeneity in cancer cells often affects the efficacy of targeted therapy, leading to the frequent emergence of drug resistance [47]. 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 [48]. Therefore, we 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, we identified the DEDGs, 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, we selected drugs from the DrugBank database to target these disease modules. Patients' responses to drugs, however, differ greatly from one another due to the diversity of molecular profiles. To address this further, we developed a DL-based framework to predict drug responses using gene expression, mutations, and IC50 values from large-scale experimental data. We design the novel framework by combining LASSO-based feature selection, autoencoder-based data integration, and then prediction using the DNN. We 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, we showed that cancer cell lines from various subtypes of glioma exhibit varying degrees of drug sensitivity. Prior research has documented the drug response model for a specific form of cancer [49,50]. However, the present study demonstrates that DL-based models can also be utilized for predicting drug responses for a distinct subtype of cancer. The existing drug response prediction models mainly relied on mono-omics data [51,52]. However, we have taken a step further by using gene expression and mutation data for the construction of the model. Furthermore, we have demonstrated within a single framework that the tool sets of network medicine and algorithms of DL might possess enhanced therapeutic significance in the context of personalized therapy.

However, 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 we expect that this problem will be solved soon because the size of datasets is growing rapidly. 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. We hope 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.

Ethics statement

This article does not contain any studies with human participants performed by any authors.

Data availability

Cancer patient transcriptome and clinical data are available in UCSC Xena database Cancer patient transcriptome and clinical data are available in UCSC Xena database (https://xenabrowser.net/datapages/). Mutational data is available in COSMIC (the Catalogue Of Somatic Mutations In Cancer) (https://cancer.sanger.ac.uk/cosmic/download). Cancer cell drug sensitivity data is available in the Genomics of Drug Sensitivity in Cancer (GDSC) (https://www.cancerrxgene.org). The external datasets were retrieved from CCLE (https://portals.broadinstitute.org/ccle).The drug-target information is available at the drugBank database (https://go.drugbank.com/).

CRediT authorship contribution statement

Sana Munquad: Writing – review & editing, Methodology, Formal analysis. **Asim Bikas Das:** Writing – review & editing, Writing – original draft, Supervision, Investigation, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

A.B.D. was supported by Science and Engineering Research Board (Mathematical Research Impact Centric Support), DST, Government of India. A.B.D. thank the National Institute of Technology, Warangal, for providing the research scholar scholarship.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2024.e27190.

References

[1] M.A. Qazi, D. Bakhshinyan, S.K. Singh, Deciphering brain tumor heterogeneity, one cell at a time, Nat. Med. 25 (10) (2019) 1474–1476.

[2] D.N. Louis, A. Perry, P. Wesseling, D.J. Brat, I.A. Cree, D. Figarella-Branger, et al., The 2021 WHO classification of tumors of the central nervous system: a summary, Neuro Oncol. 23 (8) (2021) 1231–1251.

[3] D.J. B, R.G. V, A. KD, Y. WK, S. SR, L.A. C, et al., Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas, N. Engl. J. Med. 372 (26) (2015) 2481–2498.

[4] Q.T. Ostrom, H. Gittleman, P. Farah, A. Ondracek, Y. Chen, Y. Wolinsky, 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).

[5] Q. Wang, B. Hu, X. Hu, H. Kim, M. Squatrito, L. Scarpace, et al., Tumor evolution of glioma-intrinsic gene expression subtypes associates with immunological changes in the microenvironment, Cancer Cell 32 (1) (2017) 42–56.e6.

[6] S. Munquad, T. Si, S. Mallik, A.B. Das, Z. Zhao, A deep learning-based framework for supporting clinical diagnosis of glioblastoma subtypes, Front. Genet. (2022) 13.

[7] S. Munquad, T. Si, S. Mallik, A. Li, A.B. Das, Subtyping and grading of lower-grade gliomas using integrated feature selection and support vector machine, Brief Funct. Genomics 21 (5) (2022) 408–421.

[8] P. Zhang, Q. Xia, L. Liu, S. Li, L. Dong, Current opinion on molecular characterization for GBM classification in guiding clinical diagnosis, prognosis, and therapy, Front. Mol. Biosci. 7 (2020).

[9] D. Ostroverkhova, T.M. Przytycka, A.R. Panchenko, Cancer driver mutations: predictions and reality, Trends Mol. Med. 29 (7) (2023) 554–566.

[10] J. Foo, L.L. Liu, K. Leder, M. Riester, Y. Iwasa, C. Lengauer, et al., An evolutionary approach for identifying driver mutations in colorectal cancer, PLoS Comput. Biol. 11 (9) (2015).

[11] L.Y.H. Lee, J. Loscalzo, Network medicine in pathobiology, Am. J. Pathol. 189 (7) (2019) 1311–1326.

[12] S.D. Ghiassian, J. Menche, A.L. Barabási, 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 (4) (2015).

[13] J. Wu, Q. Zhang, G. Li, Identification of cancer-related module in protein-protein interaction network based on gene prioritization, J. Bioinf. Comput. Biol. 20 (1) (2022).

[14] E. Guney, J. Menche, M. Vidal, A.L. Barábasi, Network-based in silico drug efficacy screening, Nat. Commun. 7 (2016).

[15] X. Gan, Z. Shu, X. Wang, D. Yan, J. Li, S. Ofaim, et al., Network medicine framework reveals generic herb-symptom effectiveness of traditional Chinese medicine, Sci. Adv. 9 (43) (2023).

[16] A. Sharma, J. Menche, C. Chris Huang, T. Ort, X. Zhou, M. Kitsak, et al., A disease module in the interactome explains disease heterogeneity, drug response and captures novel pathways and genes in asthma, Hum. Mol. Genet. 24 (11) (2015) 3005–3020.

[17] E.O. Mahgoub, W.C. Cho, M. Sharifi, M. Falahati, H.A. Zeinabad, H.E. Mare, et al., Role of functional genomics in identifying cancer drug resistance and overcoming cancer relapse, Heliyon 10 (1) (2024) e22095.

[18] D. Ostroverkhova, T.M. Przytycka, A.R. Panchenko, Cancer driver mutations: predictions and reality, Trends Mol. Med. 29 (7) (2023) 554–566.

[19] D.S. Wishart, Y.D. Feunang, A.C. Guo, E.J. Lo, A. Marcu, J.R. Grant, et al., DrugBank 5.0: a major update to the DrugBank database for 2018, Nucleic Acids Res. 46 (D1) (2018) D1074–D1082.

[20] S.A. Forbes, D. Beare, P. Gunasekaran, K. Leung, N. Bindal, H. Boutselakis, et al., COSMIC: exploring the world's knowledge of somatic mutations in human cancer, Nucleic Acids Res. 43 (Database issue) (2015) D805–D811.

[21] C. Arnedo-Pac, L. Mularoni, F. Muiños, A. Gonzalez-Perez, N. Lopez-Bigas, OncodriveCLUSTL: a sequence-based clustering method to identify cancer drivers, Bioinformatics 35 (22) (2019) 4788–4790.

[22] M.J. Goldman, B. Craft, M. Hastie, K. Repečka, F. McDade, A. Kamath, et al., Visualizing and interpreting cancer genomics data via the Xena platform, Nat. Biotechnol. 38 (6) (2020) 675–678.

[23] O. Basha, R. Barshir, M. Sharon, E. Lerman, B.F. Kirson, I. Hekselman, et al., The TissueNet v.2 database: a quantitative view of protein-protein interactions across human tissues, Nucleic Acids Res. 45 (D1) (2017) D427–D431.

[24] W. Yang, J. Soares, P. Greninger, E.J. Edelman, H. Lightfoot, S. Forbes, et al., Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells, Nucleic Acids Res. 41 (2013) (Database issue).

[25] F. Zhang, M. Wang, J. Xi, J. Yang, A. Li, A novel heterogeneous network-based method for drug response prediction in cancer cell lines, Sci. Rep. 8 (1) (2018).

[26] F. Iorio, T.A. Knijnenburg, D.J. Vis, G.R. Bignell, M.P. Menden, M. Schubert, et al., A landscape of pharmacogenomic interactions in cancer, Cell 166 (3) (2016) 740–754.

[27] R. Muthukrishnan, R. Rohini, LASSO: a feature selection technique in predictive modeling for machine learning, in: 2016 IEEE International Conference on Advances in Computer Applications, ICACA 2016, 2017, pp. 18–20.

[28] J. Song, Z. Xu, L. Cao, M. Wang, Y. Hou, K. Li, The discovery of new drug-target interactions for breast cancer treatment, Molecules 26 (24) (2021).

[29] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, et al., TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems, 2016.

[30] N. Stransky, M. Ghandi, G.V. Kryukov, L.A. Garraway, J. Lehár, M. Liu, et al., Pharmacogenomic agreement between two cancer cell line data sets, Nature 528 (7580) (2015) 84–87.

[31] A. Behdenna, J. Haziza, C.A. Azencott, A. Nordor, pyComBat, a Python tool for batch effects correction in high-throughput molecular data using empirical Bayes methods, bioRxiv (2021) 995431.

[32] J.T. Leek, W.E. Johnson, H.S. Parker, A.E. Jaffe, J.D. Storey, The sva package for removing batch effects and other unwanted variation in high-throughput experiments, Bioinformatics 28 (6) (2012) 882–883.

[33] H. Tegally, K.H. Kensler, Z. Mungloo-Dilmohamud, A.W. Ghoorah, T.R. Rebbeck, S. Baichoo, Discovering novel driver mutations from pan-cancer analysis of mutational and gene expression profiles, PLoS One 15 (11) (2020).

[34] A.L. Barabási, N. Gulbahce, J. Loscalzo, Network medicine: a network-based approach to human disease, Nat. Rev. Genet. 12 (1) (2011) 56–68.

[35] M. Kang, E. Ko, T.B. Mersha, A roadmap for multi-omics data integration using deep learning, Briefings Bioinf. 23 (1) (2022).

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

[37] E. Delen, O. Doğanlar, 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 (4) (2020) 444–454.

[38] R. Chen, M. Zhang, Y. Zhou, W. Guo, M. Yi, Z. Zhang, et al., The application of histone deacetylases inhibitors in glioblastoma, J. Exp. Clin. Cancer Res. 39 (1) (2020).

[39] E. Galanis, K.A. Jaeckle, M.J. Maurer, J.M. Reid, M.M. Ames, J.S. Hardwick, et al., Phase II trial of vorinostat in recurrent glioblastoma multiforme: a north central cancer treatment group study, J. Clin. Oncol. 27 (12) (2009) 2052–2058.

[40] F.C. Kipper, A.O. Silva, A.L. Marc, G. Confortin, A.V. Junqueira, E.P. Neto, et al., Vinblastine and antihelmintic mebendazole potentiate temozolomide in resistant gliomas, Invest. N. Drugs 36 (2) (2018) 323–331.

[41] S. Vairy, G. Le Teuff, F. Bautista, E. De Carli, A.I. Bertozzi, A. Pagnier, 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 (1) (2020).

[42] L.R. Schaff, M. Kushnirsky, A.L. Lin, S. Nandakumar, C. Grommes, A.M. Miller, et al., Combination Olaparib and temozolomide for the treatment of glioma: a retrospective case series, Neurology 99 (17) (2022) 750–755.

[43] A. Junca, C. Villalva, G. Tachon, P. Rivet, U. Cortes, K. Guilloteau, et al., Crizotinib targets in glioblastoma stem cells, Cancer Med. 6 (11) (2017) 2625–2634.

[44] Y.K. Banasavadi-Siddegowda, S. Namagiri, Y. Otani, H. Sur, S. Rivas, J.P. Bryant, et al., Targeting protein arginine methyltransferase 5 sensitizes glioblastoma to trametinib, Neurooncol. Adv. 4 (1) (2022).

[45] A. Drilon, G.R. Oxnard, D.S.W. Tan, H.H.F. Loong, M. Johnson, J. Gainor, et al., Efficacy of selpercatinib in RET fusion-positive non-small-cell lung cancer, N. Engl. J. Med. 383 (9) (2020) 813–824.

[46] S. Zhao, Z. Zhang, J. Zhan, X. Zhao, X. Chen, L. Xiao, et al., Utility of comprehensive genomic profiling in directing treatment and improving patient outcomes in advanced non-small cell lung cancer, BMC Med. 19 (1) (2021).

[47] E.Y. Rosen, H.H. Won, Y. Zheng, E. Cocco, D. Selcuklu, Y. Gong, et al., The evolution of RET inhibitor resistance in RET-driven lung and thyroid cancers, Nat. Commun. 13 (1) (2022).

[48] B.V.S.K. Chakravarthi, S. Nepal, S. Varambally, Genomic and epigenomic alterations in cancer, Am. J. Pathol. 186 (7) (2016) 1724–1735.

[49] L. Parca, G. Pepe, M. Pietrosanto, G. Galvan, L. Galli, A. Palmeri, et al., Modeling cancer drug response through drug-specific informative genes, Sci. Rep. 9 (1) (2019).

[50] R. Qureshi, S.A. Basit, J.A. Shamsi, X. Fan, M. Nawaz, H. Yan, et al., Machine learning based personalized drug response prediction for lung cancer patients, Sci. Rep. 12 (1) (2022).

[51] E.A. Clayton, T.A. Pujol, J.F. McDonald, P. Qiu, Leveraging TCGA gene expression data to build predictive models for cancer drug response, BMC Bioinf. 21 (Suppl 14) (2020).

[52] S. Chawla, A. Rockstroh, M. Lehman, E. Ratther, A. Jain, A. Anand, et al., Gene expression-based inference of cancer drug sensitivity, Nat. Commun. 13 (1) (2022).