Peshnaja: a framework for predicting survivability of glioblastoma patients using ML and SEER data

Aleema Ashfaq[1], Bilal Wajid[1,2], Faria Anwar[3], Fahim Gohar Awan[2], Ali Anwar[4], Muhammad Ali Subhani[1], Imran Wajid[5], and Abdul Rauf Anwar[6]
[1]Ibn Sina Research & Development Division, Sabz-Qalam, Lahore 54000, Pakistan
[2]Department of Electrical Engineering, University of Engineering & Technology (UET), New Campus, Lahore, Pakistan.
[3]Out-Patient Department, Mayo Hospital, Lahore, Pakistan.
[4]Department of Computer Science & Engineering, University of Minnesota, Minneapolis, MN 55455, USA.
[5]School of Social Sciences, Istanbul Commerce University, Istanbul, Turkey.
[6]CENIR MEG-EEG, Paris Brain Institute, Paris, France.

Abstract — **Glioblastoma is a common and fatal tumor presenting a poor survival rate. To choose the best course of treatment, patients and providers need to predict the survival rate of patients. Historically, statistical methods have helped analyze clinical features to forecast survival, while recently the same is being accomplished by applying artificial intelligence techniques. However, most of these works are limited to predicting 1-, 2-, or 10-year survivability with several of these works simulating data for balancing the dataset. Hence, there is a need for fine-grained prognosis without tempering the data. To achieve the same, we employ data from Surveillance, Epidemiology, and End Results (SEER) along with an ensemble of classification and regression models to develop a fine-grained model to predict the survival period of glioblastoma patients. The proposed framework titled 'Peshnaja' presents higher resolution in the prognosis of glioblastoma while showcasing an accuracy of 70% with an overall RMSE of 2.65. Moreover, a comparison of Peshnaja with other frameworks shows that we did not impute missing values nor employed synthetic data to force good results, thereby keeping Peshnaja true to the existing data.**

Keywords: Cancer survival analysis, Glioblastoma, Machine learning, Classification, survival rate, SEER

I. INTRODUCTION

Glioblastoma (brain tumors) are mostly secondary tumors, with primary tumors occurring within lungs, breast, colon, kidney, or melanoma [1]. Patients affected by brain cancer exhibit unregulated cell growth resulting in headaches, reduced hearing and vision, seizures, and difficulty maintaining balance [2]. As of today (July 2023), the 5-year survival for Brain Cancer patients stands at a meager 36%, as it is difficult to treat [3].

Brain tumors are usually diagnosed using magnetic resonance imaging (MRI) which helps in not only diagnosing the disease but also its subtypes, prognosis, and management. However, MRIs are not fruitful for predicting the survival of a patient [4]. Therefore, scientists are moving towards cohort studies to predict survival periods. One such database is provided Surveillance, Epidemiology, and End Results (SEER) which details cancer incidence and survival in the USA [5]. Researchers are not only improving the survivability of brain cancer by applying novel Machine Learning approaches on these databases to obtain accurate results, but also concentrating their efforts on developing a finer-grained predictive model to anticipate how long the patient will survive.

II. LITERATURE REVIEW

Fewer studies have employed SEER data for predicting the survivability of glioblastoma patients. Earlier works employ traditional statistical methods for survival analysis. In 2019, Senders [6] designed an online calculator for 1-year survival prediction of Glioblastoma patients using Cox proportional hazards regression (CPHR) and the accelerated failure time (AFT) algorithms. CPHR with the best performance in terms of discrimination (concordance index= 0.70) was deployed in the calculator.

In 2020, another study by Yang [7] developed nomograms to predict 3-year and 5-year survival using Univariate and multivariate Cox analysis models with 0.759 and 0.768 concordance indexes respectively. Although statistical methods help in understanding the relationship between a limited number of variables, they cannot effectively handle the complex datasets required to generate diagnostic and prediction models that can improve clinical decision-making [8].

From 2021 onwards, there was a substantial increase in the use of supervised machine learning and deep learning algorithms offering reliable prediction results across complex variables. A study done in 2021 by Samra [9] recommends a framework to determine (i) short (6 months with 86% accuracy), (ii) intermediate (12 - 18 months with 70% accuracy), and (iii) long-term survival (2 years with 81% accuracy) based on SEER data. Even though Samra presents good accuracy, they employed simulated data to balance short, intermediate, and long-term survival classes, which are not recommended in clinical studies.

In 2022, another researcher B. Bakirarar [10] proposed a hybrid model for predicting one-year survival, with 74% accuracy, and two-year survival with 84% accuracy.

This year, 2023, another work presented by G. Nath [11] improved accuracies for three survivability periods, i.e., 1 year (89%), 5 years (90%), and 10 years (92% accuracy). However, this work also employed the use of simulated data for balancing different classes, which is not recommended in clinical studies.

Again, a paper by G. Nath [12] presents a web-based tool for predicting 10-year survivability of glioblastoma patients with an accuracy of 98%. However, as this work also employs simulated data, and only presents a binary classification framework with survivability less or greater than 10 years, the work does not hold any clinical merit.

TABLE I. DETAILED COMPARISON CARRIED OUT ENCOMPASSING STUDIES REPORTING SURVIVAL PREDICTION OF GLIOBLASTOMA PATIENTS USING SEER DATA. NOTE: LAST TWO STUDIES DO NOT REPORT ACCURACY SCORES SINCE THEY CONDUCTED STATISTICAL ANALYSIS

Year	Ref.	Time Frame taken	Registries	Survival Prediction Classes	Use of simulated date	Feature selection	Parameters	Accuracy score
2023	[12]	1975-2018	1 – 18	10-year (120 months)	SMOTE	20 features using LASSO	Accuracy, AUC	98.9%
2022	[11]	1975-2018	1 – 9	12, 60, 120 months	SMOTE for train data only	12 features using Anova, RF, Sequential Forward search	AUC	73.1
2022	[10]	2007-2018	1 – 18	12, 24 months	--	7 features using Kaplan Mier method	Accuracy, F measure, MCC	84% and 74%
2021	[9]	2004-2015	1 – 17	6, 12, 18, 24 months	Miss Forest, SMOTE	31 features using literature and RF	AUROC, accuracy, sensitivity, specifity	86%
2020	[7]	2004 – 2015	1-17	3-year, 5-year	--	Univariate and multivariate Cox regression model	3-year (c-index=0.759), 5-year (c-index=0.768	--
2019	[6]	2005-2015	1 – 18	12 months	Missing data imputed	--	C-index=0.70	--

III. RESEARCH GAP

As highlighted, most of these works convert their target variable ("Survival months") into binary classification, with the majority opting for 1 or 2 – year survivability, while others even carry out 10 – year survival prediction. Moreover, several of these works employ simulated data via balancing techniques like SMOTE, and RUS.

The collective use of bins, along with simulated data to present higher accuracies renders most of these works irrelevant from a clinical standpoint. Therefore, this humble effort aims to employ SEER data to develop a clinically useful, fine – grained model that helps doctors assist their patients by predicting the survival of glioblastoma patients, all done, without simulating data to fill – in missing entries.

IV. METHODS

[1] **Data**: Data was derived from SEER *[13]*. Specifically, through its “Case listing sessions,” we obtained 166,516 records from Cancer incidence database comprising 22 registries ranging from 2000 – 2021, with a total of 78 features.

[2] **Preprocessing and Filtering:** Features pertaining to brain cancer were retained, while attributes corresponding to other cancers were discarded. The following steps were conducted:

(a) Removed records containing empty features (blank/NA).
(b) Discarded data points with unknown survival months.
(c) Excluded features unrelated to glioblastoma.
(d) Retained one of each duplicate feature. For instance, “Primary site,” and “Primary site labeled” are duplicate features, where one is numeric and the other is categorical, both providing the same information.
(e) Converted categorical features into equivalent numeric values.
(f) Eliminated patient data, where the subject passed away for causes other than glioblastoma.
(g) Removed features which were constant in the entire dataset.

The above helped derive 9,960 records comprising the following 34 features:

TABLE II: LIST OF ALL FEATURES RETAINED AFTER PREPROCESSING. HERE ‘C’ STANDS FOR CATEGORICAL FEATURES WHILE ‘N’ DENOTES FEATURES THAT ARE NUMERIC

#	**Feature Attributes**			
	Feature Names	**Symbol**	**Feature Specification**	**Type**
				C/N
1.	Race and origin recode	O	It includes the five mutually exclusive race and ethnicity categories (NHW, NHB, NHAIAN, NHAPI, Hispanic) which SEER uses for reporting cancer statistics.	C
2.	Race	R_R	It is based on race variables: White (W), Black (B), American Indian/Alaska Native (AI), and Asian Pacific Islander (API)	C
3.	Race/ethnicity	R_E	The race of patients belonging to more diverse groups	C
4.	Aya site Recode	α	A site/histology recode that is used to analyze data on adolescent and young adults	C
5.	Sex	XY	Gender	C
6.	Primary Site - labeled	P	Sites where primary tumor originated. This provides the primary site code in ICD-O-3 and a descriptive primary site label.	C
7.	Grade Recode (thru 2017)	G	Appearances of cancer cells and how fast they may grow.	C
8.	Diagnostic Confirmation	T_T	Methods used to confirm the presence of brain cancer.	C
9.	Laterality	L	Describes the site of the body on which the reportable tumor originated.	C
10.	Combined Summary Stage 2004	S_{04}	A descriptor of the extent cancer has spread, taking into account the size of the tumor, depth of penetration, metastasis	C
11.	RX Summ – Surg Prim Site	$R_{X(P)}$	Surgical procedure to remove or destroy tissues of the primary site.	C
12.	RX Summ – Systemic/Sur Seq (2007)	$R_{X(07)}$	The sequence of any systemic therapy and surgery given as first course of therapy for those patients who had both systemic therapy and surgery.	C
13.	RX Summ – Surg/Rad Seq (2006)	$R_{X(06)}$	The order in which surgery and radiation therapies were administered for those patients who had both surgery and radiation.	C
14.	COD to site rec KM	C_{KM}	This is a recode based on underlying cause of death to designate cause of death into groups similar to the incidence site recode with KS and mesothelioma	C
15.	Radiation recode (2003)	γ	Records the type of Radiation treatment delivered	C
16.	Chemotherapy recode (2004)	C_{04}	Records the chemotherapy given as a part of the first course of treatment or the reason that chemotherapy was not given i.e., (yes, no/unk).	C
17.	EOD Schema ID Recode (2010)	E	Extent of Disease (EOD) is a set of three data items that describe how far a cancer has spread at the time of diagnosis.	C
18.	ICCC site recode extended 3rd edition/IARC 2017	$ICCC$	A site/histology recode that is mainly used to analyze data on children	C
19.	SEER Brain and CNS Recode	S_β	To analyze brain tumors according to major histological categories.	C

#	Feature Attributes			
	Feature Names	Symbol	Feature Specification	Type
				C/N
20.	Site recode - rare tumors	C_R	Table of recodes for rare cancer sites.	C
21.	ICD-O-3 Hist/behav	B	Definitions of major cancer sites based on the primary site and histology.	C
22.	Histologic Type ICD-O-3	H_{ICD}	Microscopic composition of cells and/or tissues for specific primary. Used for staging and treatment determination	C
23.	IHS Link	I	Indian Health Service (IHS) Link reports the result of linkage between the registry database and the Indian Health Service patient registration database	C
24.	Reason for no surgery	$R_{\bar{S}}$	The reason why surgery was not performed (if not)	C
25.	PRCDA 2020	P	This data item identifies whether or not the county of diagnosis is served by Purchased/Referred Care Delivery Area (PRCDA)	C
26.	Survival months	S	Number of months that patient is alive from date of diagnosis.	N
27.	Months from diagnosis to treatment	M	Estimates month of diagnosis to treatment, based on other known dates for that patient, when actual month of diagnosis is unknown.	N
28.	Age recode with single ages and 90	A	Age at time of diagnosis.	N
29.	Total number of in situ/malignant tumors for patient	T_M	The number of malignant tumors in the patient’s lifetime	N
30.	Total number of benign/borderline tumors for patient	T_B	The number of benign tumors in the patient’s lifetime	N
31.	Year of follow-up recode	F	Records year of last follow up	N
32.	Year of death recode	Y_Y	Records year of Death	N
33.	Year of diagnosis	Y_D	The year of diagnosis is the year the tumor was first diagnosed by a recognized medical practitioner, whether clinically or microscopically confirmed.	N
34.	SS seq # 2000 - mal ins (most detail)	SS	Site specific sequence number of the tumor associated with the site classification scheme in the variable Site - mal+ins (most detail). Based on all the tumors in SEER.	N

[3] **Binning Target variable:** We chose “survival months” as the target variable for each patient.

(a) For experiment 1, we employed regression, thereby using ‘survival months’ as-is.
(b) For experiment 2, we binned the 'survival months' into three classes (i) 0-7 months, (ii) 8 - 19 months, and (iii) 19 - 140 months. Equal distribution was ensured.
(c) In experiment 3, we employed ensemble learning, using cart-classification approach, wherein each layer comprised of binary classification, see figure 7.

[4] **Normalization**: We first performed 'Min-Max Scaler' normalization on the dataset.

[5] **Train – test – validation split:** We split the data into two parts i.e., (i) 90% as the combined train/test set, and (ii) 10% as the validation set. Thereafter, we used the train/test split using 10× cross-validation. These splits were conducted using random sampling.

[6] **Feature Selection**: In experiment 1, we conducted both sequential forward and sequential backward search. The performance graph shown below shows that beyond 3 features there is hardly any difference in performance indicated by the graph’s straight line. Therefore, we decided to retain the first 5 features for further processing.

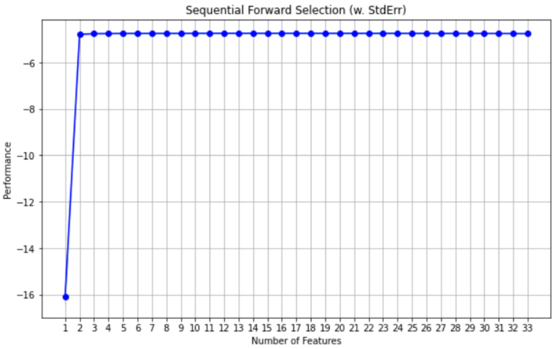

Figure 1: Performance Graph for Sequential Forward search

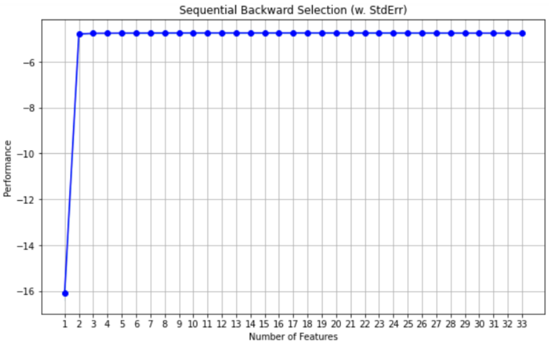

Figure 2: Performance Graph for Sequential Backward search

However, in experiments 2, we employed the Random Forest Classifier (RFC) to determine the top 15 features as per their “feature importance scores.” In experiment 3, RFC was again employed for Layers 1 and 2 during classification, while carrying out sequential searches for Layer 3.

[7] **Machine learning Models**: We applied several tools to the dataset.

(a) For classification, we applied (1) Logistic Regression (LR), (2) Decision tree (DT), (3) Linear discriminant analysis (LDA), (4) Support vector machine (SVM), (5) Gradient Boost classifier (GB), (6) Random Forest classifier (RFC), (7) Gaussian Naïve Bayes (GNB), (8) Perceptron (Perc), (9) Ad-booster (AB), (10) Quadratic Discriminant (QDA), (11) One vs. rest classifier using Logistic Regression (OvR-LR), (12) One vs. rest classifier using perceptron (OvR-P), (13) Multi-Layer Perceptron (MLP), (14) Artificial Neural Network (ANN), and (15) Voting classifier (Vote).

(b) For regression, we applied (1) Linear Regression (LR), (2) SVM, (3) Ridge regression (RR), (4) Elastic-net regression (Elas), (5) LASSO, (6) LARS, (7) Automatic Relevance Determination (ARD), (8) Gradient Boosting regression (GBR), (9) Stochastic Gradient Descent (SGD), and (10) Random Forest regression (RFR), (11) Decision Tree, and (12) Linear SVM.

[8] **Measurements:** As the train/test set was employed using 10x Cross-Validation, therefore, our results report the average accuracy for the training and test sets. Whereas we report the accuracy, as-is, for the validation set. As for regression, we report RMSE.

V. RESULTS

We conducted the following three experiments to devise Peshnaja:

[1] **Exp. 1: Regression**: The first experiment involves applying regression models, and then comparing them as per their RMSE values. Hereinbelow, table III highlights the performance of the models with respect to the training, test, and validation sets. As shown in Table III, as the overall RMSE values are large, we need to try a different approach for obtaining a better solution.

TABLE III: The table shows the performance of different regression models when applied to SEER dataset. For ease, the table is sorted as per the RMSE values of the validation set with the top model at #1.

S. No.	Models	RMSE scores for Regression	
		Train set	**Validation set**
1.	Gradient Boosting	4.33	4.32
2.	Random Forest	1.85	4.52
3.	Ridge regression	4.73	4.79
4.	Linear Regression	4.73	4.79
5.	ARD regression	4.73	4.79
6.	LASSO Regression	4.74	4.79
7.	SGD Regression	4.74	4.80

S. No.	Models	RMSE scores for Regression	
		Train set	Validation set
8.	Linear SVR	4.94	5.10
9.	Decision Tree Regression	0.51	5.93
10.	Elastic net	6.32	6.73
11.	SVR	5.88	6.82
12.	LARS regression	17.02	18.31

[2] **Exp. 2: Multi-Class Classification:** As regression models in experiment 1 presented large RMSE values, we opted to bin 'survival months' as equally as possible into three bins i.e., (a) 0 - 7 months, (b) 8 - 19 months, and (c) 20-140 months, and try a 3-class classification framework.

We applied different classification models on the training/test set (90% of the data) using 10x cross-validation and verified these models on the validation set (remaining 10% of the data). The results are shown in Table IV. As for the voting classifier, we took the top 3 models and used them together. Upon inspecting the confusion matrices shown in Table V, we conclude that the classifiers exhibit a bias for class 1 (0 - 7 months) over classes 2 and 3, due to unbalanced data, as shown in Figure 3.

TABLE IV: The table shows the performance of different classification models when applied to SEER dataset. For ease, the table is sorted as per the accuracy score of the validation set with the top model at #1.

S. No.	Models	Accuracy scores	
		Train set	Validation set
1.	Gradient Boost	59	61
2.	AdaBoost	59	60
3.	Multi-Layer Perceptron	47	58
4.	Random Forest	56	57
5.	Support Vector Machine	56	56
6.	Logistic Regression	56	56
7.	Linear Discriminatinant Analysis	56	56
8.	One vs Rest using LR	55	56
9.	Gaussian Bayes	54	55
10.	Perceptron	44	52
11.	Decision Tree	49	49
12.	Quadratic Discriminant Analysis	45	43
13.	One vs Rest using Perceptron	46	43
14.	**Voting Classifier**	**59**	**60**

TABLE V: Confusion matrices for the validation set reveal, (i) the classifiers favor class 1 over classes 2 and 3, and (ii) the performances of the classifiers for classes 2 and 3 is poor. (Top 3)

Models	Gradient Boost	AdaBoost	MLP	*Voting Classifier*
Confusion Matrices	$\begin{bmatrix} 267 & 82 & 16 \\ 194 & 177 & 61 \\ 37 & 95 & 167 \end{bmatrix}$	$\begin{bmatrix} 260 & 89 & 16 \\ 97 & 169 & 66 \\ 39 & 90 & 170 \end{bmatrix}$	$\begin{bmatrix} 284 & 80 & 1 \\ 119 & 197 & 16 \\ 46 & 147 & 106 \end{bmatrix}$	$\begin{bmatrix} 265 & 85 & 15 \\ 98 & 172 & 62 \\ 38 & 93 & 168 \end{bmatrix}$

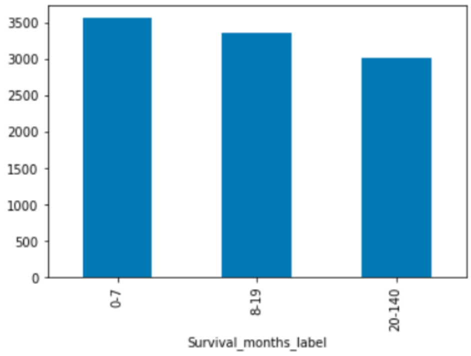

Figure 3: Distribution of records in 3 classes reveal (i) most glioblastoma patients die within the first year of diagnosis, and (ii) the data is unbalanced.

[3] **Exp. 3: Multi-layer Classification:** To improve the previous solution, we opted to undertake multi-layered classification, ensuring that classes within each layer contain the same number of patients.

(a) **Layer 1:** Data is first classified into 2 classes, i.e., (a) '0 - 12' months, and (b) 'above 12 months'. This ensures an (almost) equal distribution of records within each class, as shown in Figure 4. Tables VI and VII present the results of the classification models when applied to layer 1. As shown in Table VI, SVC, ANN, LOG, and ONE-Log are all equivalent presenting 90% accuracy on the validation set. Moreover, as shown in Table VII, confusion matrices do not exhibit the bias shown in the previous experiment.

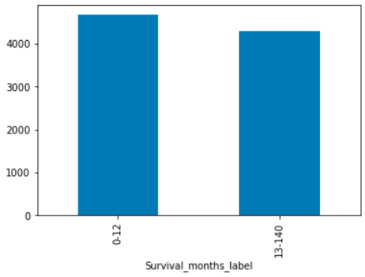

Figure 4: Equal Distribution of Records during Layer 1 Classification as Class A (0-12 months) and Class B (13-140 months)

TABLE VI. LAYER 1 CLASSIFICATION RESULTS FOR CLASS A AND CLASS B

S. No.	**Models**	**Accuracies scores**	
		Train set	**Validation set**
1.	**Support Vector Classifier**	**87**	**90**
2.	**ANN**	**87**	**90**
3.	**Logistic Regression**	**87**	**90**
4.	**One vs Rest using Logistic regression**	**87**	**90**
5.	Gradient Boost	87	89
6.	AdaBoost	86	88
7.	Linear Discriminatinant Analysis	85	88
8.	Perceptron	81	86
9.	One vs Rest using Perceptron	81	86
10.	Random Forest	84	85
11.	Multi-Layer Perceptron	85	84
12.	Decision Tree	83	84
13.	Quadratic Discriminant Analysis	65	79
14.	Gaussian Bayes	70	72
15.	**Voting classifier**	**87**	**90**

TABLE VII. Top 3. COMPARISON OF CONFUSION MATRIXES OF CLASSIFICATION MODELS

Models	SVC	ANN	Log	Voting
Confusion Matrices	$\begin{bmatrix}527 & 2\\ 93 & 374\end{bmatrix}$	$\begin{bmatrix}505 & 24\\ 76 & 391\end{bmatrix}$	$\begin{bmatrix}511 & 18\\ 80 & 387\end{bmatrix}$	$\begin{bmatrix}519 & 10\\ 84 & 383\end{bmatrix}$

(b) **Layer 2:** Classes (A) and (B) are further divided into two classes, as shown in Figure 8. Class A is further divided into Class 1 (0 – 5 months) and Class 2 (6 – 12 months) while Class B is divided into Class 3 (13 – 23 months) and Class 4 (24 – 140 months). Here again, the division of the classes ensures (almost) equal distribution of patients within each class shown in Figures 6 – 8. Finally, Class 4 is divided into Class 4A (24 – 36 months) and Class 4B (37 – 140 months). This additional step is performed to obtain survival prediction till 3 years at maximum, while survival periods above 3 years are grouped in one individual class (Class 4B). Table VIII presents the results of different classifiers when applied to Layer 2.

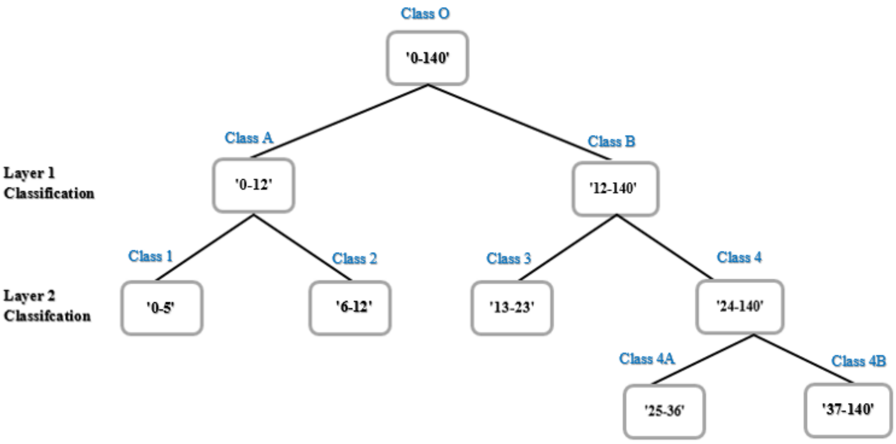

Figure 5: Distribution of Records in Layer 1 and Layer 2 involving Classification framework

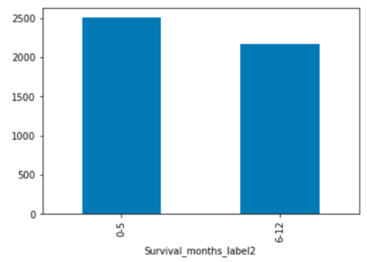

Figure 6: Distribution of Records in Class 1 and Class 2

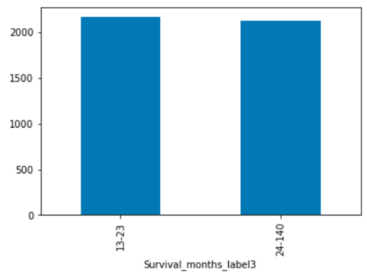

Figure 7: Distribution of Records in Class 3 and Class 4

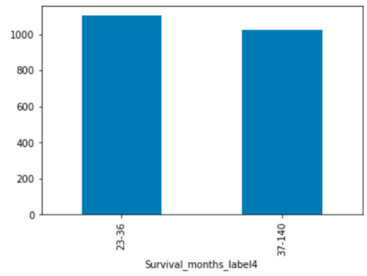

Figure 8: Distribution of Records in Class 4a and Class 4b

TABLE VIII. LAYER 2 CLASSIFICATION RESULTS WITH HIGHLIGHTED RESULTS SHOWING TOP 3 MODELS, FURTHER EMPLOYED FOR VOTING CLASSIFER

S. No.	**Models**	**Accuracy Scores**		
		Class 1 vs. 2	**Class 3 vs. 4**	**Class 5 vs. 6**
1.	Logistic Regression	78	79	83
2.	Decision Tree	64	74	83
3.	Linear Discriminatinant Analysis	79	81	83
4.	Support Vector Classifier	79	81	82
5.	Gaussian Bayes	68	59	66
6.	Random Forest	73	76	84
7.	Gradient Boost	77	80	83
8.	Perceptron	67	81	81
9.	AdaBoost	75	81	83

S. No.	Models	Accuracy Scores		
		Class 1 vs. 2	Class 3 vs. 4	Class 5 vs. 6
10.	Quadratic Discriminant Analysis	57	64	72
11.	One vs Rest using Logistic regression	78	79	83
12.	One vs Rest using Perceptron	67	81	81
13.	MLP	79	64	83
14.	ANN using Keras Classifier	79	80	83
15	**Performance of Voting Classifiers**	**79**	**81**	**83**

To obtain accuracies of Peshnaja, we first employed the top classifiers for each layer. For instance, we chose SVC for Layer 1 with an accuracy of 90% on the validation set, while presenting a confusion matrix with 527 true positives (TP), and 374 true negatives (TN).

Thereafter, these 527 TPs were fed to Class ‘1 vs. 2’ classification (in Layer 2), while the 374 TNs were fed to Class ‘3 vs. 4’ classification (also in Layer 2).

Finally, the TNs obtained from Class ‘3 vs. 4’ classification is directed to Class ‘4A vs. 4B’ classification. The step – by – step evaluation of accuracy is shown in Table (X-XIII).

TABLE IX. BEST MODEL FROM EACH CLASSIFCATION IN LAYER 2, ALONG WIH ITS ACCURACY SCORES AND CONFUSUON MATRICES

S. No.	Classification	Classifier	Accuracy	Confusion Matrix
1.	Class 1 vs. Class 2	LDA	79	$\begin{bmatrix} 251 & 47 \\ 60 & 169 \end{bmatrix}$
2.	Class 3 vs. Class 4	AB	81	$\begin{bmatrix} 133 & 0 \\ 69 & 172 \end{bmatrix}$
3.	Class 4A vs. Class 4B	RFC	84	$\begin{bmatrix} 37 & 5 \\ 22 & 108 \end{bmatrix}$

Table X: Layer 1 Binary Classification predicting survival period till 12 months (1 year), Accuracy = 90%

	0 – 12	13 – 140
0 – 12	527	2
13 – 140	93	374

Table XI: 527 TPs of Layer 1 are used in Class 1 vs. 2 Classification

	0 – 12		13 – 140
	0 – 5	6 – 12	
0 – 5	251	47	2
6 – 12	60	169	
13 – 140	93		374

Table XI: 374 TNs of Layer 1 were directed to classification in Layer 2 predicting survival period of either (0 – 5, 6 – 12) or (12 – 24, 25 – 140) months, with an accuracy of 72.7%

	0 – 12		13 – 140	
	0 – 5	6 – 12	13 – 24	25 – 140
0 – 5	251	47	2	
6 – 12	60	169		
13 – 24	93		133	0
25 – 140			69	172

Table XII: 172 TNs of Class 4A vs. 4B were fed for further classification into Class 4A vs. 4B, presenting an accuracy of 70.0%

	0 – 12		13 – 140		
	0 – 5	6 – 12	13 – 24	25 – 36	37 – 140
0 – 5	251	47	2		
6 – 12	60	169			
13 – 24	93		133	0	
25 – 36			69	37	5
37 – 140				22	108

$$\text{Peshnaja's Accuracy} = \frac{\sum(\mathrm{TP} + \mathrm{TN})}{\sum(\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN})} = \frac{251 + 169 + 133 + 37 + 108}{996} = \frac{698}{996} = 70.0\ \%$$

(c) **Layer 3**: Finally, to obtain a fine-tuned model, wherein we can predict the exact number of months a glioblastoma patient may survive, we applied different regression models separately to each class and results are recorded in Table XIV.

The accuracy score for the proposed Peshnaja framework comes to be 70.0%. However, the RMSE scores of the individual five classes show excellent scores, as shown above in Table XIV. The overall solution, exhibited in Figure 9, is a fine-tuned model capable of predicting the exact number of months a glioblastoma patient is expected to survive.

TABLE XIII. COMPARISON OF RMSE SCORES ON VALIDATION SETS FOR ALL CLASSES

#	**Models**	**Class 1**	**Class 2**	**Class 3**	**Class 4A**	**Class 4B**
1.	Linear Regression	1.45	1.88	3.18	2.78	4.76
2.	SVR	1.44	1.93	3.06	3.68	13.57
3.	Ridge regression	1.45	1.88	3.18	2.78	4.77
4.	Elastic net	1.45	1.94	3.05	3.50	6.81
5.	LASSO regression	1.45	1.90	3.08	2.83	4.75
6.	LARS regression	1.49	1.95	3.08	3.77	19.12
7.	Linear SVR	1.43	1.90	3.02	3.32	9.64
8.	ARD Regression	1.44	1.89	3.18	2.73	4.73
9.	Gradient Boosting	1.42	1.86	3.15	3.17	4.64
10.	SGD Regression	1.45	1.88	3.14	2.82	4.77
11.	Random Forest	1.57	2.04	3.21	3.50	4.78
12.	Decision Tree Regression	2.10	2.79	3.45	4.31	5.94

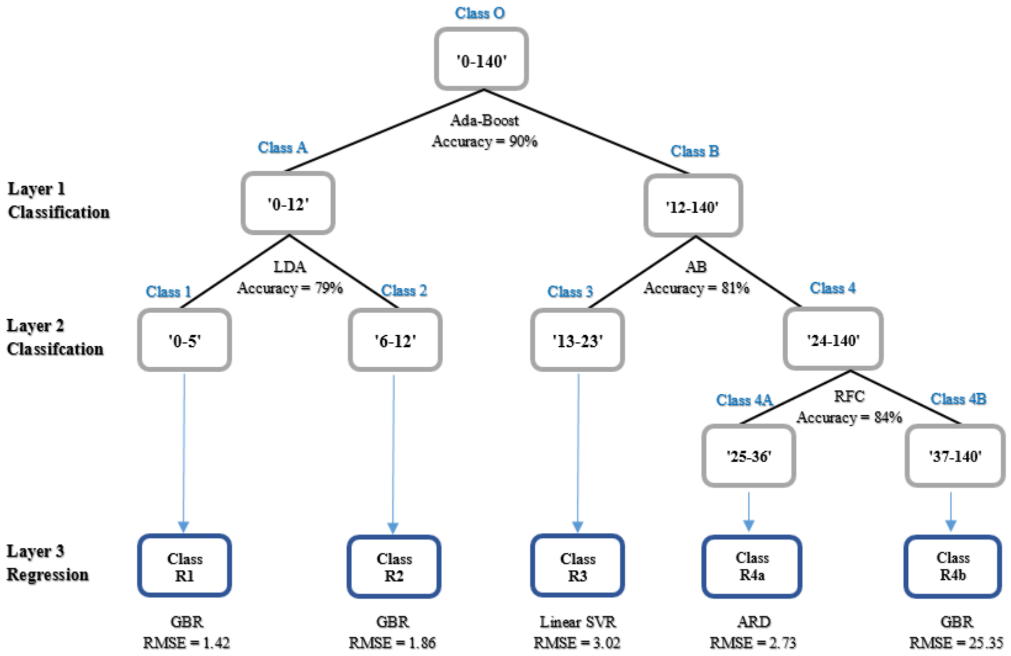

Figure 9: Complete Schematic diagram of proposed model showing its 3-layered architecture along with division of classes. Best Model along with their performance parameter is also mentioned

Finally, to show the goodness – of – fit of Peshnaja, hereinbelow, we took two samples from each class and compared Peshnaja's prediction with actual values to see its performance (see Table XV). The RMSE of Peshnaja is shown for different survival month periods. We can see that for the first year (0 – 12 months) Peshnaja shows excellent RMSE of 1.61, dipping slightly for the third year to 2.09, to final 2.65 RMSE for the 10 year (0 – 140 months) survival period, (see Table XVI).

TABLE XIV. COMPARISON BETWEEN ACTUAL VALUES AND PESHNAJA'S PREDICTED VALUES

Class	**Time period (months)**	**Model**	**Survival Months**	
			Actual Value	**Predicted Value**
Class 1	0 – 5	RFR	1	1.4
			4	3.22
Class 2	6 – 12	LR	6	8.5
			11	8.9
Class 3	13 – 24	SVM	18	17.30
			22	18.66
Class 4A	25 – 36	Linear SVM	28	31.3
			34	33.1
Class 4B	37 – 140	GBR	42	43.36
			89	92.16

TABLE XV. OVERALL RMSE SCORES WITH RESPECT TO DIFFERENT SURVIVAL PERIODS

S. No.	Survival months	RMSE scores
1	0 – 5	1.42
2	0 – 12	1.61
3	0 – 24	2.04
4	0 – 36	2.09
5	0 – 120	2.65

VI. Discussion

This paper introduces Peshnaja – a clinically relevant, finely–tuned framework for predicting the survivability of Glioblastoma patients.

We conducted three experiments to achieve the desired accuracy. The first experiment employed regression on the entire dataset resulting in high RMSE scores, proving that regression alone is unfit to provide a decent solution.

Secondly, we carried out a 3-ary classification with an equal distribution of data. The three classes (0 – 7 months, 7 – 19 months, and 20 – 140 months) present skewed class limits owing to the low survivability of glioblastoma patients. However, no model could fit the data well, showing poor accuracy, indicating that simplistic classification alone is unsuitable for this problem.

Lastly, we embarked on developing a prediction framework employing both classification and regression. The resulting ensemble learning model comprised a binary tree, wherein each level of the tree divided the data into equal sets, allowing us to try different classification models at each layer. Once suitable classes (survival periods) were made, we applied different regression models to each survival period to determine the exact possible survival month for each glioblastoma patient.

While developing Peshnaja, we (i) avoided filling in missing values using synthetic data, and (ii) did not forcefully balance classes to achieve equal class limits. The above two techniques, if done, improve the accuracy, making the solution theoretically relevant, but rendering the entire exercise clinically useless, as generating synthetic data for cancer is highly deplorable.

The resulting framework, Peshnaja presents a decent accuracy of 70% with an overall RMSE of 2.65, allowing for exact prediction of survival months making it clinically relevant. A comparison of Peshnaja with other frameworks is shown in Table XVII.

TABLE XVI: COMPARISON OF PESHNAJA WITH OTHER GLIOBLASTOMA STUDIES. ALL THESE STUDIES HAVE DEVELOPED THEIR MODELS USING SEER DATA. COMPARISON SHOWS THAT PESHNAJA IS (I) FINE-TUNED, AND (II) DOES NOT EMPLOY SYNTHETIC DATA, RENDERING IT CLINICALLY RELEVANT.

Components	Peshnaja	G. Nath [12]	G.Nath [11]	B. Bakirarar [10]	Samara [9]	Yang [7]	Senders [6]
Fine – tuned results.	✓						
Did not employ synthetic data.	✓					✓	✓
Did not imputed missing values.	✓	✓	✓	✓		✓	

We employed a systematic approach for feature extraction. For Layers 1 and 2, we used RFC to extract important features from the dataset. Whereas, for Layer 3, we used Sequential forward and Backward searches. By conducting feature selection for each classifier and regressor separately, we greatly improved the accuracy of our model.

Table XVII: TABLE SHOWING FEATURES EMPLOYED FOR THE FINAL MULTI-TIER MODEL (‘PESHNAJA’)

	Features	**Classification**				**Regression**				
	Symbols	**O**	**A**	**B**	**4**	**R1**	**R2**	**R3**	**R4A**	**R4B**
1.	O		✓	✓	✓		✓			
2.	R_R									
3.	R_E									
4.	α									
5.	XY	✓	✓	✓	✓	✓	✓	✓		✓
6.	P	✓	✓	✓	✓	✓	✓			
7.	G			✓	✓			✓		
8.	T_T								✓	
9.	L	✓	✓	✓	✓	✓		✓		
10.	S_{04}		✓	✓	✓	✓	✓		✓	
11.	$R_{X(P)}$	✓	✓	✓	✓					
12.	$R_{X(07)}$	✓	✓				✓			✓
13.	$R_{X(06)}$	✓	✓			✓			✓	
14.	C_{KM}						✓	✓	✓	
15.	γ		✓		✓	✓				
16.	C_{04}	✓	✓			✓		✓		✓
17.	E									
18.	$ICCC$					✓				
19.	S_β									✓
20.	C_R									
21.	B	✓	✓	✓	✓			✓		
22.	H_{ICD}	✓		✓	✓					✓
23.	I									
24.	$R_{\bar{S}}$									
25.	P	✓	✓	✓	✓					
26.	S	Target Variable								
27.	M	✓	✓	✓	✓	✓	✓			✓
28.	A	✓	✓	✓	✓	✓	✓		✓	✓
29.	T_M		✓	✓						
30.	T_B								✓	✓
31.	F	✓	✓	✓	✓	✓	✓	✓	✓	✓
32.	Y_γ	✓	✓	✓	✓					
33.	Y_D	✓	✓	✓	✓	✓	✓	✓	✓	✓
34.	SS									
Total Features (N)		15	18	16	16	12	10	8	8	10

Lastly, Peshnaja is trained on SEER dataset ranging from year 2000 – 2021. We did not employ the entire dataset available in SEER (1975 – 2021) as some important features like ‘Radiation’ and ‘Chemotherapy’ are only available from 2000 onwards.

VII. CONCLUSION

This humble effort presents Peshnaja – a clinically relevant, finely–tuned framework for predicting the survivability of Glioblastoma patients. Peshnaja presents a decent accuracy of 70% with an overall RMSE of 2.65, allowing for exact prediction of survival months making it clinically relevant. A comparison of Peshnaja with other frameworks shows that we did not impute missing values nor employed synthetic data to force good results, thereby keeping Peshnaja true to the existing data, with state – of – the – art finesse.

As cancer treatment is expensive, Peshnaja could help oncologists determine the likelihood of a patient's survival, which in turn will help the patient and their doctor determine the optimal course of the treatment, based on available finances, and expected chances of survival.

As of now, Peshnaja is limited to survival time prediction for Glioblastoma. In future, the authors wish to extend the framework for other types of cancer.

Lastly, the authors chose the word Peshnaja for the proposed framework as it is a combination of two words 'predict' and 'survival.' The word for predict in Urdu is پیشن گوئی (transliterated as Peshangui), while the same in Persian is پیش بینی (transliterated as Peshbini). Similarly, the word for survival in Arabic is نجاة (transliterated as Naja). Hence, we decided to combine the two words 'predict' and 'survival' to 'Peshnaja.'

Acknowledgment: The authors thank (i) the SEER program for sharing their dataset for the development of Peshnaja; (ii) the website (www.indifferentlanguages.com) for helping us choose the name Peshnaja; and lastly (iii) Sabz-Qalam (Grant no. SQ-2022-Cancer) for supporting this work.

Conflict of Interest: The authors declare that they have no conflict of interest.

Author Contribution: The following table summarizes each authors contribution:

#	Items	AA	BW	FA	FGA	AA	MLS	IW	ARA
1	Propose the problem		×	×					
2	Literature review and gap analysis	×		×				×	
3	Obtain the data	×							
4	Design the experiments	×	×		×	×			×
5	Analyze the data	×							
6	Build the model	×	×						
7	Keeping the work clinically relevant		×	×					
8	Write the paper	×					×	×	
9	Edit the paper	×	×		×	×	×	×	×
10	Supervise the project		×						

Please note, that both AA and BW should be considered joint first authors.

REFERENCES

[1] J. S. Barnholtz-Sloan, A. E. Sloan, F. G. Davis, F. D. Vigneau and P. Lai, "Incidence Proportions of Brain Metastases in Patients," *Journal of Clinical Oncology,* vol. 22, pp. 2865-2872, 2004.

[2] V. Shah and P. Kochar, "Brain cancer: implication to disease, therapeutic strategies and tumor targeted drug delivery approaches," *Recent patents on anti-cancer drug discovery,* vol. 13, pp. 70-85, 2018.

[3] "Surveillance, Epidemiology, and End Results (SEER)," National Cancer Institute,DCCPS, Surveillance Research Program,, April 2023. [Online]. Available: https://seer.cancer.gov/statfacts/html/brain.html). [Accessed July 2023].

[4] I. A. Tewarie, J. T. Senders, S. Kremer, S. Devi, W. B. Gormley, O. Arnaout, T. R. Smith and M. L. D. Broekman, "Survival prediction of glioblastoma patients—are we there yet? A systematic review of prognostic modeling for glioblastoma and its clinical potential," *Neurosurgical Review,* vol. 44, pp. 2047-2057, 2021.

[5] D. Cox and D. Oakes, Analysis of survival data Chapman and Hall, 1984.

[6] J. Senders, P. Staples, A. Mehrtash, D. Cote, M. R. D. Taphoorn, W. Gormley, T. Smith, M. Broekman and O. Arnaout, "An online calculator for the prediction of survival in glioblastoma patients using classical statistics and machine learning.," *Neurosurgery,* vol. 86, p. E184, 2020.

[7] Y. Yang, M. Yao and S. Long, "Prognostic Nomograms for Primary High-Grade Glioma Patients in Adult: A Retrospective Study Based on the SEER Database," *BioMed Research International,* vol. 2020, 2020.

[8] K. Shameer, K. W. Johnson, B. S. Glicksberg, J. T. Dudley and P. P. Sengupta, "Machine learning in cardiovascular medicine: are we there yet?," *Heart,* vol. 104, pp. 1156-1164, 2018.

[9] K. Samara, Z. Al Aghbari and A. Abusafia, "GLIMPSE: a glioblastoma prognostication model using ensemble learning—a surveillance, epidemiology, and end results study.," *Health Information Science and Systems,* vol. 9, pp. 1-13, 2021.

[10] B. Bakirarar, E. Egemen and F. YAKAR, "Machine learning model to identify prognostic factors in glioblastoma: a SEER-Based analysis," *Pamukkale Medical Journal,* vol. 16, pp. 338-348, 2022.

[11] G. Nath, A. Coursey, J. Ekong, E. Rastegari, S. Sengupta, A. Dag and D. Delen, "Determining the Temporal Factors of Survival Associated with Brain and Nervous System Cancer Patients: A Hybrid Machine Learning Methodology," *International Journal of Healthcare Management,* pp. 1-15, 2023.

[12] G. Nath, A. Coursey, Y. Li, S. Prabhu, H. Garg, S. Halder and S. Sengupta, "An interactive web-based tool for predicting and exploring brain cancer survivability.," *Healthcare Analytics,* vol. 3, p. 100132, 2023.

[dataset][13] *Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) SEER*Stat Database: Incidence - SEER Research Data, 22 Registries, Nov 2021 Sub (2000-2021) - National Cancer Institute, DCCPS, Surveillance Research Program,* Aprill 2023.

Highlights:

- **Need:** Glioblastoma is a deadly tumor with poor survival rates. As cancer treatment is expensive, Peshnaja could help oncologists determine the likelihood of a patient's survival, which in turn will help determine the optimal course of the treatment.
- **Proposed model:** Peshnaja is an ensemble learning, hybrid model combining 4 classifiers and 5 regressors, showcasing 70% accuracy, and an overall RMSE of 2.65. Together, Peshnaja can predict exact survival months for glioblastoma patients.
- **Improvement:** Most of the previous works limit their survival model to 1 –, 2 –, or 10 – year survival prediction, hence limiting their usability in clinic. Moreover, several of these works simulate data to balance their proposed classes, rendering their work unreliable. Here, we did not impute missing values nor employed synthetic data to force good results. Moreover, Peshnaja 70% accuracy and an overall RMSE of 2.65, allows for exact prediction of survival months, making it significantly better than previous works and clinically relevant.