Multimodal Feature-Based Drug Response Prediction Using Light Gradient Boosting Machine and Gene Expression Analysis in Brain Tumors

Awais Raza Zaidi[1*], Abdul Majid[2], Muhammad Bilal[3], Tuba Majid[4]
awaisraza_19@pieas.edu.pk*, abdulmajiid@pieas.edu.pk, muhammadbilal@pieas.edu.pk, tmajid@ethz.ch
Pakistan Institute of Engineering & Applied Sciences[1,2,3], Department of Mechanical and Process Engineering, ETH Zurich[4]

ABSTRACT

Due to their recurrence and complex biology, brain tumors remain among the most challenging cancers and significantly contribute to global cancer mortality. With continuous development in precision oncology, accurately predicting patient-specific drug responses is essential for effective treatment and drug design. In this study, we propose a novel multimodal framework utilizing a Light Gradient Boosting Machine (LightGBM) regressor for brain tumors drug response prediction. The model integrates both biological and chemical data features using three modality-specific Variational Autoencoders (VAEs) to encode gene expression, gene mutation, and drug molecular fingerprint features respectively. The integrated feature representations are used by the LightGBM regressor to predict the half-maximal inhibitory concentration IC_{50} of drugs. Reliable results are obtained using subject-level stratified nested cross-validation. Our model has yielded improved RMSE and correlation R^2 values 1.12 and 0.76, respectively. These results are statistically significant ($p<0.05$) as compared to several existing models. Furthermore, using proposed model five FDA-approved drugs with the most accurately predicted IC_{50} values were identified. Using the statistical analysis of Glioblastoma (GBM) cell lines, we explored several over-expressed genes: EGFR, MKI67, BIRC5, TOP2A, AURKA and under-expressed genes: GFAP, MBP, NEFL, SLC1A2, PLP1., highlighting their biological roles in tumor progression and suppression. For clinical perspective, we have carried out the Survival analysis that showed that highly expressed tumor genes did not significantly affect normal patient survival ($p > 0.05$). It is anticipated that this study would be useful in precision oncology for anticancer drug development.

Key words: Precision Oncology, SMILES, Autoencoders, LightGBM, Differential Gene Expression

1. Introduction

Brain tumors are highly heterogeneous diseases that is characterized by complex biological mechanisms and limited therapeutic effectiveness. Brain tumours contribute significantly to global cancer mortality. As shown in Figure 1, mortality from brain tumours is high across all continents. This figure highlights that brain tumour deaths are highest in Asia and the lowest in Oceania. However, brain tumour death in Europe has a high impact, while other continents have moderate levels of mortality. This mortality differences reflect healthcare quality across the population[1].

The abnormal growth of tumor cells is mainly caused by DNA damage and uncontrolled invasion into surrounding tissues. Such rapid and irregular cell activity leads to malfunctioning of vital organs and leads to severe health complications. Genetic disruptions in brain tumors can harm several biological mechanisms such as cell apoptosis [2], DNA repair mechanisms and cell cycle checkpoints.

The Glioblastoma multiforme (GBM) is one of the most aggressive and hard to treat forms in the brain tumors. This type of tumor is the most common primary brain tumor [3]. In adults, brain tumors often progress rapidly and show resistance to existing treatments. Despite rigorous treatment protocols, GBM frequently develops drug-resistant clones [4]. This complicates the therapeutic decision-making efforts and reduce treatment efficacy. Due to its selective nature, the blood–brain barrier (BBB) poses a major challenge in the effective treatment of brain tumors [5]. This prevents over 95% of therapeutic compounds from entering the central nervous system. Even when the BBB becomes leaky in glioblastoma, drug build up stays low because transporters like P-gp push drugs back into the blood.

In GBM tumour tissues, gene expression levels play major clinical and biological role in pathway dysregulation. The variation in gene expression level would help to identify significantly unregulated genes. This would highlight their potential role in cancer initiation and progression. These genes are widely recognized as key biomarkers of cell proliferation. As compared to the gene expression in the normal tissue, genes in the brain tumour (EGFR, MKI67, BIRC5, TOP2A, and AURKA) have shown over/under expression. They actively involved in tumour initiation and progression. These highly expressed genes can be potential biomarker for therapeutic targets. Among these genes, AURKA is a key regulator of mitosis and is strongly associated with uncontrolled cell proliferations[6][7].
In this study, through the statistical analysis of GBM cell lines, we identified over-expressed and under-expressed genes by comparing the expression levels of brain tumour samples with normal tissue samples. We also highlighted the biological roles of these highly expressed genes. Further, in the study, we statistically found under expressed genes such as GFAP, MBP, NEFL, SLC1A2, and PLP1. Their abnormal expression highlights their potential biomarker and therapeutic relevance. The detailed analysis of these over/under expressed genes is given in the differential gene expression analysis. Furthermore, we analysed these genes using the cancer genome atlas (TCGA)[8] data that revealed their statistical significance between tumour and normal samples. Although, their survival analysis through Kaplan-Meier (KM) curves indicates no strong correlation between their expression levels and patient outcomes [9].

Due to less availability of genomic profile, cancer is being treated through conventional means. Still major cancer treatment methodologies include surgery, chemotherapy, immunotherapy, radiotherapy etc. However, these methods have certain limitations explained in [10]. In this scenario, the development in precision oncology would be helped by accurately predicting patient-specific drug responses[11]. It is very important for effective drug design and treatment strategies. However, with the increasing availability of large genomic and chemoinformatic datasets such as GDSC[12], CCLE [13], TCGA [14], ChEMBL [15], DrugBank [16], and PubChem [17]. This publicly available dataset has accelerated research in the computational oncology for personalized cancer therapies.

In the study, we have developed several traditional machine learning (ML) models (support vector machines [18]and random forests [19], multi layer perceptron [20]) for drug response prediction. However, these simple models often faced challenges in the high dimensionality of the biological data and complex heterogeneity present in the genomic data. In this scenario, we employed advanced ML method such as variational Autoencoders (VAEs) for useful feature extraction [21], [22], [23]. We developed VAE-based Gradient Boosting Machine (LightGBM) regressor to enhance the prediction accuracy of drug response prediction.

In this study, first, we developed multimodal learning framework that integrates genomic and chemoinformatic data. This framework employs three modality-specific VAEs to encode gene expression, gene mutation, and drug molecular fingerprint data. The generated features from multimodal framework are processed through LightGBM regressor. Our prediction model has utilized light gradient boosting machine for predicting drug response of minimal inhibitory concentration IC_{50} values. The VAEs multimodal framework preserved the biologically important features while modelling complex data. The proposed LightGBM model has performed better due to its leafwise tree growth and histogram-based tuning modelling. Our model has performed well in comparison with several drug response prediction models such as KMBTL Kernelized Multi-task Learning [24], the hybrid convolutional framework, the neighbour-based collaborative filtering model, and the Weighted Graph Regularized Collaborative Matrix Factorization (WGRCMF) [25]. The detailed description of these prediction models and conventional models such as MLP, SVR and RF are given in the supplementary materiel file. We evaluate and compare performance of drug response prediction models in terms of root mean squared error (RMSE) and coefficient of determination R^2.

Main Contributions:

- In the study, we developed a novel VAE-LightGBM regressor by integrating genomic and chemical features to predict drug response.

- Using proposed regressor model, we identified most accurately predicted IC50 values of drugs such as Erlotinib, Linifanib, Pictilisib, and Acetyleshikonin. Furthermore, by analyzing GDSC dataset based on the lowest IC50 values, we identified most potent FDA-approved drugs such as Midostaurin, GW-2580, FMK, Imatinib, and A-443654.

- From the statistical analysis using volcano plot, we identified several over-expressed genes and under-expressed genes. Furthermore, we compared their gene expression in tumor samples with normal tissue samples. The biological role of these genes is highlighted.

- For clinical perspective, the survival analysis of these over/under expressed genes is carried out.

Remaining part of the paper is summarized as follows:

The remainder of this research is structured as follows: In Sect. 2, we will provide our proposed ML methodology. Section 3 provides detailed feature extraction and dataset preprocessing. The discussion of the experiment results is covered in Sect. 4, followed by the conclusion and future directions.

2. Proposed methodology of LightGBM model

We developed LightGBM model using VAE based multimodal learning framework for drug response prediction of brain tumor cell lines. The proposed model uses three different VAE to handle three biological features gene expression, gene mutation, and cheminformatics features like molecular fingerprints of FDA approved anticancer drugs. We obtained gene expression and gene mutation biological features from GDSC dataset. From a biological perspective, gene expression patterns illustrate the cellular context of medication response from cancer cell lines of patients suffering from brain tumor. The variable threshold-based filtering method was employed to handle the curse of dimensionality. This helped to prevent model overfitting while handling large data instances. These preprocessed biological and cheminformatics features are integrated and fed into the proposed model for the prediction of drug response as IC_{50} values. Figure 2 represents the data-processing module and the proposed framework for regression model prediction. The data from gene expression, gene mutation, and drug chemical structures are preprocessed and feed into VAE framework separately with the aim of useful feature extraction. These features are then concatenated into a single vector that are used to train the LightGBM regressor model to predict drug response for IC_{50} values.

2.1 Data Preprocessing stage:

Traditional drug screening methods such as 2D cell culture assays, face problems to capture the complete spatial complexity and cellular heterogeneity of tumor environments. This limitation highlights the need for advanced ML–based approaches to model pharmacological responses more accurately. With the availability of large-scale genomic datasets and powerful ML algorithms, we can develop reliable drug response prediction models. Public repositories like GDSC and CCLE have comprehensive pharmacogenomic data for improved cancer drug response prediction. In the current work, we selected 1,478 genes in 56 brain tumor cell lines from the GDSC dataset. The 2nd dataset is comprised on anticancer drugs that is used for brain tumor treatment. We have employed SMILES (simplified molecular input line entry system) format[26] of drugs retrieved from PubChem dataset. The molecular descriptors are converted into the binary representation using RDkit python Library [27].

In this study, first, datasets are collected from public cancer research databases. Then, we employed three types of data modality such as gene expression, gene mutation, and drug chemical structure. We retrieved gene expression profiles of 1,478 genes from 56 brain tumor cell lines using the GDSC database. This helped us to understand tumor biology and its drug response. Further, we also collected binary mutation data for the same genes and cell lines, marking 0 for wild type and 1 for mutation. In addition, we have obtained chemical structures of 220 anticancer drugs from the PubChem database. These were converted from SMILES format to ECFP4 fingerprints. Among the 56 brain tumor cell lines, most are classified as glioblastoma multiforme (GBM), however, others included low-grade glioma (LGG) and medulloblastoma [28]. This provide a diverse range of tumor subtypes for analyzing therapeutic responses. Table 1 presents an exemplary 10 anticancer drugs including their PubChem compound IDs, molecular formulas, SMILES representations, and 2D structures.

2.2 Data Preprocessing

Firstly, we removed those cell drug combinations, for which IC_{50} values were missing or redundant to retain complete data instances. Gene expression profiles were normalized through z-score normalization for symmetry among genes. Gene mutation data were binarized, where 0 indicated wild-type and 1 represented mutation status. This ensured a consistent representation that would enable effective latent feature learning using Variational Autoencoders (VAEs). In the 2nd step, for drug chemoinformatic features, SMILES notation is used for drug to transform into 1024 bit binary Extended-Connectivity Fingerprints (ECFP4) using the RDkit library. This library is widely employed for encoding molecular structure information. Next, dataset of three different modalities (gene expression, gene mutation and chemical structures of drugs) were integrated by aligning corresponding drugs with cancer cell lines.

2.3 Cross-Validation Technique

To address the limitations of the single hold-out split, we have employed subject-level stratified 5-fold cross-validation (CV). This technique will help to improve the robustness and generalizability of the model. The integrated dataset is obtained by combining the three modalities of gene expression, gene mutation, and drug chemical structures. All samples were derived from 56 unique brain tumor cell lines. For data partition, subject-level 5-fold cross-validation was performed. In the scheme, the cell lines were divided into five folds, with stratification based on tumor subtypes such as glioblastoma multiforme (GBM), low-grade glioma (LGG), and medulloblastoma. This would ensure that each fold preserved the overall subtype distribution. To prevent data leakage, we placed all drug–cell line samples from the same cell line in one-fold. This ensured that each cell line appeared only one-fold and that the test data contained only unseen cell lines.

Nested Cross-Validation for Hyperparameter Tuning: To prevent the test fold from influencing the hyperparameter tuning, we implemented a nested model selection strategy. Inside each outer fold, we created a separate inner validation set to tune the LightGBM model and apply early stopping. Training was stopped after 100 iterations if the RMSE did not improve. We performed normalization and feature scaling on the inner training data only and then applied them to the validation and test data, ensuring stringent leakage control. The full pipeline, including VAE and LightGBM training, was repeated separately for each of the five outer folds. Figure 3 represents different pre-processing stages for effective feature extraction for the development of improved prediction models.

2.3 VAE based multimodal learning framework

VAEs is generative model employed to extract useful hybrid features that capture complex and nonlinear patterns in high-dimensional gene expression data. The method compresses the input data space into a smooth latent space by effectively reducing noise. This model retains biologically meaningful features for further downstream analysis [29], [30], [31]. In the study, we employed VAEs for efficient feature extraction to predict drug response for brain tumor cell lines. The hybrid feature space is developed using three distinct modalities for gene expression, gene mutation data, and cheminformatics data features like molecular fingerprints of cancer drugs. Each of these distinct variational autoencoders have uniform architecture configuration consisting of an encoder, a feature representation feature space. We have used encoder with three fully connected layers containing 256,128 and 64 neurons respectively. Each layer of encoder followed by a batch normalization and a dropout layer at a rate of 0.2 to resist 20% of neurons for training that help to avoid over fitting. The encoder output is divided into two parallel dense layers with 32 neurons each. One layer generates the mean vector (μ), and the other produces the log-variance vector ($\log \sigma^2$). A latent feature vector (z) is then created using stochastic sampling that enable backpropagation through the network as follows in Eq. (1):

$$z = \mu + \sigma \cdot \Phi, \text{ where, } where\ \Phi \in N(0, I), \text{ and } \sigma = \exp(0.5\log(\sigma^2)) \quad \text{Eq. (1)}$$

This produces an enriched 32D latent space representation z for each input. The decoder mirrored the encoder architecture having three layers with 64, 128, and 256 neurons, respectively. The Batch normalization process and the dropout layers were used in the same way as in the encoder so that the feature space could be normalized that helped to prevent the model over-fitting. The sample latent feature space z was reconstructed for gene expression data. The reconstruction loss was calculated using mean squared error (MSE) as follows in Eq. (2):

$$MSELoss = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2 \quad \text{Eq. (2)}$$

Similarly, for gene 6 binary cross entropy (BCE) as follows in Eq. (3):

$$BCE = -\frac{1}{N}\sum_{i=1}^{N}[y_i\log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)] \quad \text{Eq. (3)}$$

where, N represents the total number of samples, yi denotes the true label for each sample (either 0 or 1), and yi^ indicates the predicted probability that the sample belongs to class 1. This measure indicates how well the predicted probabilities match the actual binary outcomes. To make the reconstructed features more like the original input features, we used Kullback-Leibler (KLD) divergence [32]. This measure assesses how much the reconstructed feature space is different from the original input space. This measure is calculated as follows in Eq. (4):

$$KLD = \frac{-1}{2}\sum_{i=1}^{d}(1 + \log(\sigma^2{}_i) - \mu_i^2 - \sigma_i^2) \quad \text{Eq. (4)}$$

The total training loss for all three VAEs is represented as follows in Eq. (5):

$$L_{VAE} = L_{recon} + \beta \cdot L_{KL} \quad \text{Eq. (5)}$$

where β represents a coefficient of regularization that creates balance between the reconstruction accuracy (L_{recon}) and latent space regularization. After the successful completion of VAE training, the gene expression variational autoencoder (VA_{ge}) has generated a useful 32D gene expression latent vector (LV_{ge}). Similarly, the gene mutation autoencoder (VA_{gm}) has yielded a latent vector for the gene mutation (LV_{gm}) of 32D feature space. Finally, the third autoencoder (VA_{chem}) has generated a latent vector (LV_{chem} = 32D) for the drug cheminformatics feature space. These 32D vectors from each autoencoder were then integrated to form a 96D hybrid feature space for each drug and cell line pair. These features were then used to train and test the LightGBM regressor for that specific fold. The proposed model has effectively utilized this hybrid feature space to accurately predict IC_{50} values.

2.4 VAE based hybrid feature learning

All three variational autoencoders were trained separately, for each outer cross-validation fold, using the inner training and validation splits. The VAEs showed robust and accurate reconstruction performance. For Gene Expression, we found the value of 0.0034 for mean minimum validation loss across all folds. However, the mean Binary Cross Entropy (BCE) losses for Gene Mutation and Drug Cheminformatics were 0.00004316 and 0.00002114, respectively. This indicate effective learning of high-quality binary feature spaces. Figure 4 (A-C) show the loss curve for training and validation data using gene expression, gene mutation, and drug cheminformatics. From these figures, we observe closely convergence between the training and inner validation data. These figures with their exponential decreasing loss trend highlight effective VAE learning in the high-dimensional feature space across all modalities. For gene expression data, we noted a minor gap around 0.05 between training and validation losses. In contrast, the gene mutation and drug cheminformatics losses have

demonstrated stable learning for the binary features. In general, we obtained a convergence around 50 epochs that informed us of the early stopping criterion.

3. LightGBM model development

We used the LightGBM algorithm, based on decision trees, to predict IC_{50} drug response. It can effectively learn complex nonlinear relationships between genes and drugs. The model was trained using hybrid features derived from gene expression, gene mutation, and drug cheminformatics features. During training, LightGBM employed the gradient boosting process. The weak learners were added step by step to reduce the prediction error. We used the RMSE measure as the loss function for continuous IC_{50} values. The integrated hybrid features were efficiently used by the LightGBM regressor to accurately predict the drug response in brain tumor cell lines.

LightGBM employed rectified linear unit (ReLu) activation function for efficient learning of hybrid feature representation that avoid the vanishing gradient problem as well. The LightGBM regressor integrated with VAE tries to improve next prediction ŷ iteratively as follows in Eq. (6):

$$\hat{y}^t = \hat{y}^{(t-1)} + \eta \cdot f_t(x) \quad \text{Eq. (6)}$$

Here, η represents the learning rate, which controls the weight update process. $f_t(x)$ denotes the output of each trained residual. A small learning rate (η = 0.005) was chosen to ensure gradual and stable learning across iterations. We used 5000 boosting rounds that achieved sufficient training depth and model flexibility. Unlike other tree-growing methods, LightGBM grows trees leaf-wise instead of level-wise to improve accuracy. In this way, the leaf nodes are expanded to provide the greatest reduction in the loss function. To manage model complexity, the maximum tree depth was set at 7 and the maximum number of leaves limited to 64. Moreover, to avoid model over-fitting problem, we used optimal number of samples (40) for a split.

We set the random training subsample value at 0.6 and the random feature subset parameter (colsample_bytree) at 0.6. This would help to minimize bias and enhance model generalization. This setting also improved regularization during training. Both L1 and L2 regularization parameters were set to 0.12. Here, ω_i denotes the weight of each leaf, α and λ represent the regularization strengths of L1 and L2, respectively. These penalty parameters (α and λ) regulates the leaf weight and model complexity as follows in Eq. (7):

$$L_{total} = L_{RMSE} + \alpha\sum|\omega_i| + \lambda\sum\omega_i^2 \quad \text{Eq. (7)}$$

We incorporated an early stopping criterion to enhance the efficiency of the training process. This approach continuously observed the RMSE values on the validation dataset. The training process is stopped after 100 iterations if there is no appreciable improvement is obtained. This strategy helped to reduce the computational costs and prevented the model from learning noise or irrelevant data. Figure 5 illustrates the detailed implementation strategy of the LightGBM model development.

4. Results and Discussions

In this section, we explain different statistical analyses to understand the biological and clinical impact of different gene expressions in brain tumors. First, we describe performance analysis of the proposed prediction model with several baseline models and existing models. Their performance is evaluated in terms of RMSE and R^2 metrics. Then, we explain the most accurately predicted drug and their therapeutic impact. From GDSC dataset, we observed the most potent drugs having minimum IC_{50} values. Their IC_{50} values highlight the minimal inhibitory concentration to obtain the maximum sensitivity or resistivity of drug. In the second phase, the statistical gene analysis is carried out to identify the over-expressed and under-expressed genes that has potential to contribute to brain tumor progression and suppression. Finally, we compare the gene expression levels in brain tumor samples with those in normal brain tissues. For clinical perspective, we carried out survival analysis of over/under expression genes with normal brain tissues.

4.1 Comparative Analysis of prediction models

Table 2 highlights the performance of our proposed model with several existing models SVR, RF and MLP, Hybrid CNN, NCFGER, KMBTL and WGRMF. All these models are trained using the same integrated biological and chemoinformatic feature representations. Their predictive performance was evaluated using metrics like RMSE and R^2. The reliable results are obtained using five-folds subject-level stratified cross-validation. Our model has yielded improved RMSE and R^2 values of 1.12 and 0.76, respectively with low standard deviations (SDs) across the folds. Among the existing models, MLP (1.14) and Random Forest (1.15) have achieved relatively lower RMSE values as compared to SVR (1.31), WGRMF (1.37), and Hybrid CNN (1.46). This indicate MLP and Random Forest obtained better prediction accuracy. The proposed VAE-LightGBM model achieved the lowest RMSE value at 1.12. This demonstrates the better performance of our model as compared to all other models in terms of RMSE. The statistical test was carried out using paired t-test, where the proposed model has obtained t-value of 4.27 with p-value 0.003. This highlights that our model prediction performance is statistically significant ($p < 0.05$).

Further, Table 2 compares the performance of our proposed model with several existing models in terms of R^2 measure. The value of the coefficient of determination (R^2) shows how well the model fits the data. A higher R^2 value means better prediction performance. Among the previous existing models, MLP (0.82) and Random Forest (0.81) have performed better than SVR (0.77), Hybrid CNN (0.72), NCFGER (0.53), KMBTL (0.51), and WGRMF (0.41) models. However, our model has achieved the highest R^2 Score of 0.76 show that 76% of IC50 values are predicted accurately. Figure 6 show the scatter plot between actual and predicted IC50 values for LightGBM model with MLP, SVR and RF. This figure shows the improved performance of our model with baseline MLP, SVR and RF. This indicate the improved performance of our model in terms of R^2 as well. Further, our model has obtained t-value 4.36 with p-value of 0.025, that show significant improvement with our model. This means, overall, the proposed model predicts more accurately than the other predictors. This is due to the integration of useful VAE features employed for LightGBM model development that has enhanced the model prediction accuracy.

4.2: Biological implication of the most accurately predicted drugs

Using proposed regressor model, we identified most accurately predicted IC_{50} values of drugs such as Erlotinib, Linifanib, Pictilisib, and Acetyleshikonin. In further analysis, we identified five drugs FDA approved drugs from the testing GDSC dataset. Our model showed the best prediction of IC_{50} values for Midostaurin, GW-2580, FMK, Imatinib, and A-443654. The model has achieved the lowest RMSE values of 0.186, 0.277, 0.322, 0.330, and 0.332 for these drugs.

Table 3 presents several FDA approved drugs with their biological activity in brain tumor and their important therapeutic effects. For these drugs, our predictor effectively captured the complex relationship between the hybrid features from the VAE and their target bioactivity. The VAE-based hybrid features have captured the most relevant patterns that fit well with the data distribution. This is since our model employed the most relevant hybrid features to better fit with the underlying data distribution. The improved performance metrics show that our model can be useful for predicting new drugs with unknown bioactivities. The accurate prediction of the proposed model would help to establish improved treatment strategies.

Furthermore, by analyzing GDSC dataset based on the lowest IC_{50} values, we have identified only five most potent FDA-approved drugs such as Midostaurin, GW-2580, FMK, Imatinib, and A-443654. Their IC_{50} values highlight the minimal inhibitory concentration to obtain the maximum sensitivity or resistivity of drug.

4.3 Statistical Gene Expression Analysis

We carried out statistical gene analysis to identify the over-expressed and under-expressed genes for GDSC dataset using volcano plot as shown in figure 7. We selected several abnormal genes and explained their biological impact in brain tumor progression and suppression using recent literature. The GDSC database used cell lines to provide genomic drug response and gene-cell line associations with differential expressions. The clinical perspective of these highly expressed genes (HEG) is analyzed. Further statistical analysis is carried out using genomic TCGA dataset for HEG. This genomic data resource contains additional information related to brain tissue across normal and tumor cell lines, which are missing in the GDSC database. Then, the statistical significance of HEGs across

normal genes is carried out. The levels of HEG vary significantly across normal and brain tumor cell lines. This analysis would provide useful information for disease diagnosis. We carried out survival analysis using UALCAN [33] webserver to understate their impact on the patient survival.

4.3.1 Gene Expression Analysis: specific to brain tumor

The gene expression levels in brain tumor cells is analyzed using transcripts per million (TPM) values. Those genes with TPM values above 10 are considered highly active. On the other hand, genes with TPM values 1–10 show moderate activity and genes with less than 1 TPM are considered weakly active or inactive.

We have employed volcano plot to show the significantly expressed genes in the brain tumor GDSC dataset. X-axis represents $\log_2$ fold change that shows how much gene expression level increases or decreases between tumor and normal samples. However, on the other hand the Y-axis indicates $-\log_{10}$ (p-value) with the statistical significance of the gene expression. Those Genes with higher values on Y-axis are more significant. However, on the X-axis, genes away from-2, and 2 indicate high up-regulation / down-regulation. The red and green dots indicate the over and under expressed genes with a p-value less than 0.05. Further analyzing the graph, we obtained the over-expressed genes: EGFR, MKI67, BIRC5, TOP2A, *and* AURKA. *These genes are potentially* linked to tumor growth, cell division, and survival. We found the under-expressed genes: GFAP, MBP, NEFL, SLC1A2, *and* PLP1. *They* are normally active/normal expressions in healthy brain tissue but reduced their expression level in tumor samples. The low expression level of these genes may lead to the loss of normal brain cell functions. These expression patterns can help identify potential biomarkers to further guide targeted therapies for brain tumor patients. It is important to understand their biological roles and therapeutic implications. Table 4 shows a detailed information about these over expressed and under expressed genes for GBM, also highlighting their biological functionality, therapeutic implications and biological insights.

4.3.2 Differential Gene Expression (DGE) Analysis: Specific to brain tumor

The volcano plots in figure 7 highlight the tendencies of HEG gene. This figure helped to identify significantly up/down regulated genes in tumors, highlighting their potential role in cancer progression. However, we have selected only five sample genes (EGFR, MKI67, BIRC5, TOP2A, and AURKA) to investigate their differential expression with respect to their statistical perspective in tumor tissues comparison to normal brain samples.

Figure 8 in the form of boxplots were made in the UALCAN portal using TCGA data. The figure demonstrates behavior of five HEG genes of EGFR, MKI67, BIRC5, TOP2A, and AURKA in GBM samples for TCGA dataset. Overall, this figure shows the significantly higher transcript levels in primary tumor tissues compared to normal sample. The EGFR gene showed the highest expression among all other genes with a median transcript per million (TPM) value 7.195 in tumor samples versus 13.029 in normal tissue.
However, in GBM gene, this value increases up to 44.225 TPM. On the other hand, the MKI67 gene has exhibited higher expression in tumors with median value 0.519 TPM as compared to normal sample with median 0.148 TPM). This genre is known as proliferation marker that indicates their active role in cell division. Similarly, the BIRC5 gene has obtained median value of 0.359 TPM against 0.181 TPM in normal tissues. This show their upregulation pattern with tumor samples that reflect their crucial role in inhibiting apoptosis. Further, extending the analysis for TOP2A gene, we observed higher median of tumor tissue 0.954 TPM versus 0.364 TPM in normal. This gene function is associated with DNA replication and further enhanced proliferative activity. Lastly, AURKA gene has demonstrated higher expression level in tumors with median value of 7.129 TPM as compared to normal median 1.973 TPM. This gene plays important role in mitotic regulation. From this, we conclude all five HEG genes has revealed overexpression in GBM primary tumors. This highlight their potential biomarkers and possible therapeutic targets in glioblastoma progression.

4.4 Survival Analysis

We have performed survival analysis of over expressed genes EGFR, MKI67, BIRC5, TOP2A, and AURKA. For this purpose, we have employed Kaplan-Meier (KM) survival curves to examine survival rates in patient suffering from GBM. This curve is used to illustrate the survival probabilities of tumor vs normal group over time. This analysis helps the timely diagnosis of tumor to improve the patient survival rate by determining their levels of severity. Figure 9 represent survival analysis through Kaplan-Meier curves for over expressed genes. We want to perform survival analysis for the over-expressed genes EGFR, MKI67, BIRC5, TOP2A, and AURKA. For this

purpose, we have used KM survival curves for patients with glioblastoma multiforme (GBM). These curves show the survival probability of patients in the tumor group compared with the normal group over time. This analysis would help to understand how gene expression affects patient outcomes. This analysis would be useful for early diagnosis and to determine the cancer severity. Figure 9 represent the KM curves for these genes. The KM results showed that none of the over expressed genes were statistically significant with patient survival. For gene EGFR with p value 0.33 has small increase in survival probability.

This gene possesses medium survival period, but it has no statistical significance with patient survival. Similarly, both MKI67 and BIRC5 genes have exhibited survival probability values of 0.76 and 0.63, respectively. Therefore, these genes do not provide meaningful stratification between high and low gene expression groups. Although, these genes play crucial role in cell proliferation and inhibit apoptosis. But their individual expression levels have not shown direct influence on patient survival in GBM. Similarly, TOP2A genes with p value 0.67 and AURKA with p value 0.59 have shown no significant difference in survival outcome. This means that these genes are not strong markers to predict survival in GBM patients. However, their biological role is significant in DNA copying and cell division.

These genes keep their important biological roles in DNA replication and cell division. But there are not strong independent prognostic markers in GBM tumor prediction. Overall, these findings highlight that over-expressed genes such as EGFR, MKI67, BIRC5, TOP2A, and AURKA contribute to tumor growth but individually these genes may not predict patient prognosis. It is anticipated that multi-gene expression panels and their molecular subtyping could provide accurate and reliable predictions of survival outcomes in brain tumor patients.

5. Conclusion

In this study, we proposed a multi-modal variational autoencoder based framework that integrates genomic and chemoinformatic features to accurately predict drug response in brain tumor cell lines. This framework utilizes three modality-specific VAEs to encode gene expression, gene mutation, and drug molecular fingerprint information. The latent representations generated by these VAEs are fused and then processed using LightGBM regressor that enable accurate prediction of IC_{50} values. This useful integration has preserved important features to enhance the model prediction. The model demonstrated improved performance compared to several prediction models such as Support Vector Regression, Random Forest, Multi-Layer Perceptron, KMBTL, and WGRMF.

Through the analysis of the GDSC dataset, we have identified Midostaurin, GW-2580, FMK, Imatinib, and A-443654 as the most potent FDA-approved drugs for brain tumor treatment. These drugs have the minimum RMSE values 0.186, 0.277, 0.322, 0.330 and 0.332, respectively.

Furthermore, the statistical analysis of Glioblastoma (GBM) cell lines is carried out. We found several over-expressed genes: EGFR, MKI67, BIRC5, TOP2A, AURKA and under-expressed genes: GFAP, MBP, NEFL, SLC1A2, PLP1. Their abnormal expression levels have potential roles in tumor progression and suppression. Furthermore, we highlight the impact of over- and under-expressed genes on patient survival using Kaplan-Meier analysis. This analysis is based on comparisons with normal brain tissues. For clinical perspective, we have carried out the survival analysis of highly expressed genes with normal tissue samples. This analysis revealed that most expressed genes of tumor samples have no statistical significance ($p > 0.05$) with normal patient survival. In summary, the proposed silico study provided insights into how gene expression patterns contribute to brain tumors that help in identifying potential biomarkers and therapeutic targets.

6. References

[1] Mondia, M.W.L., Espiritu, A.I., and Jamora, R.D.G., "Primary brain tumor research productivity in Southeast Asia and its association with socioeconomic determinants and burden of disease", Front. Oncol., 10, p. 607777 (2020). doi:10.3389/fonc.2020.607777.

[2] Perry, A. and Wesseling, P., "Histologic classification of gliomas", in Handbook of Clinical Neurology, Vol. 134, pp. 71–95, Elsevier (2016). doi:10.1016/B978-0-12-802997-8.00005-0.

[3] Norollahi, S.E., "Role of the TLR signaling pathway in the pathogenesis of glioblastoma multiforme with an emphasis on immunotherapy", Biochem. Biophys. Rep., p. 102149 (2025). doi:10.1016/j.bbrep.2025.102149.

[4] Bal, R. and Mutton, L., "Interdisciplinary review of the qualities of glioblastoma multiforme", Sciential – McMaster Undergrad. Sci. J., 1(9), pp. 45–54 (2022). doi:10.15173/sciential.v1i9.3201.

[5] Gkountas, A.A., Polychronopoulos, N.D., Sofiadis, G.N., et al., "Simulation of magnetic nanoparticles crossing through a simplified blood–brain barrier model for glioblastoma multiforme treatment", Comput. Methods Programs Biomed., 212, p. 106477 (2021). doi:10.1016/j.cmpb.2021.106477.

[6] Cencioni, C., Scagnoli, F., Spallotta, F., et al., "The 'superoncogene' MYC at the crossroad between metabolism and gene expression in glioblastoma multiforme", Int. J. Mol. Sci., 24, p. 44217 (2023). doi:10.3390/ijms24044217.

[7] Grochans, S.S., "Epidemiology of glioblastoma multiforme: Literature review", Cancers, 14, p. 2412 (2022). doi:10.3390/cancers14102412.

[8] Asiri, A.A., "Brain tumor detection and classification using fine-tuned CNN with ResNet50 and U-Net model", Life, 13, p. 1449 (2023). doi:10.3390/life13071449.

[9] D'Arrigo, G., Leonardis, D., Abd Elhafeez, S., et al., "Methods to analyse time-to-event data: The Kaplan–Meier survival curve", J. Oncol., 2021, p. 2290120 (2021). doi:10.1155/2021/2290120.

[10] Cavalcanti, I.D.L. and Soares, J.C.S., "Conventional cancer treatment", in Advances in Cancer Treatment, pp. 29–56, Springer (2021). doi:10.1007/978-3-030-68334-4_4.

[11] Kaur, R., Bhardwaj, A., Gupta, S., et al., "Cancer treatment therapies: Traditional to modern approaches to combat cancers", Mol. Biol. Rep. (2023). doi:10.1007/s11033-023-08809-3.

[12] Ha, S., Park, J., Jo, K., et al., "Comparative analysis of regression algorithms for drug response prediction using the GDSC dataset", BMC Res. Notes, 18 (2025). doi:10.1186/s13104-024-07026-w.

[13] Sachdev, P., Ronen, R., Dutkowski, J., et al., "Systematic analysis of genetic and pathway determinants of eribulin sensitivity across cancer cell lines", Cancers, 14, p. 4532 (2022). doi:10.3390/cancers14184532.

[14] Ganini, C., "Global mapping of cancers: The Cancer Genome Atlas and beyond", Mol. Oncol. (2021). doi:10.1002/1878-0261.13056.

[15] Zdrazil, B., Felix, A., Hunter, A., et al., "The ChEMBL database in 2023", Nucleic Acids Res., 52, pp. D1180–D1192 (2024). doi:10.1093/nar/gkad1004.

[16] Knox, C., Law, M., Jewison, T., et al., "DrugBank 6.0: The DrugBank knowledgebase for 2024", Nucleic Acids Res., 52, pp. D1265–D1275 (2024). doi:10.1093/nar/gkad976.

[17] Kim, S., Thiessen, P.A., Bolton, E.E., et al., "PubChem 2025 update", Nucleic Acids Res., 53, pp. D1516–D1525 (2025). doi:10.1093/nar/gkae1059.

[18] Robert, B.M., Brindha, G.R., Santhi, B., et al., "Computational models for predicting anticancer drug efficacy", Comput. Methods Programs Biomed., 178, pp. 105–112 (2019). doi:10.1016/j.cmpb.2019.06.011.

[19] Wang, B., He, Y., Du, X., et al., "VAE-GANMDA: A microbe–drug association prediction model", Artif. Intell. Med. (2025). doi:10.1016/j.artmed.2025.103198.

[20] Zaidi, A.R., Bilal, M., Majid, T., et al., "Developing anticancer drug response system using deep learning with hybrid genomic and chemical features", Iran. J. Sci. Technol., Trans. Electr. Eng. (2024). doi:10.1007/s40998-024-00765-3.

[21] Kingma, D.P. and Welling, M., "Auto-encoding variational Bayes", in Proc. 2nd Int. Conf. Learn. Representations (ICLR 2014) (2014). Available: arXiv:1312.6114.

[22] Liao, X. and Yu, B., "DvaDRP: A dual variational autoencoder based on big data for drug response prediction", in ACM Int. Conf. Proc. Ser., pp. 78–85, Assoc. Comput. Mach. (2024). doi:10.1145/3686540.3686551.

[23] Bouchacourt, D., Tomioka, R. and Nowozin, S., "Multi-level variational autoencoder: Learning disentangled representations from grouped observations", in Proc. AAAI Conf. Artif. Intell. (AAAI 2018), pp. 2095–2102, AAAI Press (2018). doi:10.1609/aaai.v32i1.11867.

[24] Gönen, M. and Margolin, A.A., "Drug susceptibility prediction against a panel of drugs using kernelized Bayesian multitask learning", Bioinformatics, 30(17), pp. i556–i563 (2014). doi:10.1093/bioinformatics/btu464.

[25] Liu, J.X., Cui, Z., Gao, Y.L., et al., "WGRCMF: A weighted graph regularized collaborative matrix factorization method for predicting novel lncRNA–disease associations", IEEE J. Biomed. Health Inform., 25, pp. 257–265 (2021). doi:10.1109/JBHI.2020.2985703.

[26] Rajan, K., Zielesny, A. and Steinbeck, C., "STOUT V2.0: SMILES to IUPAC name conversion using transformer models", J. Cheminform., 16 (2024). doi:10.1186/s13321-024-00941-x.

[27] Wu, T., "Methodological roadmap for machine learning in deep eutectic solvent research: A framework driven review and perspective", Ind. Eng. Chem. Res., 64 (2025). doi:10.1021/acs.iecr.5c01985.

[28] Park, H.J., Moon, E.K., Yoon, J.Y., Oh, C.M., Jung, K.W., Shin, H.Y. and Park, B.K., "Incidence and survival of childhood cancer in Korea", Cancer Res. Treat., 48(3), pp. 869–882 (2016). doi:10.4143/crt.2015.290.

[29] Mohammadzadeh Vardin, T., Ghareyazi, A., Gharizadeh, A., et al., "DeepDRA: Drug repurposing using multi-omics data integration with autoencoders", PLoS One, 19 (2024). doi:10.1371/journal.pone.0307649.

[30] Adam, G., Rampášek, L., Safikhani, Z., et al., "Machine learning approaches to drug response prediction: Challenges and recent progress", npj Precis. Oncol., 4, p. 19 (2020). doi:10.1038/s41698-020-0122-1.

[31] Singh, D.P. and Kaushik, B., "A systematic literature review for the prediction of anticancer drug response using various machine learning and deep learning techniques", Chem. Biol. Drug Des., 101(1), pp. 175–194 (2023). doi:10.1111/cbdd.14164.

[32] Spineli, L.M., "Local inconsistency detection using the Kullback–Leibler divergence measure", Syst. Rev., 13 (2024). doi:10.1186/s13643-024-02680-4.

[33] Chandrashekar, D.S., "UALCAN: An update to the integrated cancer data analysis platform", Neoplasia, 25, pp. 18–27 (2022). doi:10.1016/j.neo.2022.01.001.

[34] Ilic, I. and Ilic, M., "International patterns and trends in brain cancer incidence and mortality: An observational study based on the global burden of disease", Heliyon, 9, p. e18222 (2023). doi:10.1016/j.heliyon.2023.e18222.

[35] Yashavantha Rao, H.C., Sruthi, D., Kamalraj, S., et al., "Endophytic fungi as a potential source of cytotoxic drugs: A fungal solution to cancer", in Volatiles and Metabolites of Microbes, pp. 305–323, Elsevier (2021). doi:10.1016/B978-0-12-824523-1.00015-8.

[36] Bhattacharya, S., Shinde, P., Page, A., et al., "PLGA–TPGS hybrid nanoparticles of imatinib mesylate for the treatment of glioblastoma multiforme", Curr. Med. Chem., 32, pp. 8350–8370 (2025). doi:10.2174/0109298673323270241118103546.

[37] Irfan, N., Balasubramaniyan, S., Ali, D.M., et al., "Bioisosteric replacements of tyrosine kinase inhibitors to develop potent and safe chemotherapy agents", J. Biomol. Struct. Dyn., 41, pp. 9437–9447 (2023). doi:10.1080/07391102.2022.2146751.

[38] Xie, M., Lei, X., Zhong, J., Ouyang, J. and Li, G., "Drug response prediction using graph representation learning and Laplacian feature selection", BMC Bioinformatics, 23(Suppl. 8), p. 532 (2022). doi:10.1186/s12859-022-05080-4.

[39] Li, X., "Structure of POU2AF1 recombinant protein and its role in liver cancer progression based on WGCNA and molecular docking analysis", Int. J. Biol. Macromol., 278 (2024). doi:10.1016/j.ijbiomac.2024.134629.

[40] Alzahrani, S.M., Al Doghaither, H.A., Al Ghafari, A.B., et al., "5-Fluorouracil and capecitabine therapies for the treatment of colorectal cancer (review)", Oncol. Rep. (2023). doi:10.3892/or.2023.8612.

[41] Tchounwou, P.B., Dasari, S., Noubissi, F.K., et al., "Advances in our understanding of the molecular mechanisms of action of cisplatin in cancer therapy", J. Exp. Pharmacol. (2021). doi:10.2147/JEP.S267383.

[42] Castillo Ordoñez, W.O., Aristizabal Pachon, A.F., Alves, L.B., et al., "Epigenetic regulation exerted by Caliphruria subedentata and galantamine: An in vitro and in silico approach for mimicking Alzheimer's disease", J. Biomol. Struct. Dyn., 42, pp. 11215–11230 (2024). doi:10.1080/07391102.2023.2261034.

[43] Li, X.P., Guo, Z.Q., Wang, B.F., et al., "EGFR alterations in glioblastoma play a role in antitumor immunity regulation", Front. Oncol., 13, p. 1236246 (2023). doi:10.3389/fonc.2023.1236246.

[44] Andrés Sánchez, N., Fisher, D. and Krasinska, L., "Physiological functions and roles in cancer of the proliferation marker Ki-67", J. Cell Sci. (2022). doi:10.1242/jcs.258932.

[45] Zuo, Z., Wang, P., Chen, X., et al., "SWnet: A deep learning model for drug response prediction from cancer genomic signatures and compound chemical structures", BMC Bioinformatics, 22, p. 434 (2021). doi:10.1186/s12859-021-04352-9.

[46] Pang, Y., "ATRX cooperates with TOP2B for replication fork stability and DNA damage response through G-quadruplex regulation", Nucleic Acids Res., 53 (2025). doi:10.1093/nar/gkaf939.

[47] Stefani, A., "Unweaving the mitotic spindle: A focus on Aurora kinase inhibitors in lung cancer", Front. Oncol. (2022). doi:10.3389/fonc.2022.1026020.

[48] Grube, S., Freitag, D., Kalff, R., et al., "Characterization of adherent primary cell lines from fresh human glioblastoma tissue defining glial fibrillary acidic protein as a reliable marker in glioblastoma cell culture establishment", Cancer Rep., 4 (2021). doi:10.1002/cnr2.1324.

[49] Cardona, H.J., Somasundaram, A., Crabtree, D.M., et al., "Prenatal overexpression of platelet-derived growth factor receptor A results in central nervous system hypomyelination", Brain Behav., 11 (2021). doi:10.1002/brb3.2332.

[50] Moriarty, C., Gupta, N. and Bhattacharya, D., "Role of glutamate excitotoxicity in glioblastoma growth and its implications in treatment", Cell Biol. Int., 49(5), pp. 421–434 (2025). doi:10.1002/cbin.70005.

[51] Magi, S., Piccirillo, S., Amoroso, S. and Lariccia, V., "Excitatory amino acid transporters (EAATs): glutamate transport and beyond", Int. J. Mol. Sci., 20(22), p. 5674 (2019). doi:10.3390/ijms20225674.

[52] Masciocchi, S., "Conformational antibodies to proteolipid protein 1 and its peripheral isoform DM20 in patients with CNS autoimmune demyelinating disorders", Neurol. Neuroimmunol. Neuroinflamm., 12(2), p. e200359 (2025). doi:10.1212/NXI.0000000000200359.

List of Figures

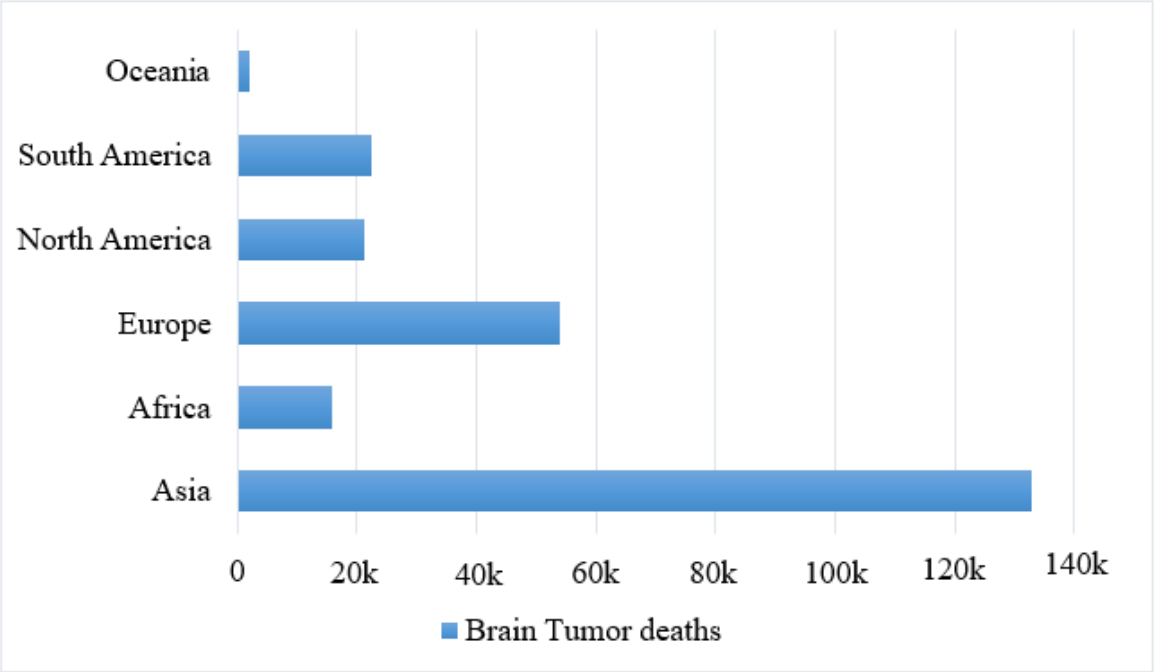


Fig.1 Brain tumor mortality rate across continents

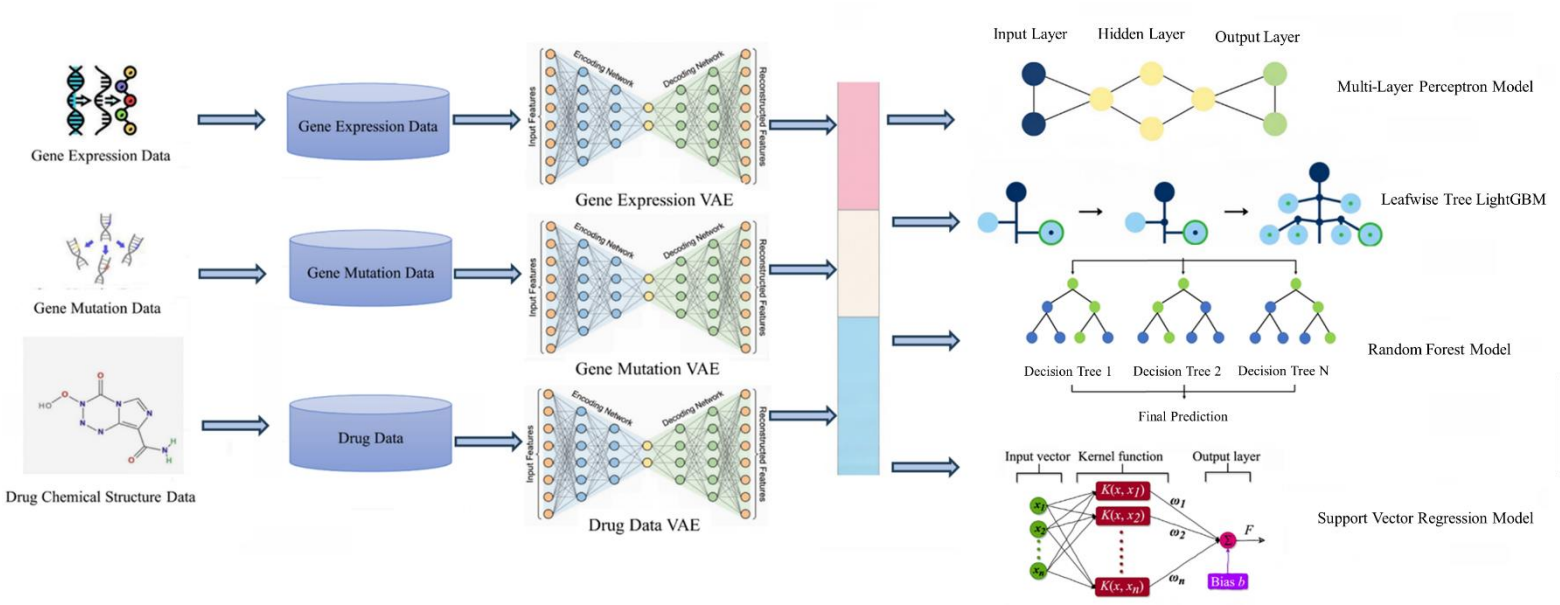


Fig. 2 Proposed framework for VAE based LightGBM prediction model development

Data Preprocessing Pipeline			
Pipeline stage	Normalization	Feature Engineering	Chemical Structure
Data Preprocessing	Z-score normalization	Gene mutation binarization	SMILES to ECFP4 conversion
Stratified K-fold cross validation	Stratification based on tumor subtype	Five folds with unique cell lines	Data points assigned to same fold
Nested cross validation	Normalization fitted on inner training data	Hyperparameter optimization (LightGBM)	Early stopping during training

Fig. 3: Different data pre-processing stages for hybrid feature extraction

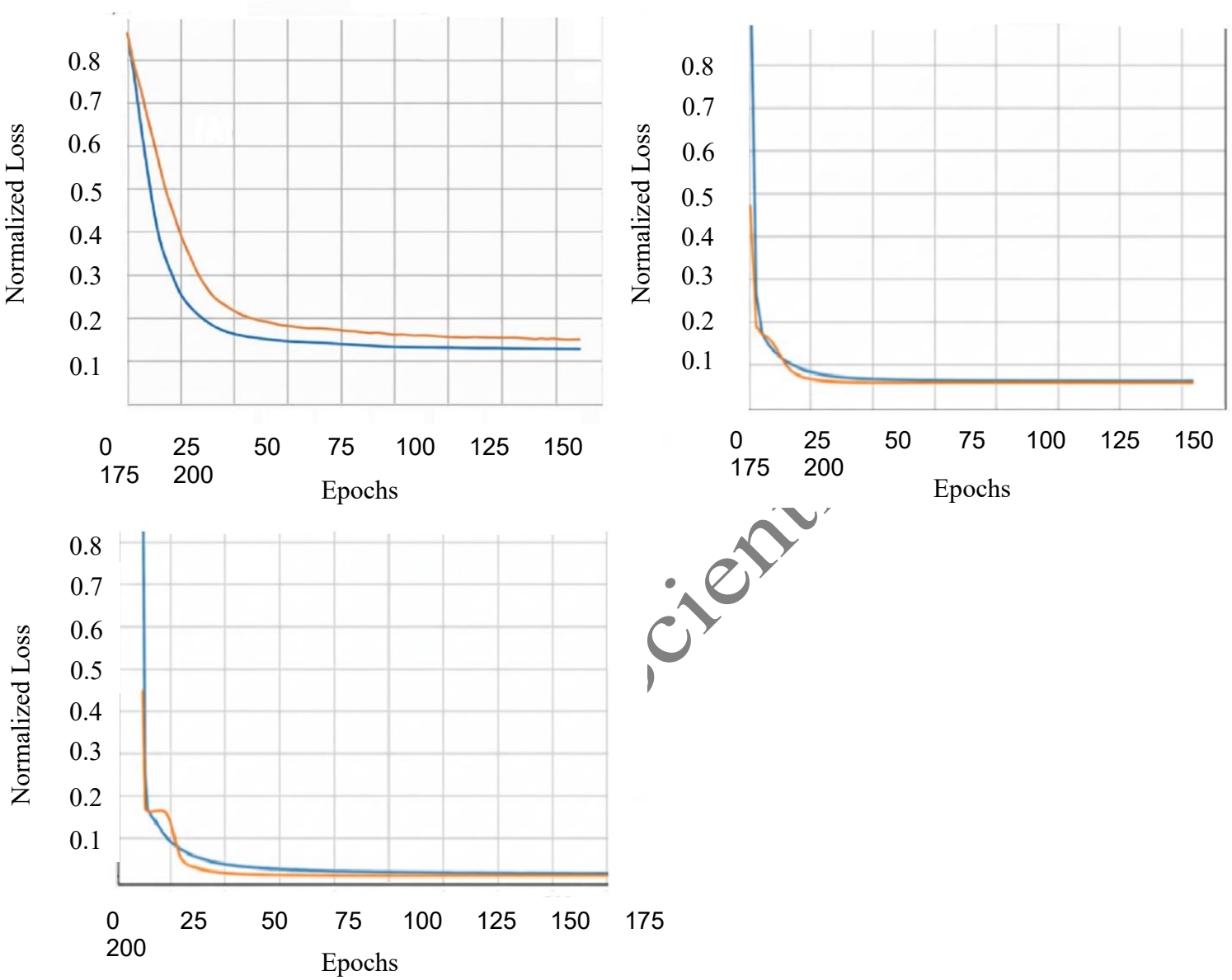


Fig. 4: Represents the loss curve for training (orange) and validation (blue) data using gene expression (top left), gene mutation (top right), and drug cheminformatics (bottom left), respectively

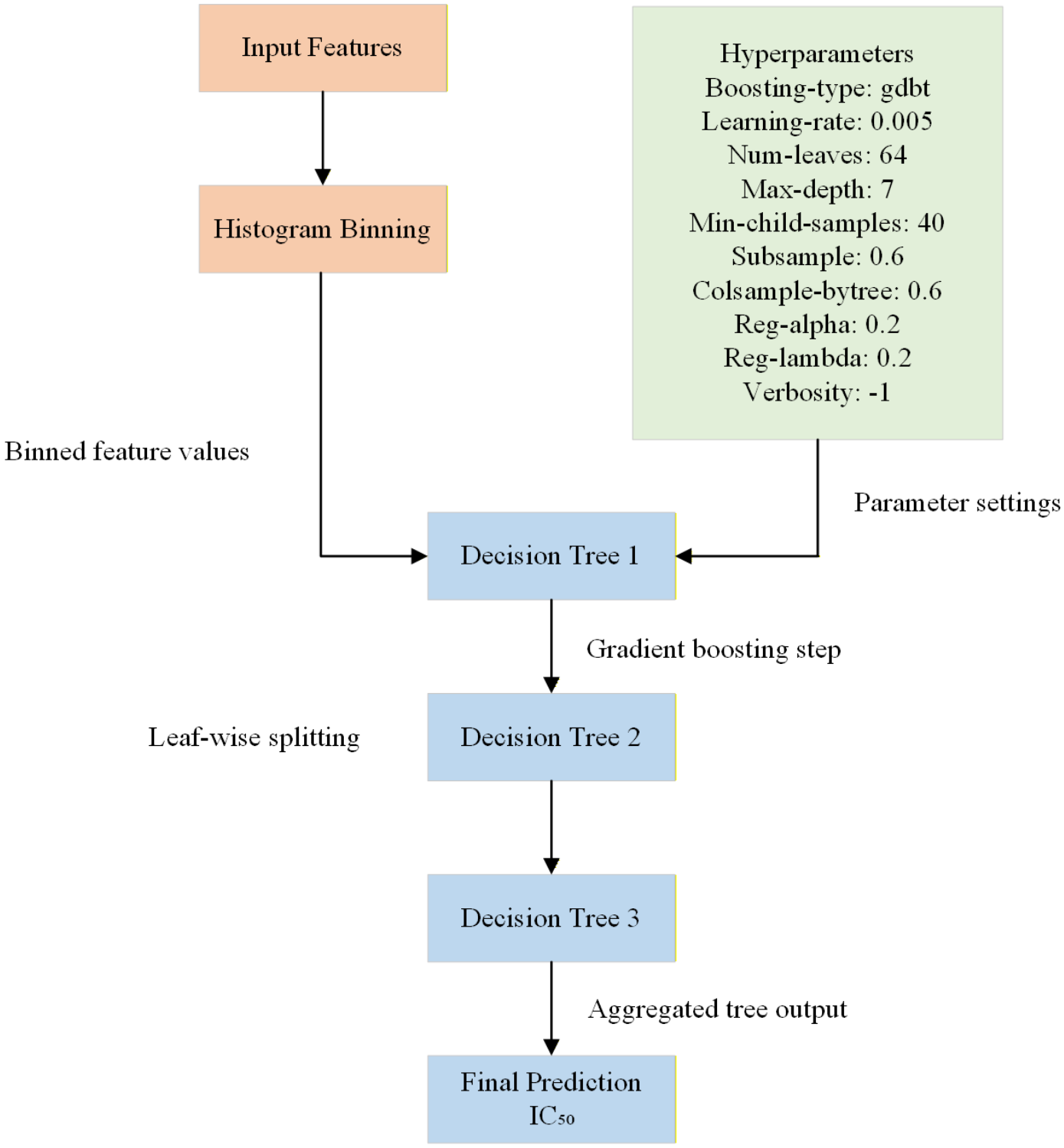


Fig. 5: Parameter setting of VAEs for the development of the proposed model

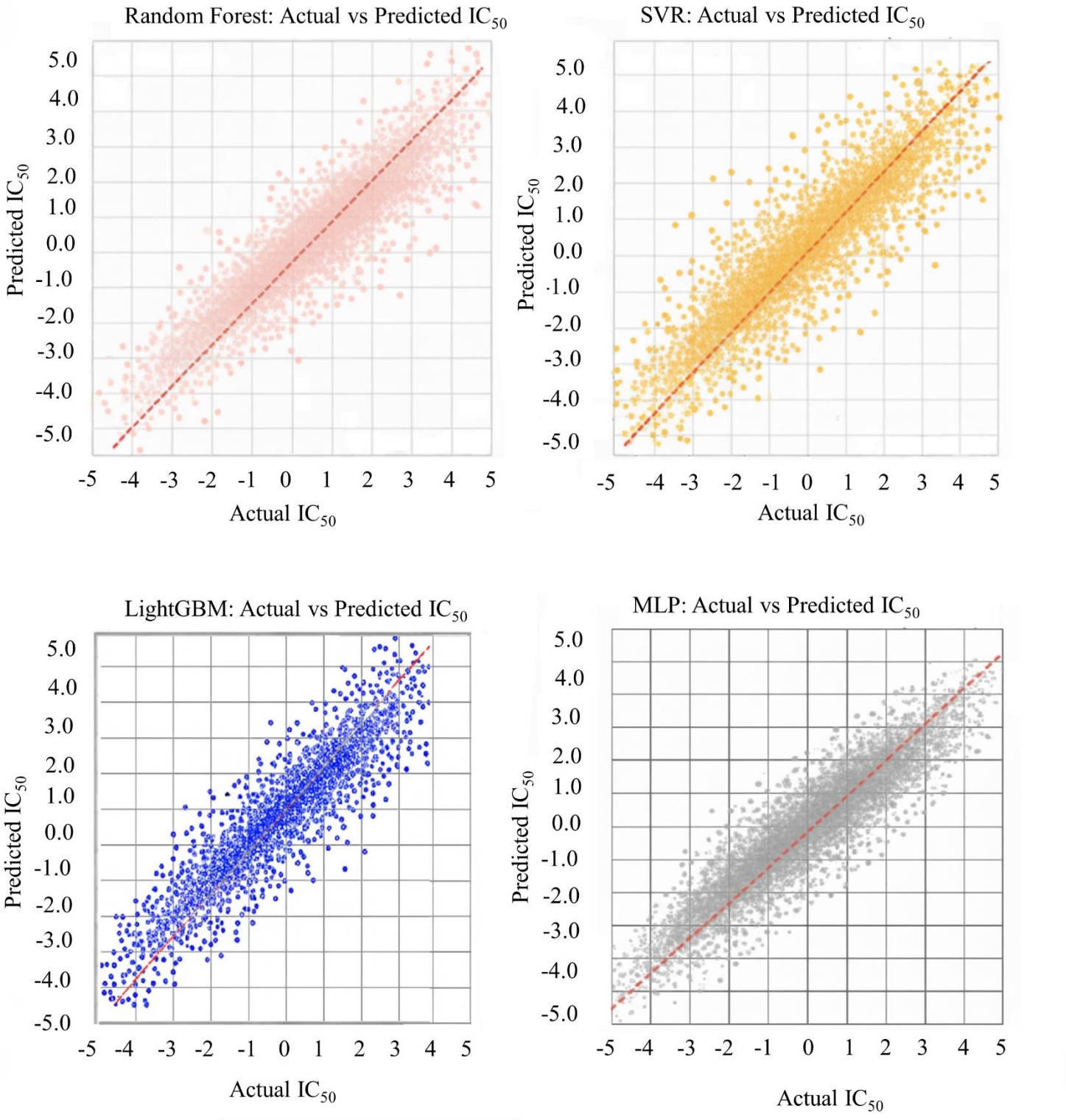


Fig.6: Scatter plot for actual and predicted IC_{50} values for LightGBM model with MLP, SVR and RF comparison models

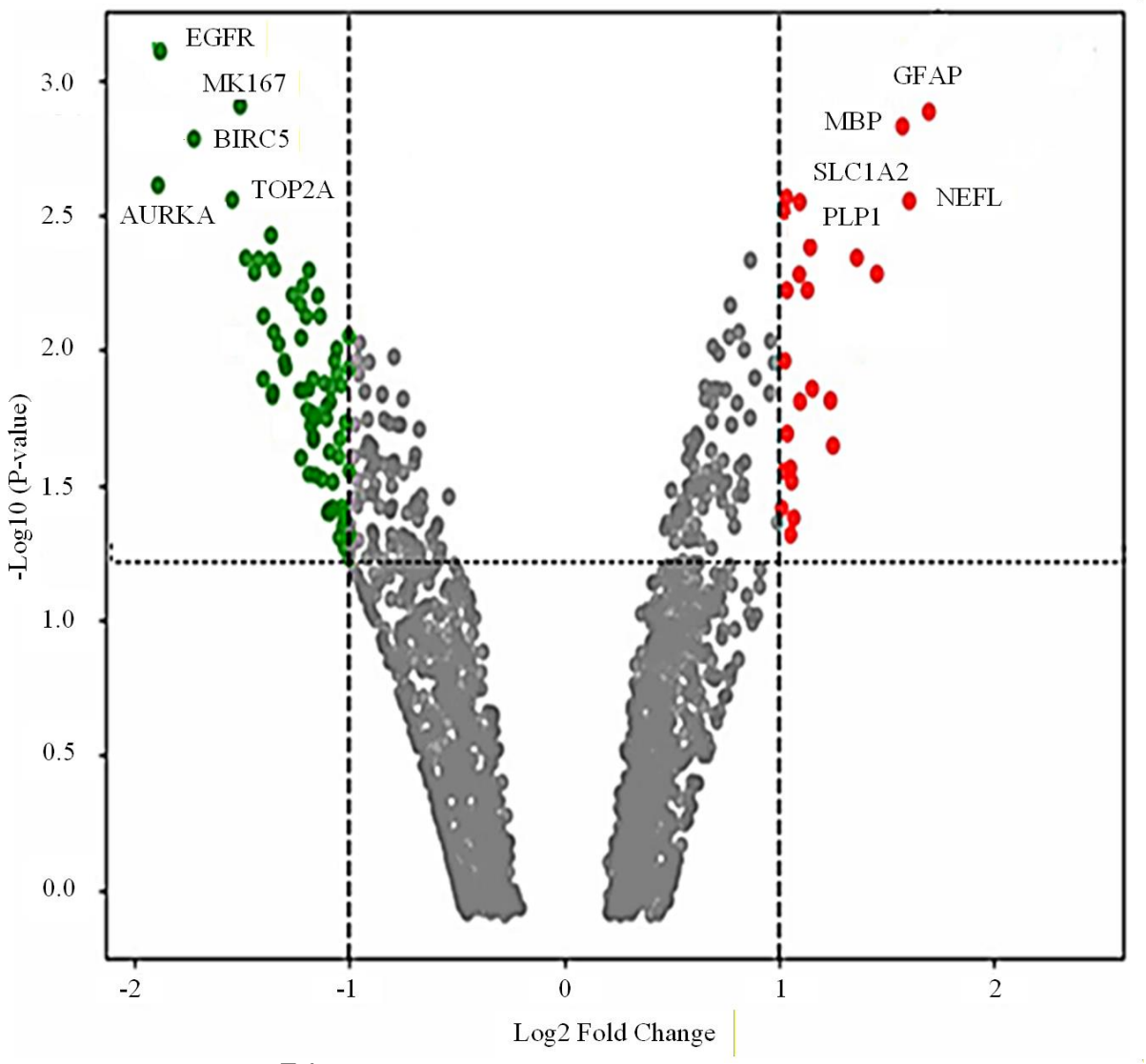


Fig.7: The Volcano plot highlight the significant (over/under) expressed genes with $p < 0.05$

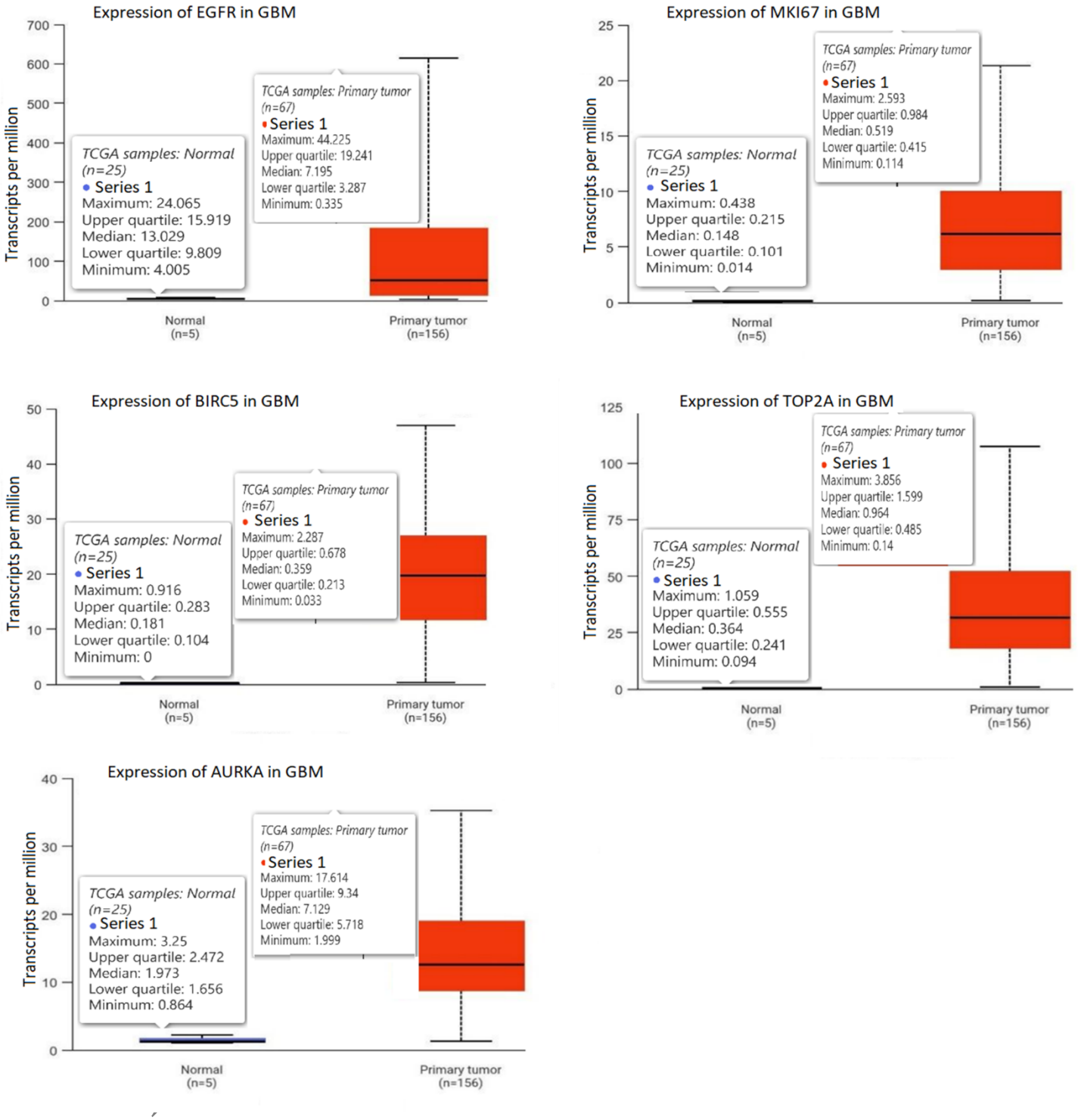


Fig. 8: Over expressed genes EFGR, MK167, BIRC5, TOP2A and AURKA in GBM Cell Lines

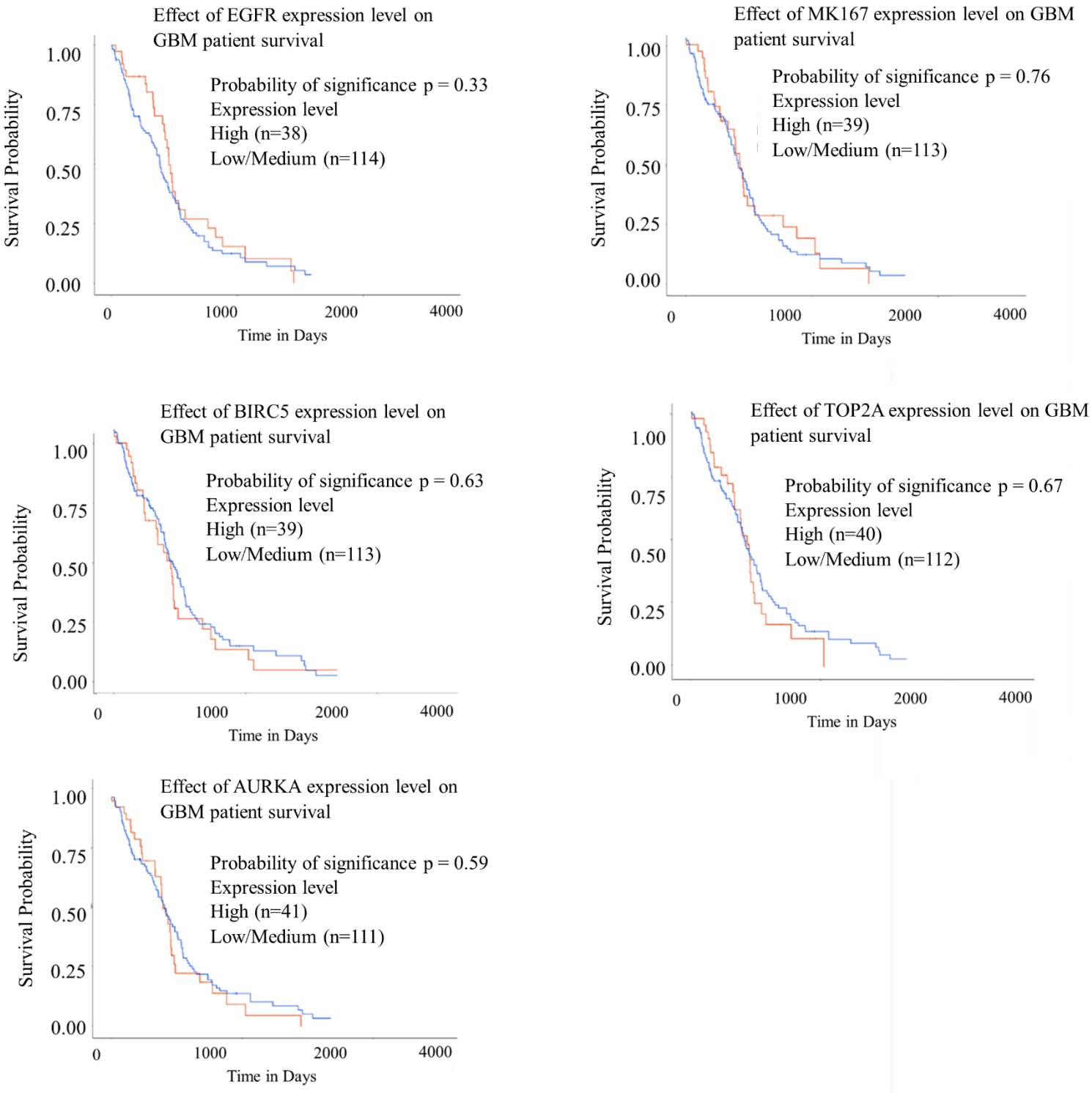
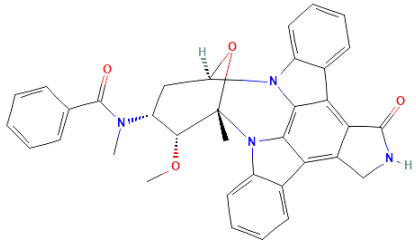
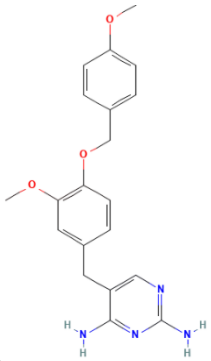
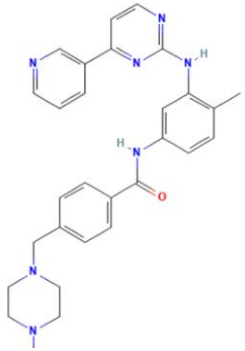
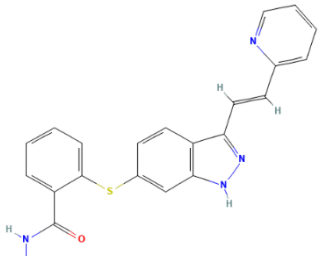
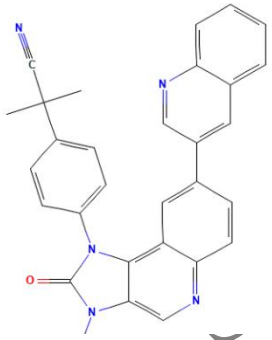
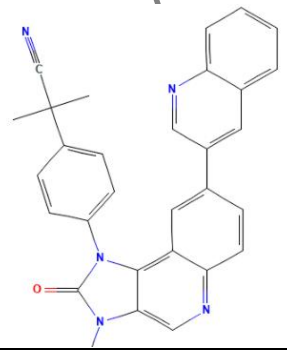


Fig. 9: Survival curve of over expressed genes: EGFR, MKI67, BIRC5, TOP2A, and AURKA

List of Tables:

Table 1: The detailed information of anticancer drugs retrieved from the PubChem dataset

Drug Name & Compound ID	Molecular Formula	Drug 2-D Structure	Isomeric SMILES Representation
Midostaurin 9829523 [35]	$C_{35}H_{30}N_4O_4$		C[C@@]12[C@@H]([C@@H](C[C@@H](O1)N3C4=CC=CC=C4C5=C6C(=C7C8=CC=CC=C8N2C7=C53)CNC6=O)N(C)C(=O)C9=CC=CC=C9)OC
GW-2580 11617559 [28]	$C_{20}H_{22}N_4O_3$		COC1=CC=C(C=C1)COC2=C(C=C(C=C2)CC3=CN=C(N=C3N)N)OC
Imatinib 5291 [36]	$C_{29}H_{31}N_7O$		CC1=C(C=C(C=C1)NC(=O)C2=CC=C(C=C2)CN3CCN(CC3)C)NC4=NC=CC(=N4)C5=CN=CC=C5

Axitinib 6450551 [37]	$C_{22}H_{18}N_4OS$		CNC(=O)C1=CC=CC=C1SC2=CC3=C(C=C2)C(=NN3)/C=C/C4=CC=CC=N4
Dactolisib 11977753 [38]	$C_{30}H_{23}N_5O$		CC(C)(C#N)C1=CC=C(C=C1)N2C3=C4C=C(C=CC4=NC=C3N(C2=O)C)C5=CC6=CC=CC=C6N=C5
(5Z)-7-Oxozeaenol 9863776 [39]	$C_{19}H_{22}O_7$		C[C@H]1C/C=C\C(=O)[C@H]([C@H](C/C=C/C2=C(C(=CC(=C2)OC)O)C(=O)O1)O)O

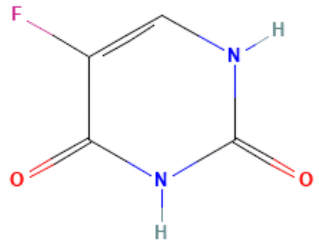
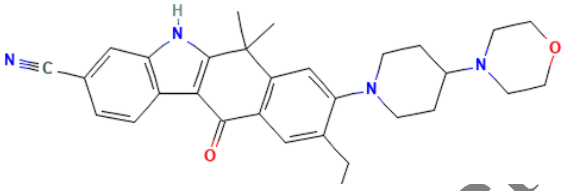
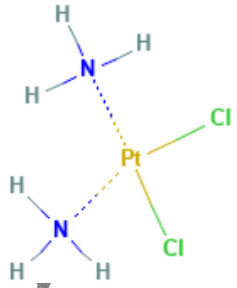
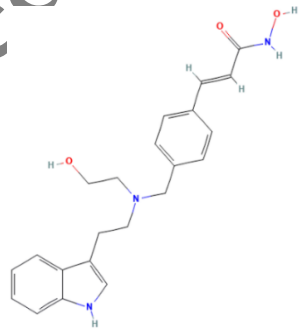
5-Fluorouracil 3385 [40]	$C_4H_3FN_2O_2$		C1=C(C(=O)NC(=O)N1)F
Alectinib 49806720	$C_{30}H_{34}N_4O_2$		CCC1=CC2=C(C=C1N3CCC(CC3)N4CCOCC4)C(C5=C(C2=O)C6=C(N5)C=C(C=C6)C#N)(C)C
Cisplatin 5460033 [41]	$Cl_2H_6N_2Pt$		N.N.Cl[Pt]Cl
Dacinostat 6445533 [42]	$C_{22}H_{25}N_3O_3$		C1=CC=C2C(=C1)C(=CN2)CCN(CCO)CC3=CC=C(C=C3)/C=C/C(=O)NO

Table 2: Comparative performance of proposed models with the other models

	Model	**RMSE**	**R^2 Score**
Existing	Hybrid CNN [34]	1.46	0.723
	NCFGER [35]	1.18	0.532
	WGRMF	1.37	0.416
	KMBTL [36]	1.26	0.511
Baseline			
	SVR	1.31	0.621
	Random Forest	1.25	0.714
	MLP	1.24	0.701
Proposed Model	VAE-LightGBM	1.12 (t = 4.27, p = 0.003)	0.761 (t = 4.36, p = 0.025)

Table 3: The most accurately predicted FDA approved drugs for brain tumor along with biological activity and their therapeutic effects

Drug name	Molecular Formula	Molecular weight (g/mol)	IC_{50} (µM)	RMSE	Biological function	Therapeutic effect
Midostaurin	$C_{35}H_{30}N_4O_4$	570.65	0.0047	0.186	Multi-targeted kinase inhibitor that blocks FLT3, KIT, and PDGFR signaling pathways, suppressing abnormal cell proliferation	Therapeutic potential in glioblastoma and other brain tumors, reducing tumor cell viability through inhibition of FLT3-mediated signaling and promoting apoptosis
GW-2580	$C_{20}H_{22}N_4O_3$	366.4	0.02486	0.277	Selective inhibitor of the colony-stimulating factor-1 receptor (CSF1R) kinase, regulates macrophage proliferation and tumor-associated microglia activation	Reduces tumor-associated macrophage activity in the brain tumor microenvironment, Limiting tumor progression and enhancing therapeutic response
FMK	$C_{21}H_{23}FN_2O_4$	467.5	0.015	0.322	Irreversible inhibitor of MAP kinase 1 (MEK1) and ERK signaling pathways	Suppresses glioma cell invasion and promotes apoptotic activity, blocking MAPK signaling cascades in brain tumor models
Imatinib	$C_{29}H_{31}N_7O$	493.603	0.1	0.330	Tyrosine kinase inhibitor targeting BCR-ABL, PDGFR, and c-KIT receptors involved in tumor growth and angiogenesis	Inhibit PDGFR-mediated pathways in gliomas, reducing tumor vascularization and enhancing sensitivity
A-443654	$C_{24}H_{23}N_5O$	397.5	0.000160	0.332	Potent and selective inhibitor of AKT kinase, a key regulator of cell growth and survival	Induces apoptosis in glioblastoma cells by inhibiting AKT-mediated signaling, leading to reduced tumor proliferation and increased cell death

Table 4: Over/Under expressed genes their function and biological impact for GBM

Expression Status	Gene	Function	Biological Insight	Ref
Over-expressed Genes	*EGFR*	Epidermal growth factor receptor	Enhances tumor proliferation and invasion in GBM	[43]
	MKI67	Nuclear proliferation marker	Marker of cell cycle progression and active mitosis	[44]
	BIRC5	Inhibitor of apoptosis (Survivin)	Prevents cell death; linked to treatment resistance	[45]
	TOP2A	DNA topoisomerase II alpha	Facilitates DNA replication and chromosome segregation	[46]
	AURKA	Aurora kinase A	Controls mitotic entry and spindle assembly; associated with tumor progression	[47]
Under-expressed Genes	*GFAP*	Glial fibrillary acidic protein	Astrocyte marker: loss indicates dedifferentiation in GBM	[48]
	MBP	Myelin basic protein	Essential for myelin sheath; typically reduced in glioma tissue	[49]
	NEFL	Neurofilament light polypeptide	Axonal structure protein: under-expression suggests neuronal identity loss	[50]
	SLC1A2	Excitatory amino acid transporter (EAAT2)	Impaired glutamate uptake; associated with excitotoxicity in GBM	[51]
	PLP1	Proteolipid protein 1	Major myelin component in CNS; reduced expression reflects disrupted oligodendrocyte function	[52]

Biographies:

Awais Raza Zaidi holds a bachelor's degree BSCS from the University of Engineering and Technology Lahore in 2012 and the master's degree MSCS from Government College, University, Lahore in 2017. He is currently a PhD Scholar at the Pakistan Institute of Engineering & Applied Sciences (PIEAS), specializing in machine learning and bioinformatics. With over six years of research experience in bioinformatics, machine learning, deep learning, neural networks, and precision oncology, his work focuses on anticancer drug response prediction using hybrid genomic and chemical features. He has published research in reputable venues, including deep learning–based drug response systems, RNN- and MLP-based models.

Abdul Majid received his M.Sc. degree from Quaid-i-Azam University in 1992, Islamabad, Pakistan, and his M.S. (in 2022) and Ph.D. (in 2026) degrees in Computer Systems Engineering from the GIK Institute of Engineering Sciences and Technology, Topi, Pakistan. He has over 23 years of academic and research experience. Currently, he is working as a Professor in the Department of Computer and Information Sciences, Pakistan Institute of Engineering and Applied Sciences (PIEAS). He completed his postdoctoral research at the GIST, South Korea in 2010. His research interests include bioinformatics, computational drug discovery, pattern recognition, and machine learning applications in biomedical engineering.

Muhammad Bilal received his B.S. degree in Computer Science from the Pakistan Institute of Engineering and Applied Sciences (PIEAS), Islamabad, in 2023. His research focuses on the development of AI-enabled biomedical applications, computational modeling, and chemical structure-based analysis for healthcare and pharmaceutical sciences. He is particularly interested in applying advanced machine learning and deep learning techniques to problems in biomedical research, drug discovery, and precision medicine. His work aims to integrate data-driven methodologies with biological and chemical insights to support more accurate prediction, analysis, and decision-making in modern pharmaceutical research.

Tuba Majid received her B.S. degree in Aerospace and Mechanical Engineering from Seoul National University in 2019 and her M.S. degree in Mechanical Engineering from Stony Brook University in 2022. Her research interests include biomimetic systems, computational methods, chemical structure-based modeling, and AI-enabled biomedical applications. Her interdisciplinary research aims to bridge advanced engineering methodologies with biomedical and pharmaceutical sciences in the field of drug discovery.