



Predicting Survival in Glioblastoma Using Gene Expression Databases: A Neural Network Analysis

Parisa Azimi^{1*}, Taravat Yazdanian², Amirhosein Zohrevand³,
Abolhassan Ahmadiani¹

1. Neurosurgeon, Neuroscience Research Center, Shahid Beheshti University of Medical Sciences, Tehran, Iran,

2. Research Fellow at the Neurological Clinical Research Institute and Healey and AMG Center for ALS, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, USA.

3. Department of Neurosurgery, School of Medicine, Babol University of Medical Sciences, Babol, Iran.

Article type: ABSTRACT

Original Article

Glioblastoma (GBM) is the most aggressive and lethal brain tumor. Artificial neural networks (ANNs) have the potential to make accurate predictions and improve decision making. The aim of this study was to create an ANN model to predict 15-month survival in GBM patients according to gene expression databases. Genomic data of GBM were downloaded from the CGGA, TCGA, MYO, and CPTAC. Logistic regression (LR) and ANN model were used. Age, gender, IDH wild-type/mutant and the 31 most important genes from our previous study, were determined as input factors for the established ANN model. 15-month survival time was used to evaluate the results. The normalized importance scores of each covariate were calculated using the selected ANN model. The area under a receiver operating characteristic (ROC) curve (AUC), Hosmer-Lemeshow (H-L) statistic and accuracy of prediction were measured to evaluate the two models. SPSS 26 was utilized. A total of 551 patients (61% male, mean age 55.5 ± 13.3 years) patients were divided into training, testing, and validation datasets of 441, 55 and 55 patients, respectively. The main candidate genes found were: FN1, ICAM1, MYD88, IL10, and CCL2 with the ANN model; and MMP9, MYD88, and CDK4 with LR model. The AUCs were 0.71 for the LR and 0.81 for the ANN analysis. Compared to the LR model, the ANN model showed better results: Accuracy rate, 83.3 %; H-L statistic, 6.5 %; and AUC, 0.81 % of patients. The findings show that ANNs can accurately predict the 15-month survival in GBM patients and contribute to precise medical treatment.

Received:

2024.05.26

Revised:

2024.06.08

Accepted:

2024.06.08

Keywords: GBM, gene expression, survival prediction, ANNs

Cite this article: Azimi P, *et al.* Predicting Survival in Glioblastoma Using Gene Expression Databases: A Neural Network Analysis. *International Journal of Molecular and Cellular Medicine*. 2024; 13(1):79-90. **DOI:** 10.22088/IJMCM.BUMS.13.1.79

*Corresponding: Parisa Azimi

Neurosurgeon, Neuroscience Research Center, Shahid Beheshti University of Medical Sciences, Tehran, Iran.

E-mail: parisa.azimi@gmail.com



© The Author(s).

Publisher: Babol University of Medical Sciences

This work is published as an open access article distributed under the terms of the Creative Commons Attribution 4.0 License (<http://creativecommons.org/licenses/by-nc/4>). Non-commercial uses of the work are permitted, provided the original work is properly cited.

Introduction

Glioblastoma (GBM) is highly heterogeneous and is an aggressive brain tumor that arises from the glial cells in the brain. These tumors are diagnosed in about 48.6% of the malignant central nervous system (CNS), are more likely to grow rapidly, often spread to other tissue, and are associated with poor prognosis (1). Despite the advances made in surgical procedures, and chemoradiotherapy, GBMs are rarely cured as these tumors are very invasive, and the median overall survival (OS) of GBM patients has remained at about 15 months (2). Therefore, other therapies and precision medicine are required for these patients. In this regard, radiomic features can be used in an artificial neural network (ANN) paradigm to predict histological and molecular confirmation, and clinical outcome measures, thus facilitating precision medicine for improving GBM patient care (3).

GBM survival prediction is important for planning the treatment strategy and assessing treatment results (4). Medical software systems have been applied to develop prediction models to estimate diagnosis, survival, and treatment effects precisely. These contain logistic regression (LR) and machine learning or artificial neural networks (ANNs). LR is a traditional statistical model generally employed in medical practice to interpret clinical data. ANNs offer a novel method for precisely assessing, which could incorporate covariates that may not be considered in traditional regression procedures to develop a predictive model by learning patterns and nonlinear relationships between potential covariates from training data. ANN is a computational model according to the functioning of biological neural networks that can be applied as nonlinear statistical data modeling tools with which the complex relationships between inputs and outputs are modeled. The ANNs try to simulate the learning process of human beings. They are made of a group of interconnected nodes (artificial neurons) that interact with each other based on predefined computational rules. Based on these rules, passing sample data (pairs of observed input/output data) through ANNs makes them modify their structure to estimate the input/output relationship pattern of the systems under study. The resultant network at the end of this learning process can then estimate or predict outputs for new inputs. The most common type of ANN is called a multilayer perceptron (MLP), which consists of 3 layers: 1) input layer; 2) hidden layer; and 3) output layer. More details are available in the scientific literature (5-7).

It is challenging to precisely estimate survival in GBM patients. Also, relationships between gene expression and survival prediction in GBM patients have been so far less discussed in scientific literature by ANNs. Hence, this study aims to create an ANN model based on the age, gender, IDH mutation status, and 31 genes selected from our previous study (4). It also sets out to determine whether ANNs perform better at predicting 15-month survival compared with logistic regression in GBM patients.

Materials and methods

Patients and data collection

The raw data on gene expression for GBM patients were downloaded from four public databases: The Cancer Genome Atlas (TCGA) (<http://xena.ucsc.edu/>), Clinical Proteomic Tumor Analysis Consortium (CPTAC) (<https://portal.gdc.cancer.gov/>), Chinese Glioma Genome Atlas (CGGA) (www.cgga.org.cn), and Mayo Clinic Brain Tumor Patient-Derived Xenograft National Resource (MAYO-PDX) (8) databases,

including 164, 99, 225, and 63 primary GBM samples from TCGA, CPTAC, CGGA, and MAYO-PDX databases, respectively. Clinical data in the cohorts included age, gender, GBM grade, overall survival, and isocitrate dehydrogenase (IDH) mutation status. Some patients with unreachable or uncertain clinical data were removed.

Gene selection for ANN and logistic regression models

In our previous study, a systematic review was performed to discover top gene expressions for survival prediction in GBM patients (4). All 613 genes (with $p < 0.05$) from this review study were included in the bioinformatic analysis. The top 31 genes including IL6, EGFR, STAT3, MMP9, CD44, FN1, CD4, TGFB1, CXCL8, CCL2, IL10, ICAM1, IL1A, CD274, KDR, SPP1, ITGB2, CDKN2A, PARP1, MYD88, AGT, NOTCH1, SERPINE1, TNFRSF1A, CDK1, CAV1, ITGB3, CDK4, FOXO3, MDM2, and PROM1, respectively, were recognized. In our other previous study, through bioinformatic analysis and an RT-qPCR method, it was suggested that the expression of cancer-testis antigens (CTAs) of CEP55 and FBXO39 was significantly higher in GBM cases compared to controls. Moreover, these genes were significantly associated with the survival of GBM cases (9). In the combined lists of genes from two studies [4, 9], 33 genes including IL6, EGFR, STAT3, MMP9, CD44, FN1, CD4, TGFB1, CXCL8, CCL2, IL10, ICAM1, IL1A, CD274, KDR, SPP1, ITGB2, CDKN2A, PARP1, MYD88, AGT, NOTCH1, SERPINE1, TNFRSF1A, CDK1, CAV1, ITGB3, CDK4, FOXO3, MDM2, PROM1, CEP55 and FBXO39 were considered to ANN and logistic regression models. Two genes, including CXCL8 and ITGB3, were not included in the final analysis because they were not listed in all databases.

ANN model

The ANN was constructed using the standard SPSS 26 approach. A multi-layer perceptron (MLP) method was chosen for the current study. MLP-ANNs consist of a set of nodes arranged in three layers: an input layer, a hidden layer with 20 nodes activated with the hyperbolic tangent activation function, and an output layer. The MLP-ANN used observed data consisting of inputs (age, gender, IDH mutation status and 31 genes) and outputs (15-month survival) to define (learn) the complex link between inputs and outputs. The patients were separated by 80%, 10% and 10% to form a training group, a test group, and a validation group, respectively. Once the MLP-ANN was trained, it was able to estimate (predict) the results (outputs) from new input datasets (6). Some experiments were run to optimize an MLP model when designing a neural network. The optimal settings were as follows:

Initial mode choice

Multi-layer Perceptron

Architecture:

Minimum number of nodes in the hidden layer, 1

Maximum number of nodes in the hidden layer, 50

Training criteria

Type of training: batch

Optimization algorithm: scaled conjugate gradient

Initial Lambda: 0.0000005

Initial Sigma: 0.00005

Interval center: 0

Interval offset: 0.5

User missing values

Exclude

Stopping rules

Maximum steps without a decrease in error: 1

Default options were applied for any other choices

For each hidden layer, the weight and bias values are randomly selected and updated with the input values for further processing in the subsequent hidden layers.

Logistic regression

Logistic regression is used to estimate the association of one or more independent (predictor) variables with a binary dependent (outcome) variable. A binary variable is a categorical variable that can take only 2 different values or levels, e.g. “dead or alive” or “0 and 1”. LR can be applied to estimate the probability of a particular outcome depending on the value(s) of the independent variable(s) (10). The conventional statistical analysis of the significance of the parameters was performed using standard logistic regression on the same data set on which the ANN was assessed.

Statistical analysis

All statistical analyses were performed using the SPSS 26 (SPSS, Inc., Chicago, IL, USA). The raw data were normalized to maximum (1) and minimum (0) separately for each gene and database. The normalized importance scores of each covariate were calculated using the selected ANN model. Evaluating the goodness of fit of ANN and LR models is crucial to ensure the accuracy of the estimated probabilities. The discrimination of a model is its ability to assign higher probabilities for the outcome to those observations that experience the outcome (11). A well-established measure of discrimination is the area under the receiver-operating characteristic (ROC) curve (12). The calibration of a model quantifies the accuracy of the estimated probabilities for the outcome. Several tests such as the Hosmer-Lemeshow (HL) test have been proposed to assess the calibration of a model (11). For comparing the ANN model and LR model, ROC curves were created and applied to compare the ANN model and the LR model. Discriminatory ability was assessed by calculating the area under the curve (AUC) from the ROC analysis.

For each pair of ANN and LR models (trained and tested on the same data sets), Hosmer-Lemeshow (H-L) statistics, AUC and accuracy rate were considered and compared by T-tests for the validation group (n= 55 patients).

Ethics

The study was approved by the Ethics Committee of Shahid Beheshti University of Medical Sciences (code: IR.SBMU.REC.1398.023, Tehran, Iran).

Results

A total of 551 (61% male, mean age=55.5 ± 13.3 years; age range=11 to 89 years) primary GBM patients from the four cohorts were divided into training (n = 441), testing (n = 55), and validation (n = 55) data sets. Table 1 shows a comprehensive overview of the demographic information of these patients. The

Table 1. Demographic information of four independent primary GBM datasets (n=551).

Demographic categories	CGGA (n=225)	TCGA (n=164)	CPTAC (n=99)	MAYO (n=63)
Age (Year)				
□ 60	51	81	41	27
≤ 60	174	83	58	36
Gender				
Female	87	58	44	29
Male	138	106	55	34
IDH-status				
IDH1-Mut	35	11	11	1
IDH1-WT	183	141	88	61
Radiotherapy				
Treated	181	52	NA	21
Un-treated	31	112	NA	42
Chemotherapy				
Treated	170	49	NA	20
Un-treated	41	115	NA	43
X1p19q_codeletion				
Codel	5	0	NA	NA
Non-codel	192	149	NA	NA
MGMTp_methylation				
methyalted	97	52	NA	NA
un-methyalted	107	71	NA	NA

characteristics of the primary GBM patients and their gene expression (n=31) can be seen in Supplementary file 1(.xlsx). The relationships between the predictor variables (input nodes), hidden variables (20 of them in a hidden stratum), and 15-month survival prediction (output nodes) are demonstrated in Figure 1.

The normalized importance scores show that the expression levels of FN1, ICAM1, MYD88, IL10, and CCL2 genes were important variables selected by the ANN model compared to other variables. The results are presented in Table 2. However, the LR analysis indicates that four variables (MMP9, MYD88, CDK4, and age) were negatively significant, while the IDH-status variable was positively significantly associated with the dependent variable in our LR model.

Time-dependent ROC analysis was used to evaluate the predictive ability of the two models in the validation phase. The AUC values for 15-month survival were 0.814 for the ANN model and 0.713 for the LR model (Figures 2 and 3). Hence, these models demonstrated good discriminatory power.

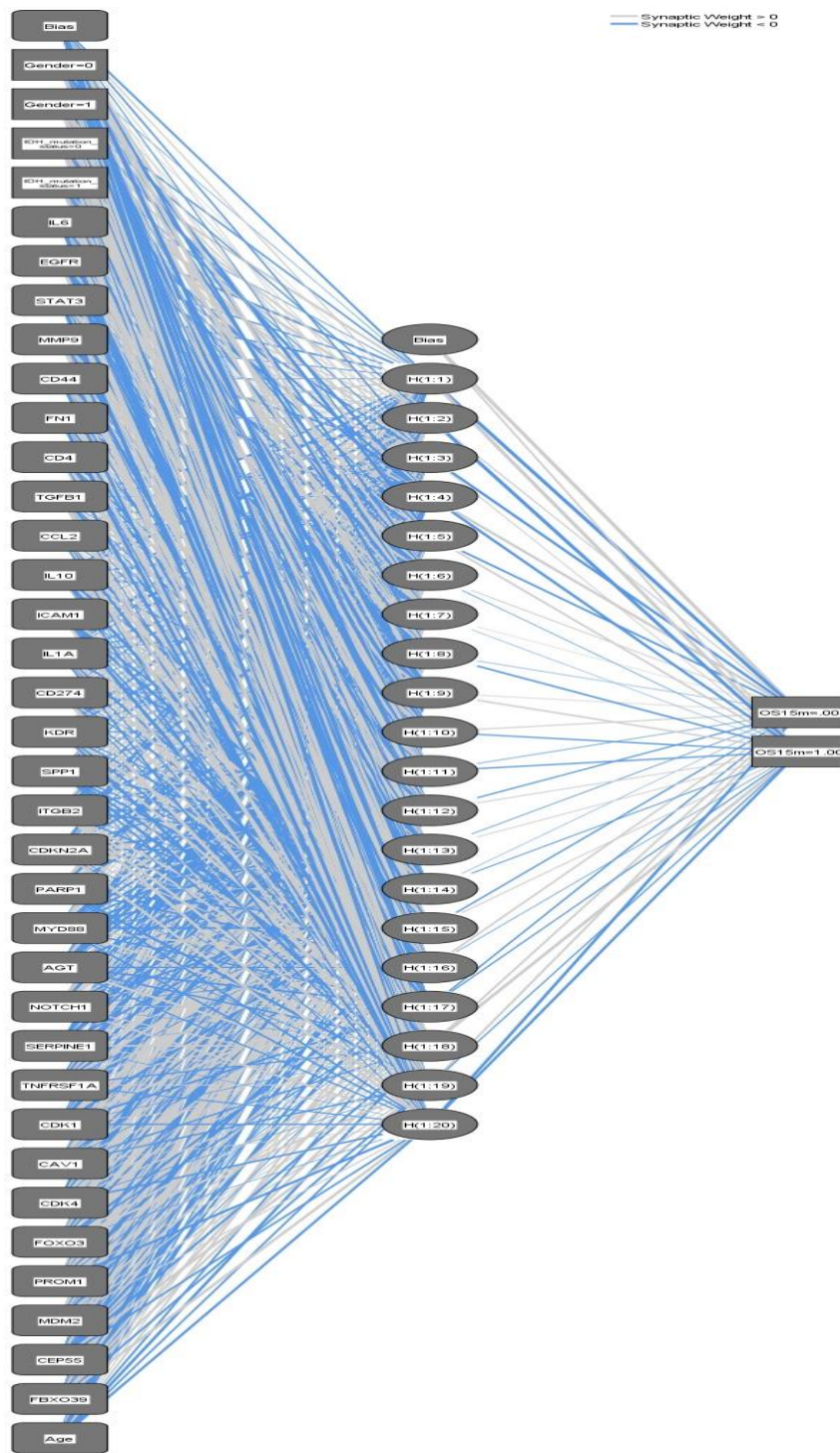


Fig.1. Artificial neural network output diagram with insets for each layer. Output figure created using SPSS v26. Input layer: Bias, input layer bias; IL6, EGFR, STAT3, MMP9, CD44, FN1, CD4, TGFB1, CXCL8, CCL2, IL10, ICAM1, IL1A, CD274, KDR, SPP1, ITGB2, CDKN2A, PARP1, MYD88, AGT, NOTCH1, SERPINE1, TNFRSF1A, CDK1, CAV1, ITGB3, CDK4, FOXO3, MDM2, PROM1, CEP55, FBXO39, Age; Gender, and IDH. Hidden layer, H (1:1), H (1:2), H (1:3), H(1:4) H (1:5), H (1:6), H (1:7), H (1:8), H (1:9), H (1:10), H (1:11), H (1:12), H (1:13), H (1:14), H (1:15), H (1:16), H (1:17), H (1:18), H (1:19), and H (1:20); Bias, hidden layer bias. Output layer: 15-month survival prediction in GBM patients.

Table 2. Normalized importance scores for top predictor variables for ANNs model.

Variables	Importance	Normalized importance
FN1	.066	100.0%
ICAM1	.063	95.3%
MYD88	.055	83.5%
IL10	.052	78.1%
CCL2	.049	73.5%
TNFRSF1A	.044	66.6%
Age	.042	63.7%
MMP9	.041	62.5%
CD4	.037	56.7%
AGT	.034	51.9%
FOXO3	.032	48.7%
ITGB2	.029	44.5%
IDH_mutation_status	.029	43.7%
EGFR	.028	42.9%
CDK4	.028	42.5%
PROM1	.028	42.3%
CEP55	.028	41.9%
CDK1	.027	41.4%
IL6	.024	35.9%
FBXO39	.023	34.9%
CAV1	.023	34.8%
SPP1	.023	34.3%
MDM2	.021	31.3%
CDKN2A	.019	29.4%
KDR	.019	29.4%
TGFB1	.018	27.9%
NOTCH1	.017	26.2%
STAT3	.017	25.8%
CD274	.016	24.3%
SERPINE1	.016	23.8%
PARP1	.015	22.9%
IL1A	.015	22.8%
CD44	.014	20.8%
Gender	.006	9.2%

The results of the comparison between the ANN model and the LR model are shown in Table 3. Compared to the LR model, the ANN model had a better accuracy rate in 83.3 % of patients, a better H-L statistic in 6.5 % of patients and a better AUC in 0.81 of patients.

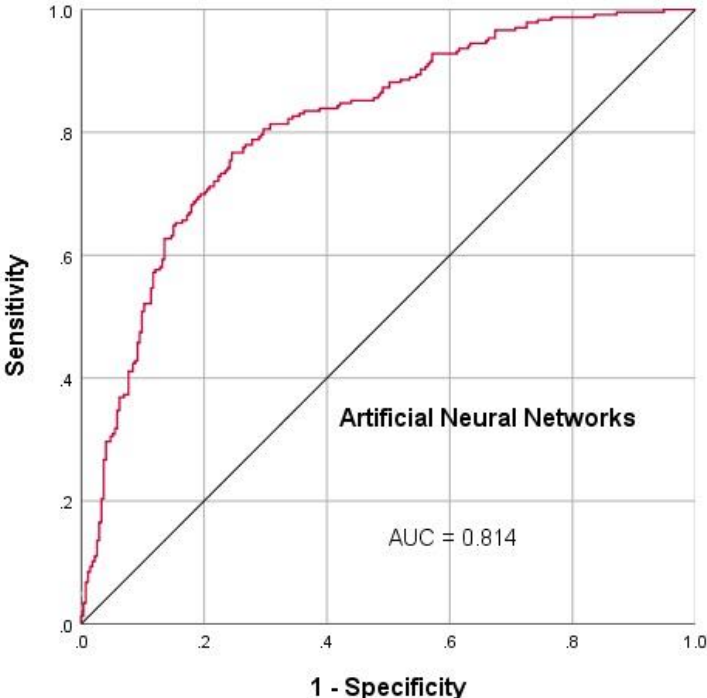


Fig.2. Valuation of the 15-month survival prediction ability of the ANN model by time-dependent ROC analysis at the validation set.

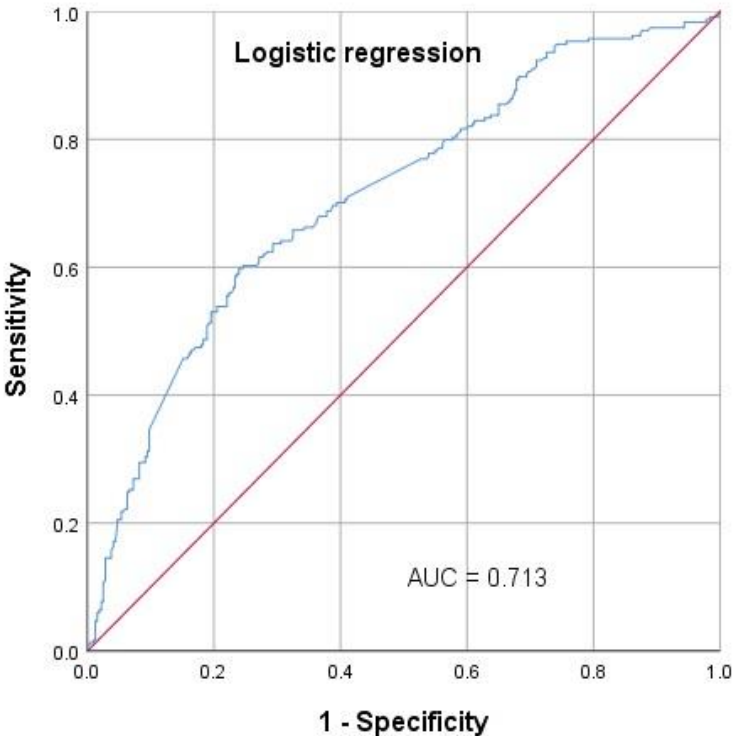


Fig.3. Valuation of the 15-month survival prediction ability of the LR model by time-dependent ROC analysis at the validation set.

[DOI: 10.22088/IJMCM.BUMS.13.1.79] [Downloaded from ijcmmed.org on 2024-10-06]

Table 3. Comparison of ANN and LR models to 15-month survival prediction in GBM patients (n=55*).

	ANN (95% C.I.)	LR (95% C.I.)	P value
Accuracy rate (%)	83.3 (80.1- 85.2)	67.7 (66.4- 74.9)	< 0.001
AUC	0.81 (0.79-0.88)	0.71 (0.67- 0.76)	< 0.001
H-L statistics	6.5 (5.1- 8.2)	10.1 (9.1- 12.4)	< 0.001

ANN =artificial neural network; LR = logistic regression; Hosmer- Lemeshow statistics = H-L statistics; AUC = area under the receiver operating characteristic. * Patients of the validation group

Discussion

The aim of the present study was to establish an ANN model from the dataset that integrates demographic and 31 gene variables to predict 15-month survival in GBM patients. Meanwhile, we compared the performance of the ANN algorithm with the LR model to evaluate which method provides better predictive performance. This study showed that the ANN model could be used to estimate 15-month survival prediction in GBM patients with a high accuracy.

To the best of the researcher's knowledge, there is no study that has analyzed the prediction of 15-month survival in GBM patients based on the ANN model using gene-expression profiles for the input layer of the ANN. In our ANN model, the normalized importance scores range from 9.2% (gender) to 100% (FN1), with 100% being the most significant predictor. The findings show that the genes FN1, ICAM1, MYD88, IL10 and CCL2, have a relatively greater impact on the predictions of the ANN model compared to other variables. Table 2 shows the important values. In addition, the ANN model had an AUC of 0.814, which was significantly higher than that of the LR model and demonstrated good discriminatory power. Therefore, the results can help to support clinical decision making and improve patient outcomes.

Many different methods have been developed to predict the survival of GBM patients. One of these is the gene risk score (GRS) model, which uses gene expression data to predict prognosis. In this regard, several GRS models with different types and numbers of genes have been presented in the literature to predict survival prognosis in these patients with different durations (13-15). However, there is little evidence on whether these GRS models achieve equality in different gene expression data sets and their objects (15). On the other hand, different machine-learning models have been developed to make accurate predictions about the survival of patients with GBM (16-19). The data in these studies came from the Surveillance, Epidemiology, and End Results (SEER) database. They used input data such as overall survival, age, gender, race, laterality, primary site, vital status, surgery, tumor size at diagnosis, and follow-up time in their machine learning models (16-19). These survival prediction studies reported accuracy rates between 70.0% and 90.66%, with different durations between 6 and 24 months (16-19), which is consistent with our study, whose accuracy rate is 83.3%. However, our ANN model is applied with different input variables. It should be noted that the best accuracy for survival prediction in neural networks depends on comprehensive data sources including histologic imaging, genomic molecular profiles, clinical data, and the type of machine learning. Moreover, the more data, the better (5). In the future, the ANN model will be developed with a combination of the aforementioned data and additional parameters for the input layer of the ANN (5).

However, the ANN model presented in the current study is an acceptable method for predicting 15-month survival in GBM patients.

One might wonder how the ANNs can help in clinical practice. It should be noted that ANN tools will never replace human experts, but they help with screening and can be used by experts to validate their diagnosis, prognosis and prediction. Despite the many attractive applications of our ANN model in clinical practice, several challenges still need to be overcome before it can be used as a support in healthcare. These include largely heterogeneous study design, the independence of ANN on probabilistic distribution, comparable quality standards in data collection across hospitals, the need for ANN algorithms to have access to sufficient amounts of data, data analysis, modeling technique, training and testing functions applied, ANN algorithms used, multidisciplinary teams including clinicians and ANN experts, and interpretation and clinical applicability of results. Although challenges of future application are addressed and corrected, the current ANN algorithms may provide an excellent framework for future developments and applications of ANN in clinical practice (5-7).

Despite the strengths of the study, there are some limitations. First, the sample size in this study is small and ANN algorithms need a large amount of training data to be effective. Second, it is important to consider that there may be other factors that can affect the prediction of 15-month survival in GBM patients that were not considered, such as imaging data, clinical data, and treatment approach. To improve the accuracy of prediction models, it may be valuable to combine different types of data and use multimodal approaches. Third, dependencies between different covariates are not considered in ANN model interpretation methods, which may lead to a correlation bias. Fourth, larger and more complete data sets could be applied to further clarify the differences between the LR and ANN models in terms of predicting GBM patient survival. However, it is expected that machine learning can be effectively used in real clinical practice in the near future with the help of high-quality neural network studies and incorporating optimal solutions. Compared with the LR method, the use of ANN prediction models can potentially improve clinicians' decision-making ability, and improve disease prognosis management and patient care.

Abbreviations

GBM: Glioblastoma

CNS: central nervous system

WHO: World Health Organization

IDH: isocitrate dehydrogenase

ROC: receiver operating characteristic

AUC: area under the curve

Mut: Mutant

OS: Overall Survival

TCGA: The Cancer Genome Atlas

CGGA: Chinese Glioma Genome Atlas

CPTAC: Clinical Proteomic Tumor Analysis Consortium

MAYO-PDX: Mayo Clinic Brain Tumor Patient-Derived Xenograft National Resource

Acknowledgments

The authors would like to thank the National Institute for Medical Research Development (NIMAD) for their support throughout the research process.

Ethics approval and consent to participate

The study was approved by the Ethics Committee of Shahid Beheshti University of Medical Sciences (Code: IR.SBMU.REC.1398.023, Tehran, Iran).

References

1. Grochans S, Cybulska AM, Siminska D, et al. Epidemiology of Glioblastoma Multiforme-Literature Review. *Cancers (Basel)* 2022;14.
2. Rong L, Li N, Zhang Z. Emerging therapies for glioblastoma: current state and future directions. *J Exp Clin Cancer Res* 2022;41:142.
3. Quazi S. Artificial intelligence and machine learning in precision and genomic medicine. *Med Oncol* 2022;39:120.
4. Azimi P, Yazdani T, Ahmadiani A. mRNA markers for survival prediction in glioblastoma multiforme patients: a systematic review with bioinformatic analyses. *BMC Cancer* 2024;24:612.
5. Azimi P, Mohammadi HR, Benzel EC, et al. Artificial neural networks in neurosurgery. *J Neurol Neurosurg Psychiatry* 2015;86:251-6.
6. Azimi P, Mohammadi HR. Predicting endoscopic third ventriculostomy success in childhood hydrocephalus: an artificial neural network analysis. *J Neurosurg Pediatr* 2014;13:426-32.
7. Azimi P, Yazdani T, Benzel EC, et al. A Review on the Use of Artificial Intelligence in Spinal Diseases. *Asian Spine J* 2020;14:543-71.
8. Vaubel RA, Tian S, Remonde D, et al. Genomic and Phenotypic Characterization of a Broad Panel of Patient-Derived Xenografts Reflects the Diversity of Glioblastoma. *Clin Cancer Res* 2020;26:1094-104.
9. Shim K, Jo H, Jeoung D. Cancer/Testis Antigens as Targets for RNA-Based Anticancer Therapy. *Int J Mol Sci* 2023;24:14679.
10. Schober P, Vetter TR. Logistic Regression in Medical Research. *Anesth Analg* 2021;132:365-6.
11. Nattino G, Pennell ML, Lemeshow S. Assessing the goodness of fit of logistic regression models in large samples: A modification of the Hosmer-Lemeshow test. *Biometrics* 2020;76:549-60.
12. Metz CE. Receiver operating characteristic analysis: a tool for the quantitative evaluation of observer performance and imaging systems. *J Am Coll Radiol* 2006;3:413-22.
13. Prasad B, Tian Y, Li X. Large-Scale Analysis Reveals Gene Signature for Survival Prediction in Primary Glioblastoma. *Mol Neurobiol* 2020;57:5235-46.
14. Yin W, Tang G, Zhou Q, et al. Expression Profile Analysis Identifies a Novel Five-Gene Signature to Improve Prognosis Prediction of Glioblastoma. *Front Genet* 2019;10:419.
15. Azimi P, Ahmadiani AH. Construction and evaluation of different glioblastoma prognosis scores based on gene expression databases. *Physiology and Pharmacology: May 28, 2024, ahead of print.*
16. Babaei Rikan S, Sorayaie Azar A, Naemi A, et al. Survival prediction of glioblastoma patients using modern deep learning and machine learning techniques. *Sci Rep* 2024;14:2371.
17. Samara KA, Al Aghbari Z, Abusafia A. GLIMPSE: a glioblastoma prognostication model using ensemble learning-a surveillance, epidemiology, and end results study. *Health Inf Sci Syst* 2021;9:5.

18. Bakırarar B, Egemen E, Dere ÜA, et al. Machine learning model to identify prognostic factors in glioblastoma: A SEER-based analysis. *Pamukkale M J* 2022;16:338-48.
19. Ghanem M, Ghaith AK, Zamanian C, et al. Deep Learning Approaches for Glioblastoma Prognosis in Resource-Limited Settings: A Study Using Basic Patient Demographic, Clinical, and Surgical Inputs. *World Neurosurg* 2023;175:e1089-e109.