 doi:10.20944/preprints202202.0051.v1

Machine Learning based Survival Group Prediction in Glioblastoma

Manasa Kalya [1,2], Alexander Kel [2,4]*, Andreas Leha[3], Kamilya Altynbekova[2], Edgar Wingender[2], Tim Beißbarth[1]

1. Department of Medical Bioinformatics, University Medical Center Göttingen, 37099 Göttingen, Germany
2. geneXplain GmbH, 38302 Wolfenbüttel, Germany
3. Institute for Medical Statistics, University Medical Center Göttingen, 37099 Göttingen, Germany
4. Institute of Chemical Biology and Fundamental Medicine SBRAS, 630090, Novosibirsk, Russia

* Correspondence: Alexander E. Kel: alexander.kel@genexplain.com

Keywords: Glioblastoma, survival prediction, Machine Learning, biomarkers, HumanPSD™, Long-term survivors.

Abstract:

Glioblastoma (GBM) is a very aggressive malignant brain tumor with the vast majority of patients surviving less than 12 months (Short-term survivors [STS]). Only around 2% of patients survive more than 36 months (Long-term survivors [LTS]). Studying these extreme survival groups might help in better understanding GBM biology. This work aims at exploring application of machine learning methods in predicting survival groups(STS, LTS). We used age and gene expression profiles belonging to 249 samples from publicly available datasets. 10 Machine learning methods have been implemented and compared for their performances. Hyperparameter tuned random forest model performed best with accuracy of 80% (AUC of 74% and F1_score of 85%). The performance of this model is validated on external test data of 16 samples. The model predicted the true survival group for 15 samples achieving an accuracy of 93.75%. This classification model is deployed as a web tool GlioSurvML. The top 1500 features which retained classification efficiency (Accuracy of 80%, AUC of 74%) were studied for enriched pathways and disease-causal biomarker associations using the HumanPSD™ database. We identified 199 genes as possible biomarkers of GBM and/or similar diseases (like Glioma, astrocytoma, and others). 57 of these genes are shown to be differentially expressed across survival groups and/or have impact on survival. This work demonstrates the application of machine learning methods in predicting survival groups of GBM.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2022 doi:10.20944/preprints202202.0051.v1

1. Introduction

The majority of patients with glioblastoma (GBM) have a short-term survival rate of fewer than 12 months (short-term survivors [STS]), however there is a minority of individuals who have a long-term survival rate of more than three years (36 months), referred to as long-term survivors (LTS)(Hwang et al., 2019a). Clinical, radiological, and histological characteristics have not been found to be predictors of long-term survival or response to therapy in studies (Davis, 2016). (Hwang et al., 2019) Machine Learning (ML) techniques are increasingly being applied in GBM research, as evidenced by a rise in the number of publications in the recent decade (Valdebenito and Medina, 2019). With enormous volumes of high-dimensional data, machine learning aids in recognizing patterns, forecasting events, and interpreting the interactions of complex biochemical networks (Valdebenito and Medina, 2019).

A biomarker is a biological marker that indicates a biological condition and can signal illness-associated molecular alterations at the molecular level which is valuable in understanding the disease state or diagnosis. ML based classification and feature selection methods have aided such a biomarker discovery (Mamoshina et al., 2018; Torres and Judson-Torres, 2019; Fortino et al., 2020; Xie et al., 2021). Some of the major examples of ML use in GBM research are the Stemness Subtype(I/II) Predictor (Wang et al., 2021), NF1 activation status predictor , GBM subtype-specific classifiers (Ensenyat-Mendez et al., 2021), and temozolomide treatment response predictor(Geldof et al., 2020). (Senders et al., 2020)Joeky et al.,2020 has developed an online survival calculator for patients with glioblastoma based on demographic, socioeconomic, clinical, and radiographic variables to predict overall survival.

Transcriptomics approaches have been demonstrated to be highly promising as they offer prognostic techniques for gaining a better knowledge of the condition. Using TCGA RNA-seq data from 129 samples, a study has used an Autoencoder (AE)-based approach for the prediction of GBM patient survival (short-term or long-term survivors) with an accuracy 89%.(Kirtania et al., 2021) In this study, we evaluated 10 ML models to build a classifier which can classify GBM patients into short-term and long-term survivo groups using transcriptomic profiles and clinical information(age) of 249 patients, pooled from 5 publicly available datasets. Random forest model has performed best with an accuracy of 80% and is deployed as a webtool - GlioSurvML. Following model identification, the top 1500 features are used for further analysis to identify important biological pathways and biomarkers.

2. Materials and Methods

2.1. Data Collection

The genome-wide expression profiles based on the Human Genome U133 Plus 2.0 array and **clinical** information of patients with GBM were collected from the public repository of the GEO database. Age information was available for 75.5% of the samples, whereas information on Gender, Karnofsky score, MGMT status, or IDH status were not available for most of them (<30%) and hence only information of age is considered along with the transcriptome to build the survival predictor.

All the datasets were pooled together leading to 176 and 73 samples corresponding to short-term survivors (STS; survival < 12 months) and long-term survivors (LTS; survival > 36 months), respectively (**Table 1).** Duplicates were not removed. Raw data, sample information, and cleaned datasets are given in **Supplementary file 1.**

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2022 doi:10.20944/preprints202202.0051.v1

Table 1. Statistics of datasets studied in this work.

	Platform	Short-term survivors	Long-term survivors
GSE53733 (Reifenberger et al., 2014)	HU133 plus 2.0 arrays	16	23
GSE108474 (Gusev et al., 2018)	HU133 plus 2.0 arrays	97	35
GSE13041 (Lee et al., 2008)	HU133 plus 2.0 arrays	20	02
GSE7696 (Murat et al., 2008)	HU133 plus 2.0 arrays	29	09
GSE43378 (Kawaguchi et al., 2013)	HU133 plus 2.0 arrays	14	04

2.2. Affymetrix microarray data pre-processing

The raw data files (. CEL format) for the above-mentioned datasets were collected from the GEO database- from here on called as GSE dataset. RMA algorithm is used in R (affy package) for background correction, quality check, and normalization to obtain log2 transformed expression values (Gautier et al., 2004). Batch correction of the pooled expression data was performed using empirical Bayes framework is performed (Leek et al., 2012). PCA plot for the batch corrected data is given in **Supplementary file 2.** This batch corrected file is used for further analysis. Multiple Affymetrix ids were summarized to genes ids by choosing the maximum out of probe intensities of multiple probes belonging to a single gene. The final expression matrix comprised 21526 probes and 249 samples is given in **Table S1-C**.

2.3 Development of a Prediction Model Using a Machine Learning Algorithm

To develop a machine learning model, we have used several functionalities of model building in python sklearn (Pedregosa FABIANPEDREGOSA et al., 2011). The dataset used to build the model contains transcriptomics profiles of 176 STS and 73 LTS and the age of the corresponding patient. Using a variance filter the top 10,000 highly variant genes are identified and were considered for model building. Labels were encoded using label encoder. Figure 1 shows the work flow of model development.The samples were first split into 80% training and 20% test data. All the downstream operations to build the predictive model were performed only on training data and is later tested on test data. The training data is scaled and quantile transformed. The scaling and quantiles were saved so that they can be applied to test data.

To deal with the problem of class imbalance during model training (training - STS:139, LTS=60), we have used the synthetic minority oversampling technique SMOTE of the imblearn package (Lemaˆıtre Lemaˆıtre et al., 2017). This oversampling strategy first randomly selects an instance from the minority class and finds its k nearest minority class neighbors. Synthetic data would then be made between the random data and the randomly selected k-nearest neighbor. With SMOTE oversampling, the number of samples in the minority class was increased to 139. On this resampled training data, we applied 10 ML models. However, only the random forest model performed better in terms of classifying the minority classes. For hyperparameter tuning of model parameters we used GridSearchCV. Models

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2022 doi:10.20944/preprints202202.0051.v1

were tuned for their hyperparameters (Table 2) for optimal performances. Hyperparameter tuning results for all ML models are given in **Table S3-A.**

Table 2. Hyperparameter tuning in ML models

Method	Parameters
Random forest	Criterion, max_depth, n_estimators
Logistic regression	penalty, Solver & C
Linear Support Vector Classification (Linear SVC)	C, kernel, gamma,
Support Vector Classification (SVC)	Kernel, C, gamma
Nu-Support Vector Classification (NuSVC)	Nu, Kernel, decision_function_shape
Naïve Bayes	var_smoothing
Classification and Regression Trees (CART)	Criterion, max_features
k-nearest neighbors (KNN)	N_neighbors,algorithm & weights
Balanced random forest	max_features, n_estimators, max_depth, criterion
Balanced Bagging	n_estimators

Hyperparameter tuned models were applied on the (20%) test data to evaluate model performances and choose the best performing classifier. The best performing model was evaluated on an external independent microarray data to evaluate the application of this classifier as a reliable tool for predicting Glioblastoma survival groups. The top best features based which retains higher classification efficiency were extracted and evaluated for biological relevance by using Gene set enrichment, Differential expression, Survival significance and their association with Glioblastoma or similar diseases.

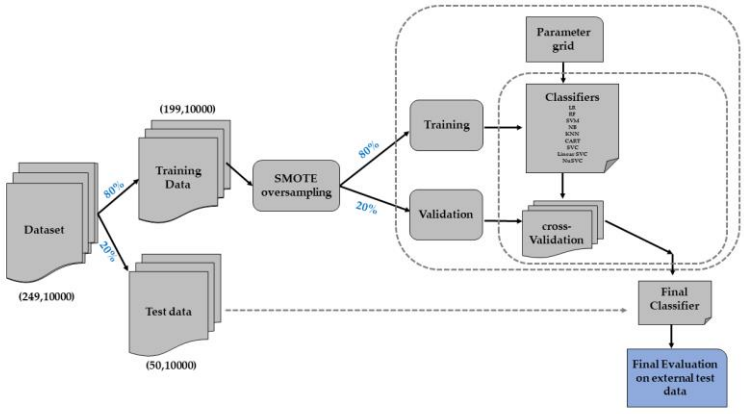


Figure 1. Workflow explaining the steps of building ML models.

2.4.Gene Enrichment Analysis

To explore the biological importance of these 1500features, gene list enrichment tool enrichR (Chen et al., 2013) is used. Enrichment for Molecular Signature Database (MSigDB) (Liberzon et al., 2011) is used.

2.5 Differential gene expression (DEG) analysis

LIMMA (Linear Models for Microarray Data) method was applied to identify differentially expressed genes (Ritchie et al., 2015) . Differential gene expression analysis for short-term and long-

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2022 doi:10.20944/preprints202202.0051.v1

term survivors is performed in GSE108474 and TCGA GBM microarray data. Clinical information and cleaned datasets of GSE108474 and TCGA GBM microarray data are given in **Supplementary 4.**

2.6. Impact on survival

Survival and Survminer libraries in R are used to perform univariate survival analysis. Univariate Cox regression for survival analysis is performed using the coxph function of the Survival package to calculate the Hazard ratio (HR) with p-value cutoff of 0.05 for significance (Therneau, 2021). KMplots are used to depict impact of genes on survival with non-overlapping 50% upper and lower quantiles. **supplementary 4**

2.7 Identification of biomarkers

Causal molecular mechanisms present a unifying principle for disease classification, analysis of clinical disorder associations, as well as prediction of disease genes, diagnostic markers, and therapeutic targets. A novel approach published (Stegmaier et al., 2010) built of 1000 causal gene-disease networks is now updated and available in the HumanPSD™ database (Wingender et al., 2007). The important features identified using the ML model can serve as biomarkers of survival/prognosis in GBM. HumanPSD™ database 2021.2 is mined to fetch information on the association of these features with GBM or similar diseases

3. Results

3.1. Development of ML model:

The genome-wide expression profiles from 5 independent experiments using Human Genome U133 Plus 2.0 arrays with corresponding clinical information of Glioblastoma patients were collected, normalized and integrated to obtain a data matrix of 176 and 73 samples corresponding to short-term survivors (STS; survival < 12 months) and long-term survivors (LTS; survival > 36 months), respectively. Top 10k highly variant genes were used for building ML models for classification. See more details in methods section (**Supplementary file 1 and 2)**

In the current work, we have used machine learning methods to predict the survival class of GBM patients using gene expression profiles.

Ten ML models such as random forest, Naïve Bayes, Support Vector Classification, Linear SVC, NuSVC, Logistic Regression, Classification and Regression Trees (CART), k-nearest neighbors (KNN), and specialized packages of imbalanced learning like Balanced Random forest and Balanced Bagging are evaluated in this study. The dataset was split into 80% training and 20% test data. To address the problem of class imbalance, SMOTE oversampling is applied during the training of the model to balance the classes. GridSearchCV upon StratifiedShuffleSplit on the oversampled training data is used for hyperparameter tuning of the models (Table S3-A and Table S3-B). The performance of all the hyperparameter tuned models on the test data is given in **Table 1**.

We found that hyperparameter-tuned random forest model (**Figure 2)** performed best out of all other models mentioned earlier, with f1_score of 86.48%, Accuracy of 80%, and AUC of 74% on test data. This corresponds to 86% of true labels in majority class and 62% true labels in minority class (**Figure 3A)**

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2022 doi:10.20944/preprints202202.0051.v1

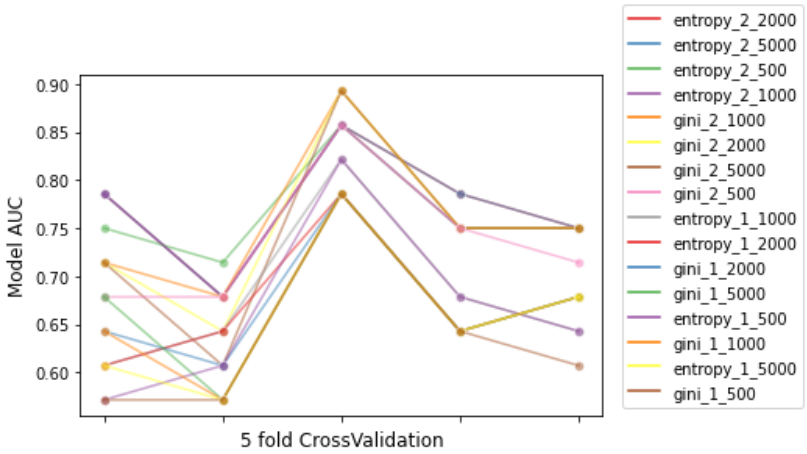
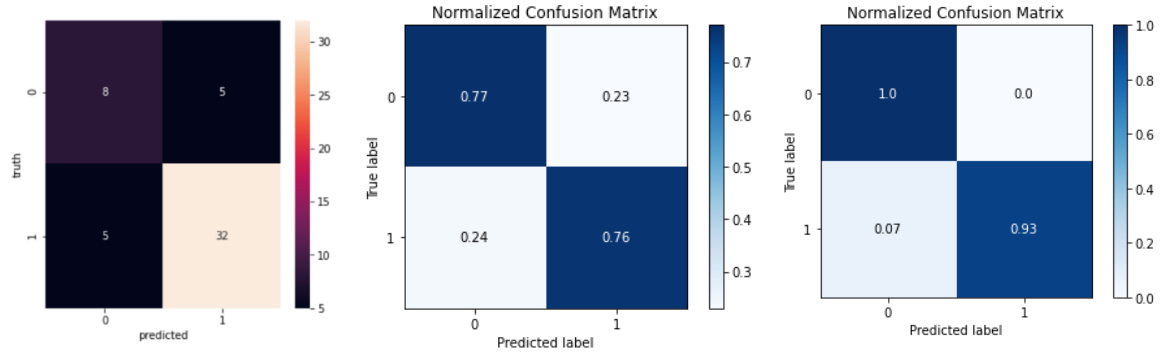


Figure 2. Hyperparameter Tuning in RF. The following hyperparameters were tuned: Tuning parameters of criterion(gini/entropy), maximum depth (1/2) and number of estimators (500/1000/2000/5000) for random forest model upon 5-fold cross validation using GridSearchCV.

The hyperparameter tuned BalancedRandomForest model performed with f1_score of 82.3%, Accuracy of 76%, AUC of 76.29% on test data. The model positively identified 77% of minority labels and 78% of majority labels (**Figure 3B)**. The linear models like LR, SVC, NuSVC, LinearSVC had lower AUC values as they identified less than 35% of the minority class (LTS) and hence were not considered in our further analysis.

Table 3. Performance of 10 ML models under study on 20% test data upon hyperparameter tuning

Hyperparameter tuned ML model	F1_Score	Accuracy	AUC
Logistic Regression	0.81	0.720	0.636
Random forest	0.864	0.800	0.740
NuSVC	0.864	0.780	0.626
SVC	0.864	0.787	0.626
Balanced random forest	0.823	0.760	0.762
Balanced Bagging	0.853	0.780	0.701
Linear SVC	0.746	0.660	0.645
Naïve Bayes	0.805	0.720	0.661
KNN	0.407	0.360	0.417
CART- Decision Trees	0.788	0.700	0.647



Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2022 doi:10.20944/preprints202202.0051.v1

Figure3. Normalized Confusion Matrix for ML models.
Normalized Confusion matrix for the classification of survival groupsis shown here.For the classes, 0(LTS) and 1(STS), the X-axis in the plot is for the predicted class and the Y-axis is for the true class. The true class elements of a row are spread across columns and the elements of the matrix are normalized row wise, i.e., sum of fractions along a row sum to 1. The only true predictions are along the diagonal, i.e., each of the i–ith element of the matrix and all other off-diagonal elements along a row are wrong predictions. The more the correctness of a class, the darker the blue hue it has in a cell of the plot of the confusion matrix. A) Normalized Confusion Matrix of Random forest model on internal (20%) test data B) Normalized Confusion Matrix for BalancedRandom forest model without oversampling C) Normalized Confusion Matrix for Random forest model on external test data

To build a robust machine learning model which can identify the survival class of the GBM patients, we tested the random forest model on an external microarray dataset (**Supplementary file 7)**. The LTS are rare events and hard to find adequate samples for testing. The external dataset containing 16 samples (1-LTS and 15-STS) was from a single experiment. Random forest model performed with an accuracy of 93.75% (AUC of 96.66%) (**Figure 3C).**

Age was found to be one of the top important (Top 7) features of the random forest model developed. The random forest model built on gene-expression and age had better sensitivity (93.75%) than the random forest model built on gene expression alone (81.25%) **(Supplementary file 7).**

3.2. Deployment of ML model:

The random forest model developed here for survival class prediction is deployed as a webtool-GlioSurvML. All information associated is given in github repository. Webtool has 2 models of RF one with including age and one without age. The webtool prints the output as a PDF report as well as an excel-table. (**Supplementary file 8)**

3.3. Feature Importances

Ranking of features/genes according to their importance in the random forest classification model discussed above is given in **Table S3-C**. The performance of the model using top 100/500/1000/1500/2000 features (**Table S3-D**) is investigated. We observed that the top 1500 features (**Table S3-E**) were sufficient enough to maintain the 80% accuracy of prediction. These genes are looked for their relevance in the disease using gene enrichment analysis, differential expression analysis, univariate survival analysis to investigate prognostic value and by utilizing existing knowledge on biomarkers of the glioblastoma.

We found that TNF-alpha Signaling via NF-kB, mTORsignalling, G2-M checkpoints, Epithelial to Mesenchymal transition are some of the top overlapping gene sets according to MsigDB **Table S3-F.**

3.4 Biomarker Identification

Exploiting the previously reported method on unifying disease mechanisms based on causal gene-disease associations as described in HumanPSD™ database (**Supplementary file 5**), we find that, out of top 1500 genes, 63 known gene expression biomarkers of Glioblasytoma and 136 gene expression biomarkers from similar diseases to Glioblastoma and 35 markers were reported both in Glioblastoma and in one of the similar diseases according to HumanPSD™ database (199 unique biomarkers in total). **Figure 4.** Based on this analysis, we propose 171(136+35) gene expression based biomarkers to Glioblastoma. According to the database, these genes were mapped to 8 diseases like Osteosarcoma, Melonoma, Ovarian neoplasm, Nasopharangeal neoplasm including Glioma , astrocytoma, brain

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2022 doi:10.20944/preprints202202.0051.v1

neoplasms. Top 10 (based on feature ranking in random forest model) of these new proposed biomarkers of Glioblastoma prognoisis are given in **Table 4**.

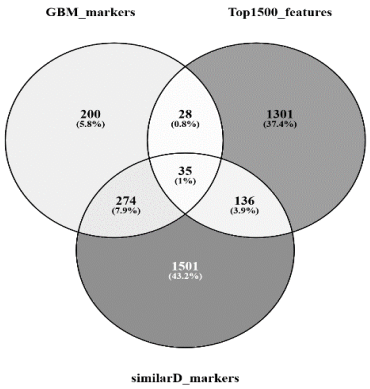


Figure 4: Venn Diagram of HumanPSD™ biomarkers and important features.
HumanPSD™ database reports 537 mRNAexpression based Glioblastoma markers, 1946 mRNA expression based biomarkers of diseases similar to GBM. Out of the top 1500 important features required for classifying the survival group of GBM, 63 Glioblastoma and 171 similar disease biomarkers were found overlapping. 35 genes were found associated with both GBM and related disease.

These biomarkers are checked for differential gene expression between STS and LTS and univariate impact on survival. The analysis is performed in GSE108474 dataset which is U133 plus 2 affymetrix platform and TCGA-GBM of 560 microarray (U133 Affy array) datasets (**Supplementary file 4).**

Table 4. Top 10 features proposed as biomarkers of prognosis in Glioblastoma in our study

Features	Feature_Rank	Molecule	Disease	Disease_Association	PMID
CBX3	25	mRNA	Osteosarcoma	increased expression of CBX3 mRNA correlates with increased neoplasm metastasis associated with osteosarcoma	22870217
GHR	29	mRNA	Melanoma	increased expression of GHR mRNA correlates with neoplasm metastasis associated with melanoma	24134847
HNRNPA2B1	38	mRNA	Brain Neoplasms	increased expression of HNRNPA2B1 mRNA correlates with oligodendroglioma tumors associated with brain neoplasms	11485829
NES	41	mRNA	Astrocytoma	increased expression of NES mRNA may correlate with disease progression associated with astrocytoma	17611714
SKP2	44	mRNA	Ovarian Neoplasms	decreased expression of SKP2 mRNA may correlate with increased response to salinomycin associated with ovarian neoplasms	23807222
RARRES2	48	mRNA	Glioma	increased expression of RARRES2 mRNA correlates with glioma	21949124
ERBB2	58	mRNA	Ovarian Neoplasms	increased expression of ERBB2 mRNA may correlate with malignant form of ovarian neoplasms	8094034
ELAVL1	63	mRNA	Ovarian Neoplasms	decreased expression of ELAVL1 mRNA may prevent increased positive regulation of gene expression associated with ovarian neoplasms	23394580
TGIF2	68	mRNA	Ovarian Neoplasms	increased expression of TGIF2 mRNA correlates with ovarian neoplasms	11006116

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2022 doi:10.20944/preprints202202.0051.v1

FZD1	80	mRNA	Ovarian Neoplasms	increased expression of FZD1 mRNA correlates with glandular and epithelial neoplasms associated with ovarian neoplasms	19148501

The information of differential gene expression (Log2FC, adj.pvalue) and survival significance (Hazard Ratio and FDR <0.05) for these 199 biomarkers in GSE108474 are given in **Supplementary File 6**. Out of these, 17 genes were significantly differentially expressed, 28 had survival significance and 12 biomarkers were both differentially expressed and had significant impact on survival.

4. Discussion

In this study, we evaluated application of 10 ML models to build a classifier to differentiate patients between STS and LTS groups based on their transcriptomic profiles and clinical information(age) from 249 patients data which is pooled from publicly available datasets. To the best of our knowledge this is the first application of its kind. Of the models evaluated, a random forest model performed best with accuracy of 80% (F1_score=86.4% AUC =74%). Furthermore, this model is evaluated on external microarray data and found to have high accuracy of 93.75% (AUC of 96.66%). The identification of age as an important feature is in line with the observation that age is an important clinical predictor for survival.We have noted that the top 1500 features alone can preserve the classification efficiency of the model and these are only used for further analysis.

The enrichment analysis revealed enrichment of TNF-Alpha via NF-kB, mTOR signalling, G2-M checkpoints, Epithelial to Mesenchymal transition signaling pathways. All of these pathways are identified as therapeutic targets in GBM (ref) and play a role in response to Temozolomide (ref), which is a first line of treatment in GBM.

Using HumanPSD™ we have identified 8 disorders which are mapped to be similar to Glioblastoma. Of these three are related to central nervous system tumors and others include ovarian, osteosarcoma, melanoma, nasopharyngeal tumors and general neoplasms. This identified overlap of GBM with gliomas and melanoma is interesting as studies have shown increased risk of gliomas in malignant melanoma patients (Scarbrough et al., 2014) and increased representation of melanoma in GBM patients (Yang et al., 2021). The gliomas and melanoma are shown to be responsive to Temozolomide which is indicative of a common potential pathophysiological pathway (Desai and Grossman, 2008).

From the HumanPSD™ we have identified 199 mRNA biomarkers that have previously been linked to Glioblastoma and/or related.

Some of the important biomarkers include retinoic acid receptor responder 2(RERRES2), Distinct Subgroup of The Ras Family Member 3 (DIRAS3), DEP Domain Containing MTOR Interacting Protein (DEPTOR), Insulin like Growth Binding Protein 5(IGFBP5) and C-Type Lectin Domain Family 2 Member B (CLEC2B). RERRES2 is a critical gene of retinoic acid signaling which is reported to be highly upregulated in STS in GBM (Barbus et al., 2011). DIRAS3 drives autophagy by Ras/AKT/mTOR pathway in GBM and is reported to be significantly downregulated in long-term survivors of GBM (Zhong et al., 2019). DEPTOR is a natural inhibitor of MTORc1 and mTORc2 which plays an important role in autophagy. Inhibitors of mTOR signaling are widely discussed as an adjuvant therapy to regulate

 doi:10.20944/preprints202202.0051.v1

autophagy in GBM (Xia et al., 2020). IGFBP5 promotes cell invasion by regulating Epithelial to Mesenchymal Transition and inhibits cell proliferation by suppressing the phosphorylation of AKT in GBM (Dong et al., 2020). Its expression was upregulated in high grades of glioma and is correlated with worse prognosis (Dong et al., 2020). CLEC2B - A rise in expression of CLEC2B was linked to a rise in the progression-free Hazard ratio (Serão et al., 2011)

Identifying the signaling pathways and biomarkers that are related to Glioblastoma, mapping to the diseases which are related to CNS or those with shared biology gives strength to our machine learning model and reinforces the idea that machine learning models can be used for understanding the biology of GBM. Our analysis has shown inclusion of clinical information i.e. age has increased the sensitivity of survival group prediction which shows the importance of adding clinical information to the machine learning models. Other clinically important variables are not added to the model due to high levels of missingness in the datasets which needs to be addressed while collecting the future data. One important limitation of the current study is that the method is applicable only for microarray platforms and extension of this model for application in RNA-seq data requires further work.

5. Conclusion

The current study presents a Machine Learning model for use in research to classify patients into Glioblastoma survival groups, deploys application as a webtool, discusses important features for relevance in the disease, proposes new plausible markers of survival in Glioblastoma.

Availability of software, data and materials: All the datasets analyzed in the current study are available from previous publications. All datasets, models and supplementary materials are available here:

https://github.com/genexplain/Manasa_KP_et_al_MLmodels_predictionofGBMsurvivorgroups

Conflicts of Interest: The authors Manasa Kalya, Andreas Leha and Tim Beißbarth are from Department of Medical Bioinformatics, University Medical Center Göttingen, Kamilya Altynbekova, Alexander Kel and Edgar Wingender are employees of geneXplain GmbH.

Author Contributions: AK is involved in conceptualization, providing resources, supervision and manuscript reviewing. MKP has conceptualized the work, performed data collection, data analysis, interpreting results and writing the manuscript. AL has extensively participated in developing the data analysis pipeline, KA has participated in making the webtool application. EW and TB are involved in supervision of work and in reviewing the draft.

Funding: This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 766069.

Acknowledgements: It is my pleasure to acknowledge Ravi Kumar Nadella who has read and given comments on every version of this paper and for his expertise, discussions and assistance in improving the manuscript.

References

Barbus, S., Tews, B., Karra, D., Hahn, M., Radlwimmer, B., Delhomme, N., et al. (2011). Differential retinoic acid signaling in tumors of long- and short-term glioblastoma survivors. *J. Natl. Cancer Inst.* 103, 598–601. doi:10.1093/JNCI/DJR036.

Chen, E. Y., Tan, C. M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G. V., et al. (2013). Enrichr: Interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* 14. doi:10.1186/1471-2105-14-128.

Davis, M. E. (2016). Glioblastoma: Overview of disease and treatment. *Clin. J. Oncol. Nurs.* 20, 1–8. doi:10.1188/16.CJON.S1.2-8.

Desai, A. S., and Grossman, S. A. (2008). Association of melanoma with glioblastoma multiforme. *https://doi.org/10.1200/jco.2008.26.15_suppl.2082* 26, 2082–2082. doi:10.1200/JCO.2008.26.15_SUPPL.2082.

Dong, C., Zhang, J., Fang, S., and Liu, F. (2020). IGFBP5 increases cell invasion and inhibits cell proliferation by EMT and Akt signaling pathway in Glioblastoma multiforme cells. *Cell Div.* 15, 1–9. doi:10.1186/S13008-020-00061-6/FIGURES/5.

Ensenyat-Mendez, M., Íñiguez-Muñoz, S., Sesé, B., and Marzese, D. M. (2021). iGlioSub: an integrative transcriptomic and epigenomic classifier for glioblastoma molecular subtypes. *BioData Min.* 14, 1–16. doi:10.1186/S13040-021-00273-8/FIGURES/5.

Fortino, V., Wisgrill, L., Werner, P., Suomela, S., Linder, N., Jalonen, E., et al. (2020). Machine-learning–driven biomarker discovery for the discrimination between allergic and irritant contact dermatitis. *Proc. Natl. Acad. Sci. U. S. A.* 117, 33474–33485. doi:10.1073/PNAS.2009192117.

Gautier, L., Cope, L., Bolstad, B. M., and Irizarry, R. A. (2004). affy--analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* 20, 307–315. doi:10.1093/bioinformatics/btg405.

Geldof, T., van Damme, N., Huys, I., and van Dyck, W. (2020). Patient-level effectiveness prediction modeling for glioblastoma using classification trees. *Front. Pharmacol.* 10, 1665. doi:10.3389/FPHAR.2019.01665/BIBTEX.

Gusev, Y., Bhuvaneshwar, K., Song, L., Zenklusen, J. C., Fine, H., and Madhavan, S. (2018). The REMBRANDT study, a large collection of genomic data from brain cancer patients. *Sci. data* 5. doi:10.1038/SDATA.2018.158.

Hwang, T., Mathios, D., McDonald, K. L., Daris, I., Park, S. H., Burger, P. C., et al. (2019a). Integrative analysis of DNA methylation suggests down-regulation of oncogenic pathways and reduced somatic mutation rates in survival outliers of glioblastoma. *Acta Neuropathol. Commun.* 7, 5. doi:10.1186/S40478-019-0744-0.

Hwang, T., Mathios, D., McDonald, K. L., Daris, I., Park, S. H., Burger, P. C., et al. (2019b). Integrative analysis of DNA methylation suggests down-regulation of oncogenic pathways and reduced somatic mutation rates in survival outliers of glioblastoma. *Acta Neuropathol. Commun.* 7, 5. doi:10.1186/S40478-019-0744-0.

Kawaguchi, A., Yajima, N., Tsuchiya, N., Homma, J., Sano, M., Natsumeda, M., et al. (2013). Gene expression signature-based prognostic risk score in patients with glioblastoma. *Cancer Sci.* 104, 1205–1210. doi:10.1111/CAS.12214.

Kirtania, R., Banerjee, S., Laha, S., Shankar, B. U., Chatterjee, R., and Mitra, S. (2021). Deepsgp:Deep learning for gene selection and survival group prediction in glioblastoma. *Electron.* 10, 1463. doi:10.3390/ELECTRONICS10121463/S1.

Lee, Y., Scheck, A. C., Cloughesy, T. F., Lai, A., Dong, J., Farooqi, H. K., et al. (2008). Gene expression analysis of glioblastomas identifies the major molecular basis for the prognostic benefit of younger age. *BMC Med. Genomics* 1. doi:10.1186/1755-8794-1-52.

Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E., and Storey, J. D. (2012). The SVA package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* 28, 882–883. doi:10.1093/bioinformatics/bts034.

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., and Mesirov, J. P. (2011). Molecular signatures database (MSigDB) 3.0. *Bioinformatics* 27, 1739–1740. doi:10.1093/bioinformatics/btr260.

Mamoshina, P., Volosnikova, M., Ozerov, I. V., Putin, E., Skibina, E., Cortese, F., et al. (2018). Machine learning on human muscle transcriptomic data for biomarker discovery and tissue-specific drug target identification. *Front. Genet.* 9, 242. doi:10.3389/FGENE.2018.00242/BIBTEX.

Murat, A., Migliavacca, E., Gorlia, T., Lambiv, W. L., Shay, T., Hamou, M. F., et al. (2008). Stem cell-related "self-renewal" signature and high epidermal growth factor receptor expression associated with resistance to concomitant chemoradiotherapy in glioblastoma. *J. Clin. Oncol.* 26, 3015–3024. doi:10.1200/JCO.2007.15.7164.

Reifenberger, G., Weber, R. G., Riehmer, V., Kaulich, K., Willscher, E., Wirth, H., et al. (2014). Molecular characterization of long-term survivors of glioblastoma using genome- and transcriptome-wide profiling. *Int. J. Cancer* 135, 1822–1831. doi:10.1002/ijc.28836.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 43, e47–e47. doi:10.1093/nar/gkv007.

Scarbrough, P. M., Akushevich, I., Wrensch, M., and Il'yasova, D. (2014). Exploring the association between melanoma and glioma risks. *Ann. Epidemiol.* 24, 469. doi:10.1016/J.ANNEPIDEM.2014.02.010.

Senders, J. T., Staples, P., Mehrtash, A., Cote, D. J., Taphoorn, M. J. B., Reardon, D. A., et al. (2020). An Online Calculator for the Prediction of Survival in Glioblastoma Patients Using Classical Statistics and Machine Learning. *Neurosurgery* 86, E184–E192. doi:10.1093/NEUROS/NYZ403.

Serão, N. V., Delfino, K. R., Southey, B. R., Beever, J. E., and Rodriguez-Zas, S. L. (2011). Cell cycle and aging, morphogenesis, and response to stimuli genes are individualized biomarkers of glioblastoma progression and survival. *BMC Med. Genomics* 4, 1–21. doi:10.1186/1755-8794-4-49/FIGURES/6.

Stegmaier, P., Krull, M., Voss, N., Kel, A. E., and Wingender, E. (2010). Molecular mechanistic associations of human diseases. *BMC Syst. Biol.* 4, 124. doi:10.1186/1752-0509-4-124/FIGURES/9.

Therneau, T. (2021). A package for survival analysis in R.

Torres, R., and Judson-Torres, R. L. (2019). Research Techniques Made Simple: Feature Selection for Biomarker Discovery. *J. Invest. Dermatol.* 139, 2068-2074.e1. doi:10.1016/J.JID.2019.07.682.

Valdebenito, J., and Medina, F. (2019). Machine learning approaches to study glioblastoma: A review of the last decade of applications. *Cancer Rep.* 2. doi:10.1002/CNR2.1226.

Wang, Z., Wang, Y., Yang, T., Xing, H., Wang, Y., Gao, L., et al. (2021). Machine learning revealed stemness features and a novel stemness-based classification with appealing implications in discriminating the prognosis, immunotherapy and temozolomide responses of 906 glioblastoma patients. *Brief. Bioinform.* 22, 1–20. doi:10.1093/BIB/BBAB032.

Wingender, E., Hogan, J., Schacherer, F., Potapov, A. P., and Kel-Margoulis, O. (2007). Integrating

pathway data for systems pathology. in *In Silico Biology*.

Xia, Q., Xu, M., Zhang, P., Liu, L., Meng, X., and Dong, L. (2020). Therapeutic Potential of Autophagy in Glioblastoma Treatment With Phosphoinositide 3-Kinase/Protein Kinase B/Mammalian Target of Rapamycin Signaling Pathway Inhibitors. *Front. Oncol.* 10, 1886. doi:10.3389/FONC.2020.572904/BIBTEX.

Xie, Y., Meng, W. Y., Li, R. Z., Wang, Y. W., Qian, X., Chan, C., et al. (2021). Early lung cancer diagnostic biomarker discovery by machine learning methods. *Transl. Oncol.* 14, 100907. doi:10.1016/J.TRANON.2020.100907.

Yang, K., Stein, T. D., Huber, B. R., Sartor, E. A., Rachlin, J. R., and Mahalingam, M. (2021). Glioblastoma and malignant melanoma: Serendipitous or anticipated association? *Neuropathology* 41, 65–71. doi:10.1111/NEUP.12702.

Zhong, C., Shu, M., Ye, J., Wang, X., Chen, X., Liu, Z., et al. (2019). Oncogenic Ras is downregulated by ARHI and induces autophagy by Ras/AKT/mTOR pathway in glioblastoma. *BMC Cancer* 19, 1–14. doi:10.1186/S12885-019-5643-Z/FIGURES/7.