

Article

Serum CD133-Associated Proteins Identified by Machine Learning Are Connected to Neural Development, Cancer Pathways, and 12-Month Survival in Glioblastoma

Thomas Joyce ¹, Erdal Tasci ¹ , Sarisha Jagasia ¹, Jason Shephard ¹, Shreya Chappidi ^{1,2}, Ying Zhuge ¹, Longze Zhang ¹, Theresa Cooley Zgela ¹, Mary Sproull ¹, Megan Mackey ¹, Kevin Camphausen ¹ and Andra V. Krauze ^{1,*} 

¹ Radiation Oncology Branch, Center for Cancer Research, National Cancer Institute NIH, 9000 Rockville Pike, Bethesda, MD 20892, USA; thomas.joyce@nih.gov (T.J.); sarisha.jagasia@nih.gov (S.J.); jasonshephard@gmail.com (J.S.); shreya.chappidi@nih.gov (S.C.); zhugey@mail.nih.gov (Y.Z.); longze.zhang@nih.gov (L.Z.); theresa.cooleyzgela@nih.gov (T.C.Z.); sproullm@mail.nih.gov (M.S.); mmackey@mail.nih.gov (M.M.); camphauk@nih.gov (K.C.)

² Department of Computer Science and Technology, University of Cambridge, 15 JJ Thomson Ave, Cambridge CB3 0FD, UK

* Correspondence: andra.krauze@nih.gov

Simple Summary: Glioblastoma (GBM) is an aggressive form of brain cancer characterized by its poor prognosis due to its resistance and recurrence capabilities. Liquid biopsies have been increasingly utilized when studying GBM as they provide extensive amounts of data at multiple time points without the invasiveness of tissue samples. We used a large-scale proteomic panel from the serum of GBM patients collected before chemoradiation therapy (CRT) to study the connections of CD133, a protein recognized as involved in GBM resistance. We used a novel machine learning process to identify 24 proteins associated with CD133 and 12-month survival, successfully grouping patients into risk groups based on protein profiles. Our results identify a potentially harmful protein profile while revealing important serum associations of CD133. These identified proteins are possible prognostic indicators and treatment entry points with heightened value because they can be frequently traced and monitored.

Abstract: Glioma is the most prevalent type of primary central nervous system cancer, while glioblastoma (GBM) is its most aggressive variant, with a median survival of only 15 months when treated with maximal surgical resection followed by chemoradiation therapy (CRT). CD133 is a potentially significant GBM biomarker. However, current clinical biomarker studies rely on invasive tissue samples. These make prolonged data acquisition impossible, resulting in increased interest in the use of liquid biopsies. Our study, analyzed 7289 serum proteins from 109 patients with pathology-proven GBM obtained prior to CRT using the aptamer-based SOMAScan[®] proteomic assay technology. We developed a novel methodology that identified 24 proteins linked to both serum CD133 and 12-month overall survival (OS) through a multi-step machine learning (ML) analysis. These identified proteins were subsequently subjected to survival and clustering evaluations, categorizing patients into five risk groups that accurately predicted 12-month OS based on their protein profiles. Most of these proteins are involved in brain function, neural development, and/or cancer biology signaling, highlighting their significance and potential predictive value. Identifying these proteins provides a valuable foundation for future serum investigations as validation of clinically applicable GBM biomarkers can unlock immense potential for diagnostics and treatment monitoring.

Keywords: CD133; PROM1; serum; proteomics; biomarkers; glioblastoma



Citation: Joyce, T.; Tasci, E.; Jagasia, S.; Shephard, J.; Chappidi, S.; Zhuge, Y.; Zhang, L.; Cooley Zgela, T.; Sproull, M.; Mackey, M.; et al. Serum CD133-Associated Proteins Identified by Machine Learning Are Connected to Neural Development, Cancer Pathways, and 12-Month Survival in Glioblastoma. *Cancers* **2024**, *16*, 2740. <https://doi.org/10.3390/cancers16152740>

Academic Editors: David Wong and Karen E. Pollok

Received: 17 June 2024

Revised: 24 July 2024

Accepted: 26 July 2024

Published: 1 August 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Glioma is the most common type of primary central nervous system cancer, characterized by the uncontrolled growth of malignant cells within the brain or spinal cord. The most aggressive form of glioma is glioblastoma (GBM), which includes astrocytic tumors given a grade IV designation by the World Health Organization (WHO) [1]. GBM is defined by its poor prognosis, as its median survival is only 15 months, with a five-year survival rate of less than 5% [2]. There is not yet a cure for GBM. However, the standard of care (SOC) management involves maximal surgical resection followed by concurrent chemoradiation therapy (CRT) and adjuvant temozolomide (TMZ) [3].

The poor prognosis of GBM is due to the tumor's resistance to treatment and the ability of certain cells within the tumor to evade therapies, particularly those with stem-like properties known as cancer stem cells (CSCs) [4]. CSCs are undifferentiated cells capable of sustaining tumor growth by generating both new CSCs and differentiated daughter cells [5]. Consequently, there is growing research interest in biomarkers that indicate the presence of CSCs due to their diagnostic and prognostic value.

CD133, or Prominin-1 (PROM-1), is a membrane-bound glycoprotein recognized as a CSC biomarker [5]. Recent research has shifted towards examining CD133's role as a potential contributor to tumor resistance and recurrence in GBM [6]. Many of these studies have been conducted utilizing in vivo and in vitro methods [6]. However, these methods are limited as they are surrogate models rather than direct human research. Other studies have used tumor tissue at the time of initial resection or biopsy. However, the invasive nature of these methods makes them difficult to repeat over the course of a disease, preventing monitoring over time.

This has spurred interest in using liquid biopsies, which involve blood, urine, or other bodily fluids as biospecimens. Liquid biopsies are cost-effective, minimally invasive, and can be collected frequently, reducing patient impact, and facilitating large-scale data acquisition. The extensive datasets generated, particularly from proteomic and metabolomic panels, have driven the adoption of machine learning (ML) to extract valuable insights [7–9]. An overarching goal of analyzing these large-scale panels is to identify and validate biomarkers that can indicate cancer presence or monitor treatment progress [10,11].

Survival is an important outcome endpoint in GBM as it is accurately calculated from the date of diagnosis to the date of death in most studies and registries, as compared to progression-free survival, which is subject to limitations in interpretation [12]. Assessment of survival at 12 months following diagnosis indicates the efficacy of treatment following SOC, given 6 to 12 months of adjuvant TMZ post-CRT. This time point indicates the disease trajectory thus rendering it a common clinical trial endpoint that can be compared across studies and with real-world data [2,13,14].

Although serum data have been explored in GBM [10,15–17], there remains a significant shortage of comprehensive large-scale data available. In this study, we utilized a large-scale proteomic panel based on serum collected from patients with pathology-proven GBM to investigate CD133. Our objective was to identify proteins related to CD133 using a clinically relevant survival endpoint. As a result, using ML, we identified proteins that were significantly associated with both CD133 and 12-month overall survival (OS). This approach combined correlation and feature selection (FS) analyses with survival and clustering evaluations for greater understanding. With further validation, our findings can enhance the understanding of the interplay between CD133 and CSCs, leading to improved prognosis, treatment monitoring, and optimization for better outcomes. The technological and clinical innovations of this study are summarized below.

1.1. Technological Innovations

- This study is the first to develop an approach to large-scale serum proteome data based on a single literature-validated protein (CD133) and a clinically meaningful endpoint (12-month OS).

- The combined use of Pearson's correlation analyses and Recursive Feature Selection, Cross-Validated (RFECV) for selecting serum proteins related to serum CD133 and 12-month OS in this study is novel.
- To ensure inclusive protein selection, we tested multiple FS methods on two base models with differing approaches to 12-month OS prediction.
- To validate our findings, we utilized two inverse but complementary processes involving correlation and FS to address the lack of large-scale data available for external validation.
- We used predicted hazard scores from the Cox Proportional Hazards Model for Gaussian Mixture Model clustering to categorize patients into risk groups based on their protein profiles.

1.2. Clinical Innovations

- This is the first study to investigate CD133 and its associations using serum proteome data.
- The findings enhance the understanding of conclusions derived from serum data while addressing whether these conclusions are transferable with other types of biospecimens.
- Almost all identified proteins are either expressed in the brain, involved in brain function, and/or related to cancer biology pathways, underscoring their likely involvement in, or alteration from, GBM.
- The results identify potentially harmful protein profiles that might predispose individuals to short survival with SOC management, warranting further validation.
- The list of identified proteins offers candidates that can potentially be used to monitor prognosis and/or treatment as new data emerge and reference ranges are validated.

2. Materials and Methods

2.1. Dataset

Serum biospecimens collected from 109 patients with pathology-proven GBM (2005–2023) treated with CRT on National Institutes of Health (NIH) Institutional Review Board (IRB) approved studies before and after completion of CRT were included in the analysis. Serum samples were obtained prior to initiation of upfront CRT (mean 6.7 days, range (0 to 24)), then frozen at -80°C for an average of 3951 days (range 239–7072 days), and finally thawed and screened using the aptamer-based SOMAScan[®] proteomic assay technology. This multiplexed, aptamer-based approach (SomaScan[®] assay) in the SomaLogic[®] research facility used samples provided by the authors to measure the relative concentrations of 7596 protein targets (7289 human) for changes in expression using approximately 150 μL of serum [18,19]. The NIH Integrated Data Analysis Platform (NIDAP) [20] was employed for storage, query operations, and analysis.

2.2. The Proposed Scheme

For the preliminary portion of this analysis, a combined evaluation utilizing Pearson's correlation and machine learning (ML) in the form of FS was developed to identify proteins related to both CD133 and 12-month OS. This was performed using two parallel analyses: Process 1 and Process 2. Process 1 conducted the correlation analysis first and followed this with an FS analysis of the remaining proteins to select those most valuable to the prediction of 12-month OS. Process 2 first conducted the FS survival evaluation and then performed the correlation analysis on the resulting proteins. The final protein list included only those proteins that emerged from both processes, ensuring robust associations of the identified molecules with serum CD133 and 12-month OS.

This final list of identified proteins was then utilized in the second portion of this analysis which began with using the Cox Proportional Hazards Model to further understand the individual protein contributions to 12-month OS. This analysis also produced predicted hazard scores for each patient that were used in Gaussian Mixture Model clustering to group

patients into risk groups based on their protein profiles. The full process is summarized in Figure 1, and all code used in these analyses can be found in the Supplemental Materials.

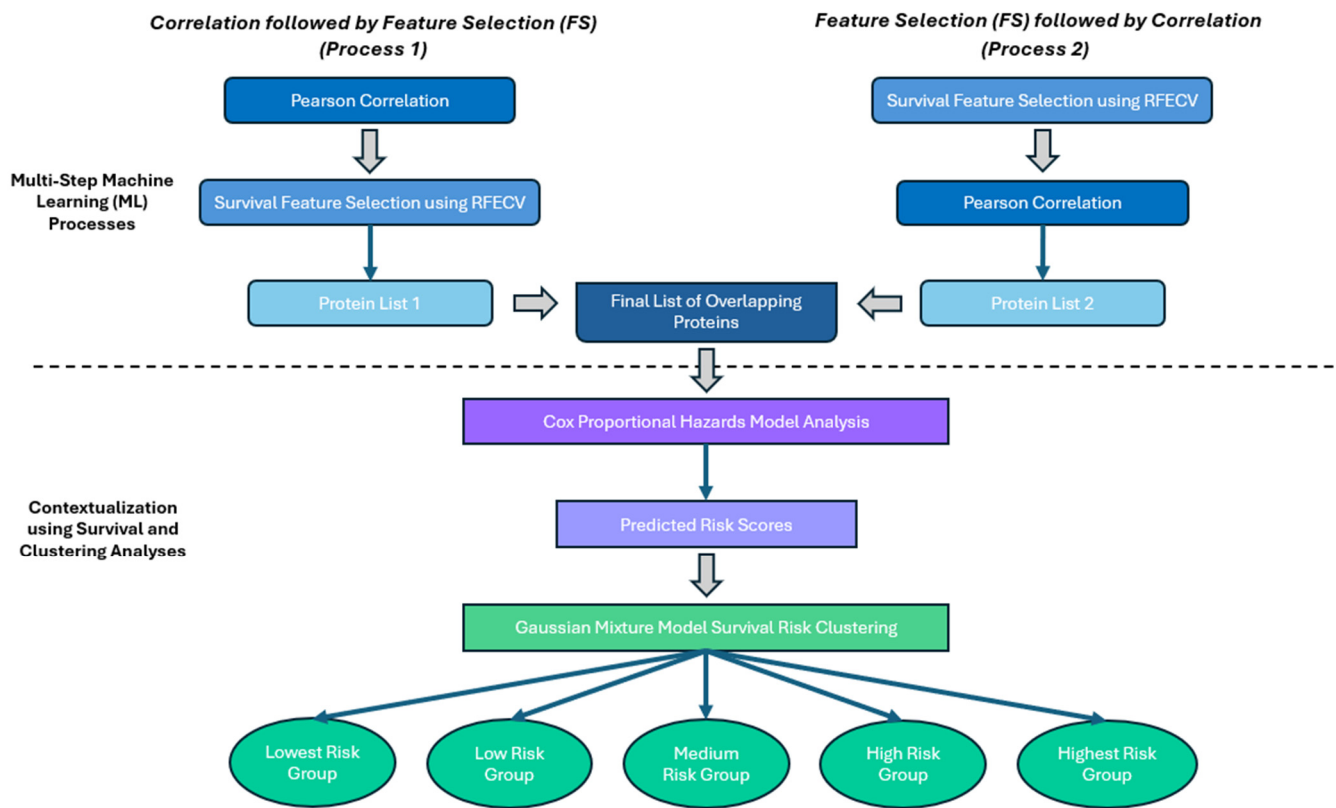


Figure 1. General overview of the proposed methodology for protein selection and analysis.

2.3. Pearson's Correlation Analysis

Pearson's correlation analyses were used to identify pairwise linear relationships between the expression levels of panel proteins and the expression levels of CD133. This method was chosen due to its universality in research, its direct explainability, and the suitability of the shape of the data following log transformation [21]. Upon executing the analysis in both processes, we eliminated any proteins with a coefficient with an absolute value of less than 0.35 from the datasets, as this is an accepted lower boundary for moderate associations [22]. False discovery rate (FDR) was also integrated into this analysis portion as proteins with an FDR greater than 0.1 were simultaneously removed. We chose an FDR of 0.1 to allow for a wider inclusion of potential signals in the FS step of Process 1, as it increased the number of proteins from 161 to 216. This threshold was then carried over into Process 2.

2.4. Base Models

Two differing base models were employed for the FS portion of these processes to enhance protein selection. The first was Logistic Regression (LR), a supervised ML model typically used in binary classification. This model assigns coefficients to each variable and uses them to make classifications with a certain confidence level [23]. The other base model was Random Forest (RF), which utilizes decision trees rather than coefficients to make its classifications. Through recursive partitioning anchored in randomization, RF creates thresholds for different variables that lead down various paths, with outcomes being the pooled consensus of these pathways [24].

Each model was trained on the training datasets of both the correlated proteins of Process 1 and the full protein dataset used in Process 2. The log-transformed version of these data were utilized for all analyses, and scaling was performed using *MinMaxScaler()*

from *sklearn.preprocessing* [25]. Additionally, a train–test split of 80% to 20% was utilized throughout. Optimal parameters for the base models were found using the *GridSearchCV*() package from *sklearn.model_selection* [26]. The penalty was kept constant at *l2* for LR, while the solvers were chosen from [*lbfgs*, *liblinear*, *sag*, *newton-cg*, *saga*], and the C values were picked from [0.01, 0.1, 0, 1, 10, 100]. For RF, the *max_depth* was decided between (3, 6, 9, *None*), the *max_features* were selected from [*log*, *sqrt*], the *max_leaf_nodes* were chosen from (3, 6, 9, *None*), and the *n_estimators* were determined from (25, 50, 75, 100, 125, 150).

For Process 1, the optimal parameters for LR were [*solver*= *liblinear*, *C* = 0.1], while for RF, they were [*max_depth* = 3, *max_features* = *log2*, *max_leaf_nodes* = 6, *n_estimators* = 125]. For Process 2, the optimal parameters for LR were [*solver* = *lbfgs*, *C* = 1], and for RF were [*max_depth* = 6, *max_features* = *log2*, *max_leaf_nodes* = *None*, *n_estimators* = 100].

Due to the imbalance of survivors and non-survivors in the training dataset, measures were taken to account and adjust for this. Following this fine-tuning, tests were carried out to determine whether the training data's stratification helped improve the base models' performance in either process. Improvements were only seen in the LR models of Process 2, so we only incorporated stratification in these tests. Additionally, LR models can incorporate "class_weights" as a parameter, which can be used to counteract specific imbalances through weight assignment, so specific weights were calculated and incorporated. For the LR models in Process 1, the weights were [0:0.613, 1:2.719], with 0 representing survivors and 1 representing non-survivors. Due to the stratification of the training data in Process 2, the weights changed slightly for this iteration to [0:0.630, 1:2.417].

2.5. Feature Selection Methods

Two FS methods with differing approaches were tested; however, only one was chosen for the full analysis. The first was the Least Absolute Shrinkage and Selection Operator (LASSO), which is known for its success when working with highly dimensional datasets. Using regularization, LASSO assigns and shrinks coefficients if unrelated to the variable of interest. By restraining the sum of these coefficients, many are shrunk to 0, which is how FS occurs [27].

The other FS method was RFECV, a wrapper-style FS model that initially classifies using all features and eliminates features in steps. After observing the initial performance of the model using all features, it ranks these features in terms of their importance to the model's ability to classify. Then, at each step, it eliminates a subset and reevaluates the model's performance while updating the rankings. This approach also accounts for the base model type, allowing for specialization in the optimal production of various models [28].

The effectiveness of the FS methods was evaluated using accuracy, Receiving Operator Characteristic (ROC)-Area Under the Curve (AUC), and Precision-Recall AUC. ROC-AUC and Precision-Recall AUC were utilized alongside accuracy to contextualize the model's performance as they provide greater insight into the actual functioning and decision-making of the model. ROC-AUC focuses more on the classification of the survivors in this analysis, while the Precision-Recall AUC focuses on the smaller, non-survivor class as it is known for its value when working with imbalanced binary classification [29,30]. As a result, the three metrics together provided a holistic view of the model's performance.

Utilizing the base models and these performance metrics, LASSO and RFECV were evaluated using both datasets, and their accuracy scores are displayed in Table 1. The *RepeatedStratifiedKFold*() package from *sklearn.model_selection* was used for cross-validation across all tests to prevent overfitting [31]. Its parameters involved a split of 10 with 3 repeats at each step. Despite similar performance on the first dataset, the LASSO FS resulted in a reduction in accuracy from 0.82 to 0.73 when tested on the second dataset using LR, which was accompanied by decreases in ROC-AUC and Precision-Recall AUC. As a result, RFECV was chosen for this analysis.

Table 1. Accuracy rates from the RFECV and LASSO FS methods employed to select the optimal approach for protein selection. Decreased accuracy when using LASSO in Process 2 resulted in the selection of RFECV for analysis.

		Correlation Followed by Feature Selection (FS) (Process 1)		Feature selection (FS) Followed by Correlation (Process 2)	
		Before FS	After FS	Before FS	After FS
Logistic Regression	LASSO	0.77	0.77	0.82	0.73
	RFECV	0.77	0.82	0.82	0.82
Random Forest	LASSO	0.59	0.73	0.55	0.68
	RFECV	0.59	0.73	0.55	0.68

Following this selection, RFECV was used in Processes 1 and 2 in their totality. RFECV allows for the modification of its “step” function, which involves the number of proteins eliminated at each pass, so to further improve its performance, various thresholds were tested. For Process 1, step values ranging from 1–5 were tried and selected due to the smaller dataset, and it was a step of 2 for LR and 3 for RF that performed best. For Process 2, due to the much larger set of values, step values between 10 and 50 were tested at increments of 10. For LR, a step of 20 was chosen to obtain the best results, while for RF, it was a step of 10. Finally, RFECV accounts for the base estimator, so LR and RF were tested separately, and the protein sets that they produced were combined. This was performed in both processes, and a final list of proteins was compiled aggregating overlapping proteins.

2.6. The Cox Proportional Hazards Model

After completing the two processes, a Cox Proportional Hazards Model analysis was conducted to contextualize the final protein list by elucidating the individual contributions of each protein to 12-month OS prediction. The *CoxPHFitter()* package from *lifelines* was used to produce parameter estimates, hazard ratios, and *p*-values [32]. Parameter estimates were evaluated based on their direction, magnitude, and accompanying hazard ratios. Selection of the most significant proteins was performed using parameter estimates and *p*-values and visualized as a volcano plot using IPA [33].

Given that these proteins were identified as major contributors to the prediction of 12-month OS, clustering of the predicted hazard scores for individual patients produced by the *CoxPHFitter()* package was utilized to group protein profiles associated with risk of death. Equation (1):

$$\beta^T X = \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p, \quad (1)$$

shows how these scores were calculated, where β represents the parameter estimate that was assigned to the protein, and X represents the actual expression value of that protein for an individual. For each patient, the parameter estimate for the protein was multiplied by the respective protein expression value, and this was performed for all proteins and then summed into one vector [34]. As shown by Equation (2):

$$h(t|X) = h_0(t)\exp(\beta^T X), \quad (2)$$

the exponential function of this vector was then taken, meaning that the further this result deviated from 1, the greater the risk of that patient’s protein expression profile. At the same time, if this value was less than 1, this meant that the patient’s profile was protective. Finally, this value was multiplied by the baseline hazard ($h_0(t)$), which is the same for all patients, to account for the specified time point (12 months) [34].

2.7. Gaussian Mixture Model Clustering

After producing predicted hazard scores for each patient, Gaussian Mixture Model clustering was performed to group patients based on their protein profiles. Gaussian Mixture Model clustering was chosen because it is more flexible than traditional k-means clustering. Rather than simply calculating the distance from a centroid, it searches for vari-

ous Gaussian distributions, which involves considering the variance of the data. Selection of the optimal number of clusters was performed utilizing Bayesian information criteria (BIC) scores, with the minimum score selected [35].

Using this clustering method, individuals with similar predicted hazard scores were grouped together. Given that the parameter estimates used in the calculation of the predicted hazard score for each patient were the same, this meant that patients in the same group had similar protein expression level values across proteins, and therefore, similar protein profiles. As a result, these clusters represented varying risk groups as predicted hazard scores are directly related to 12-month OS. This allowed for a survival analysis using Kaplan–Meier curves. It also resulted in visualizations of the protein profiles of each risk group to provide insight into which expression levels and patterns are possibly harmful.

3. Results

3.1. Correlation and Feature Selection Analysis

The breakdown of the protein elimination at each step of the processes is displayed in Figure 2. In Process 1, the 7289 original proteins were narrowed down to 216 proteins following the correlation analysis, of which 36 were chosen as most essential to 12-month OS prediction by RFECV. In Process 2, RFECV utilized 5120 proteins to predict 12-month OS, and 151 of these met Pearson’s correlation criteria. There was an overlap of 24 proteins, which comprised the final protein list, shown in Table 2. Access to the protein lists produced by each process (36 and 151) along with the final protein list (24) can be found in the Supplementary Materials.

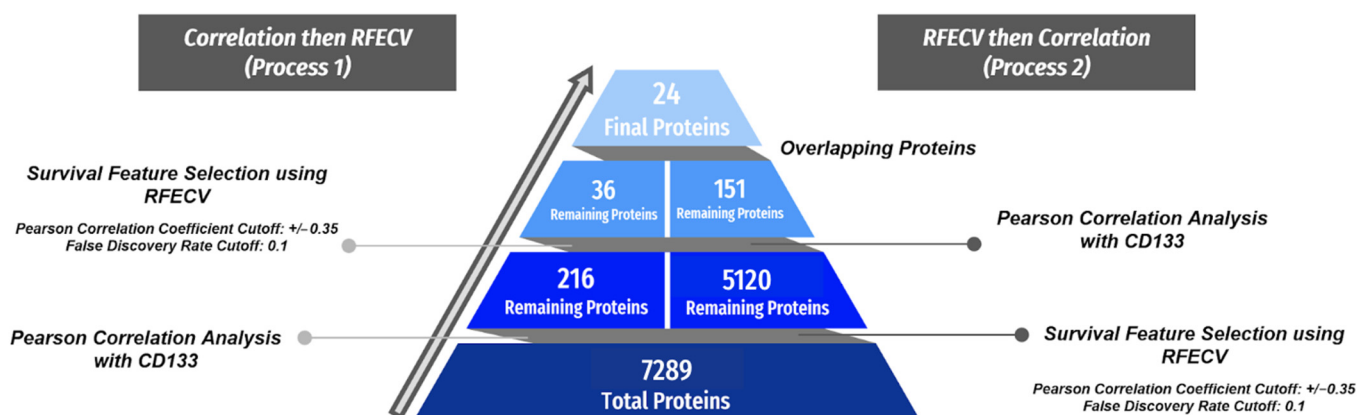
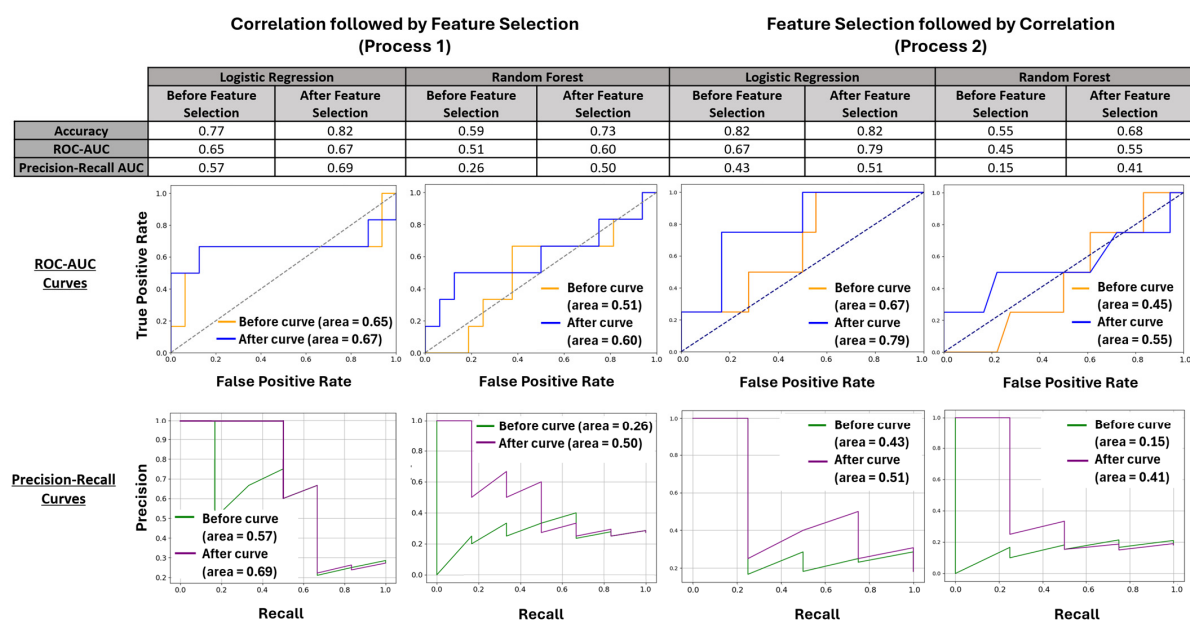


Figure 2. Summary of the initial machine learning data process with protein values. Two parallel processes were employed with overlapping proteins included in the final protein set. The left side shows the progression of Process 1, in which Pearson’s correlation analysis was completed first, followed by FS using RFECV at 12-month OS. The right shows Process 2, which followed an inverse but identical trajectory starting with FS and following this with the correlation analysis. For Pearson’s correlation analysis, proteins with a correlation coefficient with an absolute value greater than 0.35 and a false discovery rate less than 0.1 were maintained.

All models showed improvement after FS with increases in accuracy, ROC-AUC, and Precision-Recall rates in both base models in Process 1 and in the RF model of Process 2 (Figure 3). While the accuracy of the LR model in Process 2 did not increase, the model still showed improvement through higher ROC-AUC and Precision-Recall AUC values. Although the LR models outperformed the RF models in absolute metrics, the RF models demonstrated the most significant improvements in accuracy (59% to 73%) and Precision-Recall AUC (0.15 to 0.41). An LR model achieved the strongest improvement in ROC-AUC (from 0.67 to 0.79). Given the imbalanced training data, with 18 non-survivors compared to 69 survivors, the enhancements displayed in Precision-Recall AUC across these models indicated improved performance in detecting non-survivors, particularly with the RF models.

Table 2. List of proteins identified connected to serum CD133 and 12-month OS.

Entrez Gene Symbol	Target Full Name
RPA2	Replication protein A 32 kDa subunit
AMPD2	AMP deaminase 2
DLK2	Protein delta homolog 2
NEGR1	Neuronal growth regulator 1
PDCL2	Phosducin-like protein 2
POLI	DNA polymerase iota
CEACAM3	Carcinoembryonic antigen-related cell adhesion molecule 3
ITGA6	Integrin alpha-6
PCDHGA10	Protocadherin gamma-A10
SELENOW	Selenoprotein W
TIMM8A	Mitochondrial import inner membrane translocase subunit Tim8 A
CLN5	Ceroid-lipofuscinosis neuronal protein 5: Luminal domain
IL15RA	Interleukin-15 receptor subunit alpha
P3H1	Prolyl 3-hydroxylase 1
UTS2R	Urotensin-2 receptor
RGS4	Regulator of G-protein signaling 4
PTPRS	Receptor-type tyrosine-protein phosphatase S
FURIN	Furin
MSR1	Macrophage scavenger receptor types I and II: Extracellular domain
CSNK2B	Casein kinase II subunit beta
SAT2	Diamine acetyltransferase 2
TRAPPC5	Trafficking protein particle complex subunit 5
CREB3L1	Cyclic AMP-responsive element-binding protein 3-like protein 1
SCN3B	Sodium channel subunit beta-3

**Figure 3.** LR and RF were utilized in both iterations of this process and were evaluated using accuracy, ROC-AUC, and Precision-Recall AUC. The initial tables summarize the improvements made using

these processes accompanied by ROC-AUC curves and Precision-Recall AUC curves that both visualize and contextualize these values.

3.2. CD133 Serum

Pre-CRT serum CD133 expression levels measured in Relative Fluorescence Units (RFUs) for all 109 patients are displayed in Supplemental Figure S1. Levels ranged from 13100 RFUs to 3326 RFUs (mean of 4982 RFUs) with CD133 having the 6th highest levels compared with the 24 identified proteins in the analysis. The highest protein expression was observed for PTPRS (mean of 61085 RFUs), which for clinical context was significantly closer to the expression level of serum Albumin (Supplemental Figure S2). The time the samples were stored in the freezer from collection to analysis did not reveal any systematic impact on protein measurements, as shown in (Supplemental Figure S3).

3.3. The Cox Proportional Hazards Model

The results of the Cox Proportional Hazards Model indicate that proteins with parameter estimates of the highest magnitude are most impactful as seen in RPA2 (−6.67), ITGA6 (−5.27), and PDCL2 (−4.14). To identify the most significant contributors to 12-month OS, a volcano plot using the parameter estimates and *p*-values was created, shown in Figure 4a. The most significant proteins (lowest *p*-values) were RPA2 (parameter estimate = −6.67, HR = [0–0.76], *p* = 0.04) and AMPD2 (parameter estimate = −2.42, HR = [0.01–1.00], *p* = −0.05). TRPA2, AMPD2, DLK2, NEGR1, POLI, ITGA6, CLN5, P3H1, PDCL2, and CEACAM3 were the 10 proteins with the most significant contributions (Figure 4A) in the complete Cox Proportional Hazards Model analysis (Supplementary Table S1).

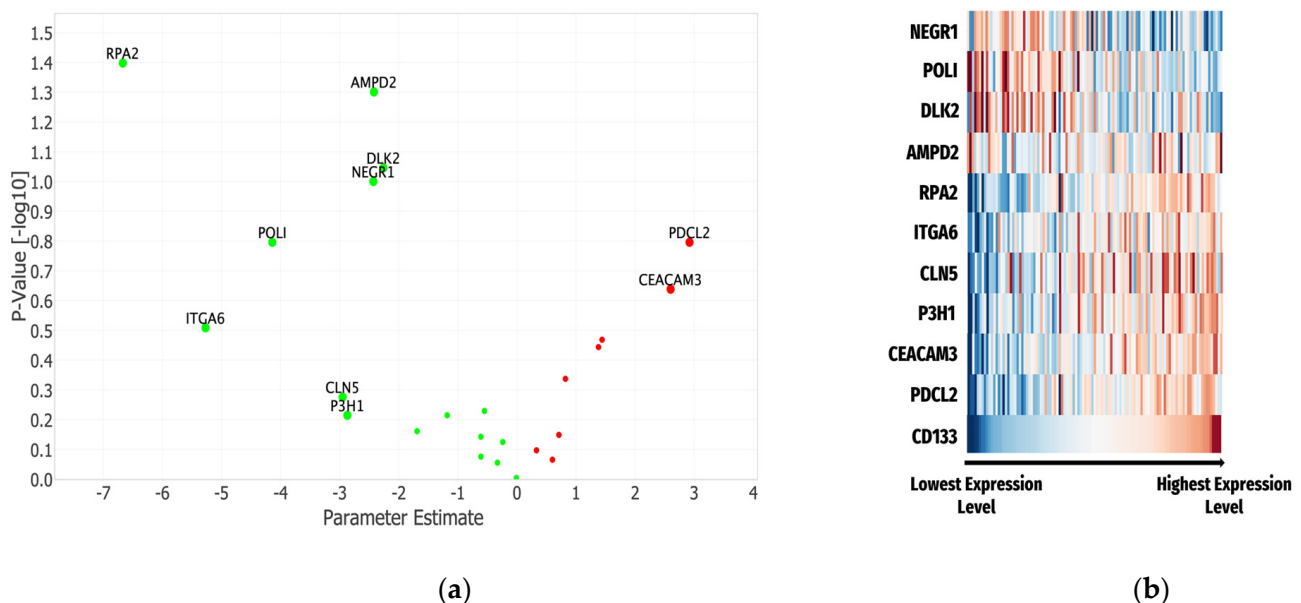


Figure 4. (a) Volcano plot of the Cox Proportional Hazards Model analysis for the proteins identified as connected to serum CD133 and 12-month OS. Top 10 proteins based on *p*-values and parameter estimates were labeled. (b) Heat map of log transformed serum protein expression levels displayed with CD133 expression for comparison. The median values for each individual expression protein were employed with each band representing the value for each individual patient. (Red to orange—indicates protein expression values greater than the median, Blue to light blue—indicates expression levels less than the median, both becoming darker with movement away from the median).

Heat maps for each of the identified molecules were created and visualized, contextualizing patterns of serum expression levels in relationship to serum CD133. Patterns involving expression levels of these proteins with CD133 were evident and expected given the correlation component of this analysis. Of the 10 proteins identified, 6 generally aligned

with the low to high expression level pattern of CD133 compared to 3 that had inverse patterns, while AMPD2 had a more ambiguous signature (Figure 4b).

3.4. Gaussian Mixture Model Clustering

The Gaussian Mixture Model clustering that produced the lowest BIC score using the predicted hazard scores resulted in five survival risk groups, labeled Lowest Risk Group ($n = 20$), Low Risk Group ($n = 18$), Medium Risk Group ($n = 49$), High Risk Group ($n = 16$), and Highest Risk Group ($n = 6$). The Kaplan–Meier survival analysis for 12-month OS was statistically significant ($p < 0.0001$) (Figure 5a) and revealed differences in survival (Figure 5b) with the Lowest Risk Group having 100% survival at 12 months, while the Highest Risk Group had 0% 12-month survival.

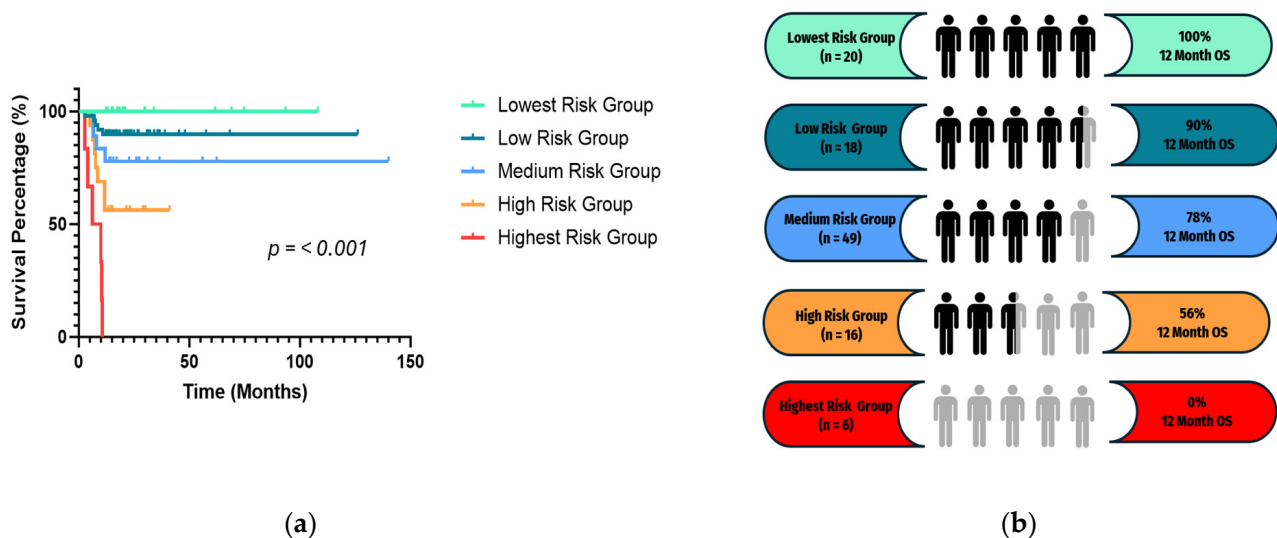


Figure 5. (a) Survival analysis of the clustering groups based on 12-month OS. (b) Summary of the 12-month OS rates in each risk group represented on the survival graph with corresponding colors.

Alongside this survival exploration, the complementary aim of the clustering analysis was to evaluate the protein profiles of individuals in each survival group to identify patterns of protein expression (Figure 6). Several proteins had opposite expression levels in the Highest Risk vs. Lowest Risk groups. CEACAM3, SAT2, TRAPPC5, PDCL2, RPA2, ITGA6, CSNK2B, and Furin all have their highest mean expression levels in the Lowest Risk group with their lowest mean expression levels in the Highest Risk group. SELENOW, TIMM8A, and PCDHGA10 revealed the opposite, with their lowest mean values in the Lowest Risk group and their highest mean values in the Highest Risk group. More variation was identified in the middle three groups, and some seemingly contradictory results between the Lowest Risk Group and the Low Risk Group, as well as between the Highest Risk group and the High Risk group were also seen. However, the table indicates that extremely high or low mean protein expression levels seem to have the most significant impact on 12-month OS.

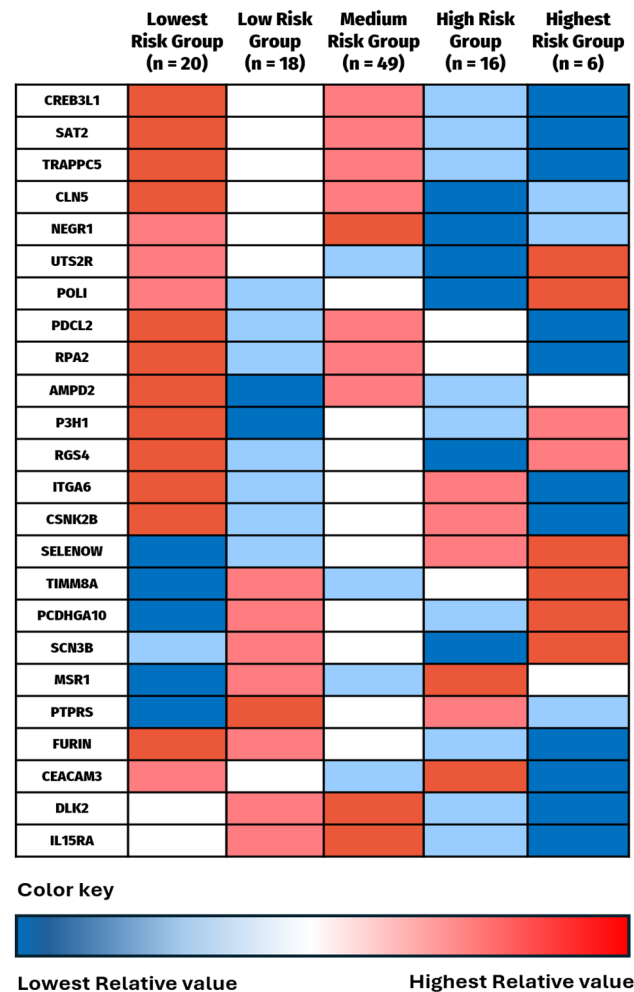


Figure 6. Comparison of mean expression values for each identified protein by survival risk group. The group with highest mean value was given the darkest red color, the second highest value lighter red, the middle value white, the second lowest value light blue, and the lowest value dark blue. This color gradient was chosen to stay consistent with the previous heat maps in Figure 4b. Numerical data are shown in Supplemental File S1.

4. Discussion

CD133 has primarily been used as a biomarker for identifying CSCs, initially gaining recognition for its potential diagnostic and prognostic potential [36–38]. However, in recent years, its role has shifted towards being considered a functional unit in tumor resistance and recurrence, as explored in our previous review [6]. Serum data have gained attention in GBM biomarker research due to its ability to evaluate samples throughout the course of a disease noninvasively. Consequently, using a large-scale proteomic panel derived from patients with pathology-proven GBM, this analysis demonstrated that CD133 can be effectively measured in serum, enabling the investigation of its potential utility and associations with other markers and survival outcomes. Although several prior studies have examined serum proteins in GBM [39–42], this is the first to focus on CD133 specifically. We achieved this by developing a novel multi-step ML process that identified proteins associated with both CD133 and 12-month OS, whose validity was confirmed through survival analyses and clustering. Although previous studies have utilized ML to identify GBM biomarkers and analyze omic panels, the selected proteins often lack direct clinical relevance, connections to known prognostic markers, and/or a clear biological mechanism. By centering this investigation on a known biomarker and a well-defined outcome endpoint (12 months), we identified proteins relevant to resistance and recurrence with a shared connection, enhancing potential clinical applicability.

Elevated CD133 in tumor tissue has largely been associated with poor prognosis and survival [38,43–45], though some studies have reported non-contributory CD133 cells that do not support this trend [46]. In this study, serum CD133 expression varied across the population without a clear correlation to 12-month patient survival. This suggests that signals from both dead and living cells might be captured. This finding is crucial for understanding the transferability of serum data-derived conclusions compared to other biospecimens, emphasizing the need for ongoing research and validation. Additionally, several of the identified final proteins showed higher expression levels than CD133, with PTPRS being the highest, indicating the potential for other proteins to serve as proxies for CD133 or its associations in future studies.

Due to the present lack of transferability across biospecimens, there are limited data available for external validation [47,48] with no serum proteomic data available in GBM for direct comparison. To address this, we aimed for enhanced internal validation utilizing two coinciding processes, resulting in the final protein list. Previous studies have utilized statistical analyses alongside FS [49,50], with multiple using Pearson's correlation coefficients in particular prior to or following FS [51,52]. Our novel approach used both which, combined with the optimization of the base models and FS methods, ensured that the selected proteins had robust associations with CD133 and 12-month OS. The effectiveness of this methodology was confirmed through analysis using the Cox Proportional Hazards Model paired with Gaussian Clustering. Similar to Liu et al. (2023), the Cox analysis was employed following FS to provide further insight into the individual survival contributions of the proteins in relation to the group. Combined FS and Cox analyses, including LASSO-Cox and Elastic-Net with Cox, have become popular ways to narrow and rank expansive datasets [53–57]. The Cox analysis identified two proteins (RPA2 and AMPD2) with statistically significant results, while the other important contributors were determined based on parameter estimates and *p*-values.

A clustering analysis was completed that separated patients into five survival risk categories based on their distinct protein profiles [58,59]. The statistically significant survival separation between the survival groups reinforced the clinical value of the final proteins while justifying the utilization of CD133 as the basis for this study. The relative mean expression value comparisons, suggested that the Highest Risk category may be harboring a harmful serum protein profile, irrespective of treatment, as none of these individuals survived to 12 months. This implied a potential predisposition to short survival, which, through future validation, may reveal that CRT is ineffective in overcoming. As a result, targeted interventions based on manipulating these protein levels or upstream mediators may allow for prolonged survival. Conversely, the Lowest Risk group achieved 100% 12-month OS, indicating a possible protective nature to this expression pattern.

Several of the proteins identified in this analysis were found to be expressed in the brain, associated with a specific role in brain function, and/or pertinent to tumor biology (Figure 7a) [60]. This notably included RGS4 (nearly exclusively expressed in the brain) [61], PCDHGA10 (most abundant in the brain) [62,63], NEGR1 (involved in synaptogenesis and neuronal arborization) [64,65], DLK2 (EGF-like Notch inhibitor associated with neuronal differentiation and stemness) [66,67] and SCN3B (associated with ion channel transport, immune response, and glioma-related epilepsy) [68]. The expression of these molecules has shown nearly exclusive or predominant neural expression [60], which provided a form of external validation of the proteins identified in this analysis, while further reinforcing the value of CD133 as the central protein and revealing the potential utility of these proteins when captured in serum (Figure 7b). Expression levels of DLK2 and IL15RA were also aligned across survival risk groups with IL15RA being relevant given its evolving role in targeted therapies to overcome resistance and improve anti-tumor effect [69]. The observed relationships implied a potential connection between CD133, CSCs, and immune and inflammatory responses that can be captured in GBM serum. CREB3L1 is expressed in the brain but is also a prominent transcription factor in cancer biology [70] that has been studied as a biomarker in glioma, given its association with tumor grade and survival [71].

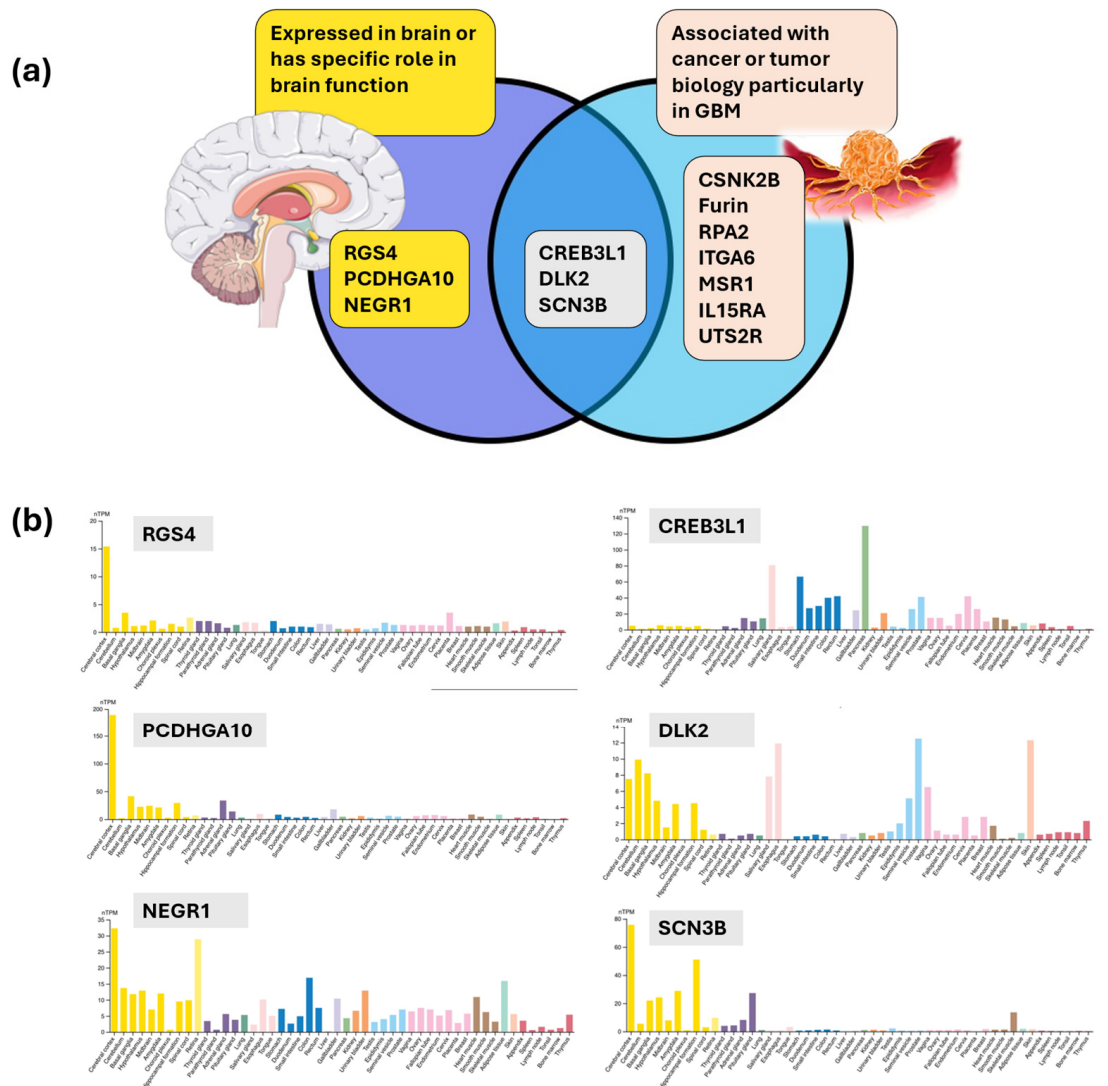


Figure 7. (a) Serum CD133 is associated with molecules expressed in the brain and associated with tumor biology, particularly in GBM [72,73]. (b) Human protein atlas RNA expression overview profiles for RGS4, PCDHGA10, NEGR1, CREB3L1, DLK2, and SCN3B illustrating tissue specificity with a predominance of association in brain tissue (yellow bars) [60].

Although the identified proteins originate from various subcellular locations and tissues of origin [60], they share associations with several biological pathways relevant to GBM, including replication and repair, metabolism and energy currency, neural development, and immune response (Figure 8). RPA2 and POLI are associated with DNA replication and repair and have recently been linked to TMZ and chemoresistance in GBM [74]. RPA2, POLI, and CSNK2B share the nucleus as a subcellular location, which may be relevant for the determination of whether dead cells are being measured. CSNK2B is also connected to cell signaling pathways, alongside ITGA6 and Furin, with documented contributions to neurodevelopmental disorders, epilepsy, and the classical GBM subtypes [75,76]. ITGA6 has also been associated with GBM subtypes and DNA damage response [77], while Furin is a component of secretory pathways connected with tumor immunity in various cancers [78]. In the survival clustering, CSNK2B, ITGA6 and Furin had their highest mean expression level in the Lowest Risk group and their lowest in the Highest Risk group.

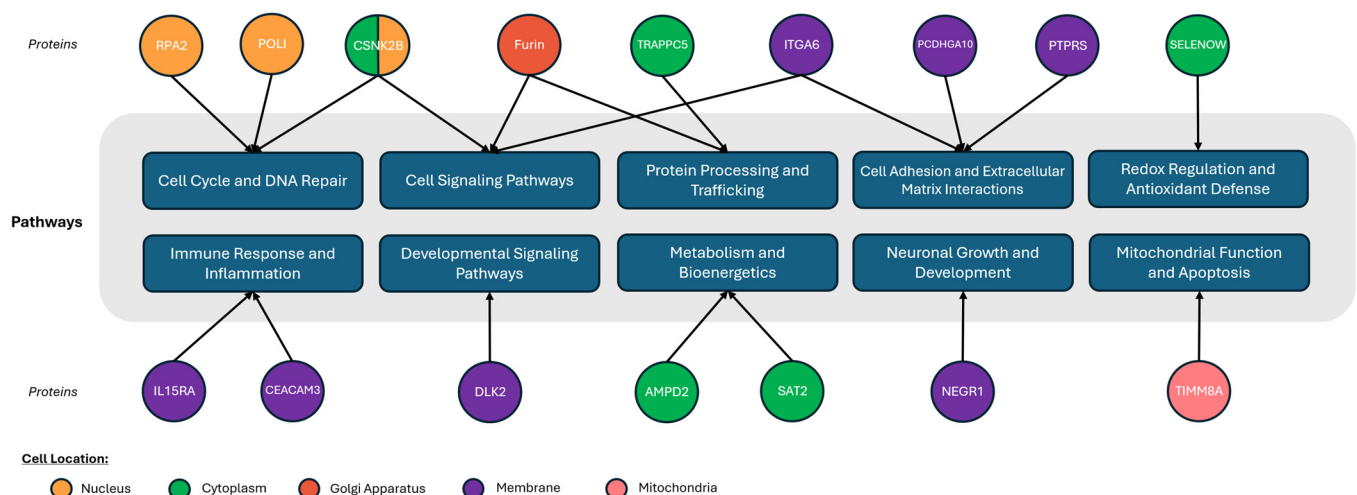


Figure 8. Subcellular location and pathway linkages for the identified proteins [74–87].

Important connections to metabolism and energy currency also involved AMPD2 and SAT2, which are associated with purine and polyamine metabolism [79,80]. Specifically, SAT2 is altered in various malignancies and associated with tumorigenesis and radiation response [81–83]. Further, SELENOW is expressed in the brain and is implicated in EGF signaling and redox regulation [84], while TIMM8A is involved in mitochondrial function with its role in cancer currently evolving [85]. PTPRS is significant in that it can distinguish CSCs from normal brain cells using single-cell RNA sequencing [86] and is one of seven genes whose mRNA expression is associated with tumor grade, with GBM exhibiting the highest levels [87]. In this analysis, however, serum PTPRS expression did not effectively associate with survival risk groups. Additional signals included TRAPPC5 and CEACAM3, the roles of which are currently being determined in cancer and GBM. Comparisons were also made with the most relevant literature proteins compiled in our previous review [6]; however, no specific shared proteins were identified. This further reinforced that serum connections may not translate to other biospecimen relationships. This is of particular relevance since currently there is no serum based GBM dataset available for comparison and comparison with proteomic datasets that are not originating from serum may produce discordant results in some signals depending on their subcellular compartment of origin or disease entity [88].

Limitations of this study include the long timeframe over which data collection occurred and the lack of independent serum data for comparison. We have, however, examined the captured serum signal of several proteins, including CD133, in relationship to the time from sample collection to analysis and have not observed any systematic impact on protein measurement [89,90]. Serum data may also have different ranges as this is an emerging field with various collection methods and measurement technologies. Some of the groups produced in the clustering analysis resulted in varying numbers of patients, some of which were low (e.g., highest risk group $n = 6$) and this aspect limits applicability without large-scale data validation. Thus, conclusions should be drawn with caution. Given the intricacy of interpretation for protein signals in serum as belonging to dead or living cells there is significant intersectionality with clinical outcomes. As large-scale serum proteomic panels grow, several aspects will improve interpretation: (1) co-measurement of large-scale proteomic signals in tissue and in serum in the same patient at the same or similar timepoint; (2) correlation with other data streams such as circulating tumor cells with viability staining; (3) comparison of large-scale proteomic signatures with normal volunteer samples. While these data are currently lacking, apoptosis or necrosis markers may provide proxy signals for the detection of cell death. This last point is subject to future directions using the current dataset as is investigating the alteration of serum proteins with treatment. In addition, the performance and interpretability of the current approach

need to be explored in additional datasets, local and public, and this work is ongoing. The emergence of comparable serum panels would allow for external validation and the ability to identify clinically applicable proteins more strongly. It would also allow for the publication of values and ranges that could be investigated in clinical and laboratory settings. However, this is contingent on the continued exploration of serum data, and although there are over 30 GBM trials currently occurring, the majority are studying tissue [88]. Serum analyses have the potential to provide clinicians with more patient information through non-invasive methods, but this depends on the continued development of this field over the next decade.

5. Conclusions

This research introduced an innovative method for investigating the relationships and survival implications of a specific protein using a serum proteomic panel. By focusing on CD133, this study revealed how this protein behaves in serum and identified other proteins associated with both CD133 and 12-month OS. These findings suggest that CD133 may play a role in tumor resistance and recurrence and that its influence can be captured in serum. Additionally, this study categorized survival risks based on protein profiles, shedding light on potentially harmful protein expression combinations that could lead to shorter survival despite CRT.

As proteomic serum data collection continues to expand and external validation becomes more feasible, tracking GBM prognosis and treatment effectiveness using non-invasive serum samples becomes increasingly likely. Comparing independently collected panels would allow us to refine ranges for these proteins, ultimately determining their clinical applicability. This study's identified proteins serve as a crucial starting point for future research that, once further built upon, have expansive clinical potential.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/cancers16152740/s1>, Figure S1: Pre-CRT serum CD133 measurements in RFU in relation to patient vital status at 12 months (black bars indicate that the patient is alive, blue bars indicate that the patient passed away prior to 12 months from diagnosis). Figure S2: Serum measurement levels of identified proteins in relationship to serum Albumin and serum CD133. Figure S3: CD133 protein value based on the 7k proteomic panel in relation to the time the samples were stored in the freezer from collection to analysis in days showing the signal measured in the CD133 protein value pre- (blue) vs. post-CRT (magenta), and the number of days in the freezer for each patient (orange). The CD133 expression does not correlate with the number of days the samples were stored in the freezer. Table S1: Complete Cox Proportional Hazards Model results with parameter estimates, *p*-values, and hazard ratios.

Author Contributions: Conceptualization, T.J. and A.V.K.; data curation, E.T., S.J., J.S., S.C., Y.Z., L.Z., T.C.Z., M.S. and M.M.; funding acquisition, K.C. and A.V.K.; investigation, T.J., E.T. and A.V.K.; methodology, T.J.; project administration, A.V.K.; software, T.J.; supervision, K.C. and A.V.K.; visualization, T.J. and A.V.K.; writing—original draft, T.J. and A.V.K.; writing—review and editing, T.J., E.T., S.J., J.S., S.C., Y.Z., L.Z., M.M. and A.V.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded in part by the NCI NIH intramural program (ZID BC 010990).

Institutional Review Board Statement: This study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board of the National Institutes of Health (NIH) (protocol code IRB00011862 and 19 April 2022).

Informed Consent Statement: Informed consent was obtained from all subjects involved in this study.

Data Availability Statement: Responsible data sharing is part of our mandate and the cardinal future direction of this dataset. We are currently aggregating several data aspects including additional samples and omic data and once this is complete, the set will be de-identified and shared. The goal is to provide one comprehensive GBM dataset that has clinical, omic and imaging data. Individual requests to access the datasets should be directed to Andra V. Krauze.

Acknowledgments: We would like to acknowledge the use of the references [91–93] in the graphical abstract figure as permitted in accordance with Wikimedia Commons use policy.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

GBM	Glioblastoma Multiforme
WHO	World Health Organization
SOC	Standard of Care
CRT	Chemoradiation Therapy
TMZ	Temozolomide
CSC	Cancer Stem Cell
PROM-1	Prominin-1
OS	Overall Survival
FS	Feature Selection
RFECV	Recursive Feature Selection, Cross-Validated
NIH	National Institutes of Health
IRB	Institutional Review Board
NIDAP	NIH Integrated Data Analysis Platform
ML	Machine Learning
FDR	False Discovery Rate
LR	Logistic Regression
RF	Random Forest
LASSO	Least Absolute Shrinkage and Selector Operator
ROC	Receiver Operating Characteristic
AUC	Area Under Curve
BIC	Bayesian Information Criteria
RFU	Relative Fluorescence Unit

References

- Nie, S.; Gurrea, M.; Zhu, J.; Thakolwiboon, S.; Heth, J.A.; Muraszko, K.M.; Fan, X.; Lubman, D.M. Tenascin-C: A novel candidate marker for cancer stem cells in glioblastoma identified by tissue microarrays. *J. Proteome Res.* **2015**, *14*, 814–822. [\[CrossRef\]](#)
- Baid, U.; Rane, S.U.; Talbar, S.; Gupta, S.; Thakur, M.H.; Moiyadi, A.; Mahajan, A. Overall Survival Prediction in Glioblastoma With Radiomic Features Using Machine Learning. *Front. Comput. Neurosci.* **2020**, *14*, 61. [\[CrossRef\]](#)
- Fernandes, C.; Costa, A.; Osório, L.; Lago, R.C.; Linhares, P.; Carvalho, B.; Caeiro, C. *Current Standards of Care in Glioblastoma Therapy*; Codon Publications: Singapore, 2017; pp. 197–241.
- Sun, T.; Chen, G.; Li, Y.; Xie, X.; Zhou, Y.; Du, Z. Aggressive invasion is observed in CD133(-)/A2B5(+) glioma-initiating cells. *Oncol. Lett.* **2015**, *10*, 3399–3406. [\[CrossRef\]](#)
- Mia-Jan, K.; Jung, S.Y.; Kim, I.Y.; Oh, S.S.; Choi, E.; Chang, S.J.; Kang, T.Y.; Cho, M.Y. CD133 expression is not an independent prognostic factor in stage II and III colorectal cancer but may predict the better outcome in patients with adjuvant therapy. *BMC Cancer* **2013**, *13*, 166. [\[CrossRef\]](#)
- Joyce, T.; Jagasia, S.; Tasci, E.; Camphausen, K.; Krauze, A.V. An Overview of CD133 as a Functional Unit of Prognosis and Treatment Resistance in Glioblastoma. *Curr. Oncol.* **2023**, *30*, 8278–8293. [\[CrossRef\]](#)
- Tseng, Y.-J.; Huang, C.-E.; Wen, C.-N.; Lai, P.-Y.; Wu, M.-H.; Sun, Y.-C.; Wang, H.-Y.; Lu, J.-J. Predicting breast cancer metastasis by using serum biomarkers and clinicopathological data with machine learning technologies. *Int. J. Med. Inform.* **2019**, *128*, 79–86. [\[CrossRef\]](#)
- Huang, W.; Zhong, Z.; Luo, C.; Xiao, Y.; Li, L.; Zhang, X.; Yang, L.; Xiao, K.; Ning, Y.; Chen, L.; et al. The miR-26a/AP-2 α /Nanog signaling axis mediates stem cell self-renewal and temozolomide resistance in glioma. *Theranostics* **2019**, *9*, 5497–5516. [\[CrossRef\]](#)
- Xie, Y.; Meng, W.-Y.; Li, R.-Z.; Wang, Y.-W.; Qian, X.; Chan, C.; Yu, Z.-F.; Fan, X.-X.; Pan, H.-D.; Xie, C.; et al. Early lung cancer diagnostic biomarker discovery by machine learning methods. *Transl. Oncol.* **2021**, *14*, 100907. [\[CrossRef\]](#)
- Müller Bark, J.; Kulasinghe, A.; Chua, B.; Day, B.W.; Punyadeera, C. Circulating biomarkers in patients with glioblastoma. *Br. J. Cancer* **2020**, *122*, 295–305. [\[CrossRef\]](#)
- Wang, J.; Bettegowda, C. Applications of DNA-Based Liquid Biopsy for Central Nervous System Neoplasms. *J. Mol. Diagn.* **2017**, *19*, 24–34. [\[CrossRef\]](#)
- Wen, P.Y.; Weller, M.; Lee, E.Q.; Alexander, B.M.; Barnholtz-Sloan, J.S.; Barthel, F.P.; Batchelor, T.T.; Bindra, R.S.; Chang, S.M.; Chiocca, E.A.; et al. Glioblastoma in adults: A Society for Neuro-Oncology (SNO) and European Society of Neuro-Oncology (EANO) consensus review on current management and future directions. *Neuro-Oncol.* **2020**, *22*, 1073–1113. [\[CrossRef\]](#) [\[PubMed\]](#)

13. Kudulaiti, N.; Zhou, Z.; Luo, C.; Zhang, J.; Zhu, F.; Wu, J. A nomogram for individualized prediction of overall survival in patients with newly diagnosed glioblastoma: A real-world retrospective cohort study. *BMC Surg.* **2021**, *21*, 238. [CrossRef] [PubMed]
14. Bakas, S.; Sako, C.; Akbari, H.; Bilello, M.; Sotiras, A.; Shukla, G.; Rudie, J.D.; Santamaria, N.F.; Kazerooni, A.F.; Pati, S.; et al. The University of Pennsylvania glioblastoma (UPenn-GBM) cohort: Advanced MRI, clinical, genomics, & radiomics. *Sci. Data* **2022**, *9*, 453. [CrossRef] [PubMed]
15. Wang, Z.; Jiang, W.; Wang, Y.; Guo, Y.; Cong, Z.; Du, F.; Song, B. MGMT promoter methylation in serum and cerebrospinal fluid as a tumor-specific biomarker of glioma. *Biomed. Rep.* **2015**, *3*, 543–548. [CrossRef] [PubMed]
16. Zhi, F.; Shao, N.; Wang, R.; Deng, D.; Xue, L.; Wang, Q.; Zhang, Y.; Shi, Y.; Xia, X.; Wang, S.; et al. Identification of 9 serum microRNAs as potential noninvasive biomarkers of human astrocytoma. *Neuro-Oncology* **2015**, *17*, 383–391. [CrossRef] [PubMed]
17. Zhao, H.; Shen, J.; Hodges, T.R.; Song, R.; Fuller, G.N.; Heimberger, A.B. Serum microRNA profiling in patients with glioblastoma: A survival analysis. *Mol. Cancer* **2017**, *16*, 59. [CrossRef] [PubMed]
18. Gold, L.; Walker, J.J.; Wilcox, S.K.; Williams, S. Advances in human proteomics at high scale with the SOMAscan proteomics platform. *New Biotechnol.* **2012**, *29*, 543–549. [CrossRef] [PubMed]
19. Tuerk, C.; Gold, L. Systematic Evolution of Ligands by Exponential Enrichment: RNA Ligands to Bacteriophage T4 DNA Polymerase. *Science* **1990**, *249*, 505–510. [CrossRef]
20. Palantir Foundry—The NIH Integrated Data Analysis Platform (NIDAP); NCI Center for Biomedical Informatics & Information Technology (CBIT); Software Provided by Palantir Technologies Inc. Available online: <https://www.palantir.com> (accessed on 10 May 2024).
21. Schober, P.; Boer, C.; Schwarte, L.A. Correlation Coefficients: Appropriate Use and Interpretation. *Anesth. Analg.* **2018**, *126*, 1763–1768. [CrossRef]
22. Taylor, R. Interpretation of the Correlation Coefficient: A Basic Review. *J. Diagn. Med. Sonogr.* **1990**, *6*, 35–39. [CrossRef]
23. Boateng, E.Y.; Abaye, D.A. A Review of the Logistic Regression Model with Emphasis on Medical Research. *J. Data Anal. Inf. Process.* **2019**, *07*, 190–207. [CrossRef]
24. Rigatti, S.J. Random Forest. *J. Insur. Med.* **2017**, *47*, 31–39. [CrossRef]
25. *MinMaxScaler*. Available online: <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html> (accessed on 10 May 2024).
26. *GridSearchCV*. Available online: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html (accessed on 10 May 2024).
27. Muthukrishnan, R.; Rohini, R. LASSO: A feature selection technique in predictive modeling for machine learning. In Proceedings of the 2016 IEEE International Conference on Advances in Computer Applications (ICACA), Coimbatore, India, 24–24 October 2016.
28. Liang, H.; Wu, J.; Zhang, H.; Yang, J. Two-Stage Short-Term Power Load Forecasting Based on RFECV Feature Selection Algorithm and a TCN-ECA-LSTM Neural Network. *Energies* **2023**, *16*, 1925. [CrossRef]
29. Sofaer, H.R.; Hoeting, J.A.; Jarnevič, C.S. The area under the precision-recall curve as a performance metric for rare binary events. *Methods Ecol. Evol.* **2019**, *10*, 565–577. [CrossRef]
30. Bowers, A.J.; Zhou, X. Receiver Operating Characteristic (ROC) Area Under the Curve (AUC): A Diagnostic Measure for Evaluating the Accuracy of Predictors of Education Outcomes. *J. Educ. Stud. Placed Risk (JESPAR)* **2019**, *24*, 20–46. [CrossRef]
31. *RepeatedStratifiedKfold*. Available online: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RepeatedStratifiedKfold.html (accessed on 10 May 2024).
32. *CoxPHFitter*. Available online: <https://lifelines.readthedocs.io/en/latest/fitters/regression/CoxPHFitter.html> (accessed on 10 May 2024).
33. Krämer, A.; Green, J.; Pollard, J., Jr.; Tugendreich, S. Causal analysis approaches in Ingenuity Pathway Analysis. *Bioinformatics* **2014**, *30*, 523–530. [CrossRef]
34. Cox, D.R. Regression Models and Life-Tables. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **1972**, *34*, 187–202. [CrossRef]
35. *Gaussian Mixture Models*. Available online: <https://scikit-learn.org/stable/modules/mixture.html> (accessed on 10 May 2024).
36. Grosse-Gehling, P.; Fargeas, C.A.; Dittfeld, C.; Garbe, Y.; Alison, M.R.; Corbeil, D.; Kunz-Schughart, L.A. CD133 as a biomarker for putative cancer stem cells in solid tumours: Limitations, problems and challenges. *J. Pathol.* **2013**, *229*, 355–378. [CrossRef]
37. Wu, B.; Sun, C.; Feng, F.; Ge, M.; Xia, L. Do relevant markers of cancer stem cells CD133 and Nestin indicate a poor prognosis in glioma patients? A systematic review and meta-analysis. *J. Exp. Clin. Cancer Res.* **2015**, *34*, 44. [CrossRef]
38. Zhang, W.; Chen, H.; Lv, S.; Yang, H. High CD133 Expression Is Associated with Worse Prognosis in Patients with Glioblastoma. *Mol. Neurobiol.* **2016**, *53*, 2354–2360. [CrossRef]
39. Popescu, I.D.; Codrici, E.; Albulescu, L.; Mihai, S.; Enciu, A.-M.; Albulescu, R.; Tanase, C.P. Potential serum biomarkers for glioblastoma diagnostic assessed by proteomic approaches. *Proteome Sci.* **2014**, *12*, 47. [CrossRef] [PubMed]
40. Nijaguna, M.B.; Schröder, C.; Patil, V.; Shwetha, S.D.; Hegde, A.S.; Chandramouli, B.A.; Arivazhagan, A.; Santosh, V.; Hoheisel, J.D.; Somasundaram, K. Definition of a serum marker panel for glioblastoma discrimination and identification of Interleukin 1 β in the microglial secretome as a novel mediator of endothelial cell survival induced by C-reactive protein. *J. Proteom.* **2015**, *128*, 251–261. [CrossRef] [PubMed]
41. Linhares, P.; Carvalho, B.; Vaz, R.; Costa, B.M. Glioblastoma: Is There Any Blood Biomarker with True Clinical Relevance? *Int. J. Mol. Sci.* **2020**, *21*, 5809. [CrossRef] [PubMed]

42. Chen, X.-Y.; Pan, D.-L.; Xu, J.-H.; Chen, Y.; Xu, W.-F.; Chen, J.-Y.; Wu, Z.-Y.; Lin, Y.-X.; You, H.-H.; Ding, C.-Y.; et al. Serum Inflammatory Biomarkers Contribute to the Prognosis Prediction in High-Grade Glioma. *Front. Oncol.* **2022**, *11*, 754920. [CrossRef] [PubMed]
43. Han, M.; Guo, L.; Zhang, Y.; Huang, B.; Chen, A.; Chen, W.; Liu, X.; Sun, S.; Wang, K.; Liu, A.; et al. Clinicopathological and Prognostic Significance of CD133 in Glioma Patients: A Meta-Analysis. *Mol. Neurobiol.* **2016**, *53*, 720–727. [CrossRef] [PubMed]
44. Abdoli Shadbad, M.; Hosseinkhani, N.; Asadzadeh, Z.; Brunetti, O.; Silvestris, N.; Baradaran, B. The Prognostic Value of CD133 in Predicting the Relapse and Recurrence Pattern of High-Grade Gliomas on MRI: A Meta-Analysis. *Front. Oncol.* **2021**, *11*, 722833. [CrossRef] [PubMed]
45. Ahmed, S.I.; Javed, G.; Laghari, A.A.; Bareeqa, S.B.; Farrukh, S.; Zahid, S.; Samar, S.S.; Aziz, K. CD133 Expression in Glioblastoma Multiforme: A Literature Review. *Cureus* **2018**, *10*, e3439. [CrossRef] [PubMed]
46. Abdoli Shadbad, M.; Nejadi Orang, F.; Baradaran, B. CD133 significance in glioblastoma development: In silico and in vitro study. *Eur. J. Med. Res.* **2024**, *29*, 154. [CrossRef]
47. Köhler, K.; Seitz, H. Validation Processes of Protein Biomarkers in Serum—A Cross Platform Comparison. *Sensors* **2012**, *12*, 12710–12728. [CrossRef]
48. Gatto, L.; Franceschi, E.; Di Nunno, V.; Tosoni, A.; Lodi, R.; Brandes, A.A. Liquid Biopsy in Glioblastoma Management: From Current Research to Future Perspectives. *Oncologist* **2021**, *26*, 865–878. [CrossRef]
49. Rakhshaninejad, M.; Fathian, M.; Shirkoobi, R.; Barzinpour, F.; Gandomi, A.H. Refining breast cancer biomarker discovery and drug targeting through an advanced data-driven approach. *BMC Bioinform.* **2024**, *25*, 33. [CrossRef] [PubMed]
50. Golestan, A.; Tahmasebi, A.; Maghsoodi, N.; Faraji, S.N.; Irajie, C.; Ramezani, A. Unveiling promising breast cancer biomarkers: An integrative approach combining bioinformatics analysis and experimental verification. *BMC Cancer* **2024**, *24*, 155. [CrossRef] [PubMed]
51. Mei, K.; Tan, M.; Yang, Z.; Shi, S. Modeling of Feature Selection Based on Random Forest Algorithm and Pearson Correlation Coefficient. *J. Phys. Conf. Ser.* **2022**, *2219*, 012046. [CrossRef]
52. Putro, I.H.; Ahmad, T. Feature Selection Using Pearson Correlation with Lasso Regression for Intrusion Detection System. In Proceedings of the 2024 12th International Symposium on Digital Forensics and Security (ISDFS), San Antonio, TX, USA, 29–30 April 2024.
53. Sarkar, J.P.; Saha, I.; Sarkar, A.; Maulik, U. Machine learning integrated ensemble of feature selection methods followed by survival analysis for predicting breast cancer subtype specific miRNA biomarkers. *Comput. Biol. Med.* **2021**, *131*, 104244. [CrossRef] [PubMed]
54. Liu, X.; Morelli, D.; Littlejohns, T.J.; Clifton, D.A.; Clifton, L. Combining machine learning with Cox models to identify predictors for incident post-menopausal breast cancer in the UK Biobank. *Sci. Rep.* **2023**, *13*, 9221. [CrossRef] [PubMed]
55. Han, W.; Qin, L.; Bay, C.; Chen, X.; Yu, K.H.; Miskin, N.; Li, A.; Xu, X.; Young, G. Deep Transfer Learning and Radiomics Feature Prediction of Survival of Patients with High-Grade Gliomas. *Am. J. Neuroradiol.* **2020**, *41*, 40–48. [CrossRef] [PubMed]
56. Lao, J.; Chen, Y.; Li, Z.-C.; Li, Q.; Zhang, J.; Liu, J.; Zhai, G. A Deep Learning-Based Radiomics Model for Prediction of Survival in Glioblastoma Multiforme. *Sci. Rep.* **2017**, *7*, 10353. [CrossRef] [PubMed]
57. Xie, C.; Yang, P.; Zhang, X.; Xu, L.; Wang, X.; Li, X.; Zhang, L.; Xie, R.; Yang, L.; Jing, Z.; et al. Sub-region based radiomics analysis for survival prediction in oesophageal tumours treated by definitive concurrent chemoradiotherapy. *eBioMedicine* **2019**, *44*, 289–297. [CrossRef] [PubMed]
58. Chen, D.; Goyal, G.; Go, R.S.; Parikh, S.A.; Ngufor, C.G. Improved Interpretability of Machine Learning Model Using Unsupervised Clustering: Predicting Time to First Treatment in Chronic Lymphocytic Leukemia. *JCO Clin. Cancer Inform.* **2019**, *3*, 1–11. [CrossRef]
59. Vivekanandan, T.; Narayanan, S.J. A Hybrid Risk Assessment Model for Cardiovascular Disease Using Cox Regression Analysis and a 2-means clustering algorithm. *Comput. Biol. Med.* **2019**, *113*, 103400. [CrossRef]
60. Atlas, T.H.P. The Human Protein Atlas. Available online: <https://www.proteinatlas.org/> (accessed on 10 May 2024).
61. Grillet, N.; Dubreuil, V.; Dufour, H.D.; Brunet, J.F. Dynamic expression of RGS4 in the developing nervous system and regulation by the neural type-specific transcription factor Phox2b. *J. Neurosci.* **2003**, *23*, 10613–10621. [CrossRef] [PubMed]
62. Echizen, K.; Nakada, M.; Hayashi, T.; Sabit, H.; Furuta, T.; Nakai, M.; Koyama-Nasu, R.; Nishimura, Y.; Taniue, K.; Morishita, Y.; et al. PCDH10 is required for the tumorigenicity of glioblastoma cells. *Biochem. Biophys. Res. Commun.* **2014**, *444*, 13–18. [CrossRef]
63. Mancia Leon, W.R.; Spatazza, J.; Rakela, B.; Chatterjee, A.; Pande, V.; Maniatis, T.; Hasenstaub, A.R.; Stryker, M.P.; Alvarez-Buylla, A. Clustered gamma-protocadherins regulate cortical interneuron programmed cell death. *eLife* **2020**, *9*, e55374. [CrossRef] [PubMed]
64. Pischedda, F.; Ghirelli, A.; Tripathi, V.; Piccoli, G. Negr1-Derived Peptides Trigger ALK Degradation and Halt Neuroblastoma Progression In Vitro and In Vivo. *Pharmaceutics* **2023**, *15*, 2307. [CrossRef] [PubMed]
65. Pischedda, F.; Piccoli, G. The IgLON Family Member Negr1 Promotes Neuronal Arborization Acting as Soluble Factor via FGFR2. *Front. Mol. Neurosci.* **2015**, *8*, 89. [CrossRef] [PubMed]
66. Naranjo, A.I.; González-Gómez, M.J.; Baladrón, V.; Laborda, J.; Nueda, M.L. Different Expression Levels of DLK2 Inhibit NOTCH Signaling and Inversely Modulate MDA-MB-231 Breast Cancer Tumor Growth In Vivo. *Int. J. Mol. Sci.* **2022**, *23*, 1554. [CrossRef] [PubMed]

67. Nueda, M.L.; Naranjo, A.I.; Baladrón, V.; Laborda, J. The proteins DLK1 and DLK2 modulate NOTCH1-dependent proliferation and oncogenic potential of human SK-MEL-2 melanoma cells. *Biochim. Biophys. Acta* **2014**, *1843*, 2674–2684. [CrossRef] [PubMed]
68. Li, L.; Zhang, C.; Wang, Z.; Guo, Y.; Wang, Y.; Fan, X.; Jiang, T. Expression changes in ion channel and immunity genes are associated with glioma-related epilepsy in patients with diffuse gliomas. *J. Cancer Res. Clin. Oncol.* **2022**, *148*, 2793–2802. [CrossRef] [PubMed]
69. Zannikou, M.; Duffy, J.T.; Levine, R.N.; Seblani, M.; Liu, Q.; Presser, A.; Arrieta, V.A.; Chen, C.J.; Sonabend, A.M.; Horbinski, C.M.; et al. IL15 modification enables CAR T cells to act as a dual targeting agent against tumor cells and myeloid-derived suppressor cells in GBM. *J. Immunother. Cancer* **2023**, *11*, e006239. [CrossRef]
70. Zhao, Y.; Yu, Z.; Song, Y.; Fan, L.; Lei, T.; He, Y.; Hu, S. The Regulatory Network of CREB3L1 and Its Roles in Physiological and Pathological Conditions. *Int. J. Med. Sci.* **2024**, *21*, 123–136. [CrossRef]
71. Liu, L.Q.; Feng, L.F.; Nan, C.R.; Zhao, Z.M. CREB3L1 and PTN expressions correlate with prognosis of brain glioma patients. *Biosci. Rep.* **2018**, *38*, BSR20170100. [CrossRef] [PubMed]
72. Fagerving, A. Nervous System—Brain 1—Smart-Servier.png. 16 November 2023. adapted with presmission from Wikimedia Commons. Available online: https://commons.wikimedia.org/wiki/File:Nervous_system_-_Brain_1_-_Smart-Servier.png (accessed on 10 June 2024).
73. Duck, U. Tumor MTK.jpg. 4 May 2022. Adapted with Permission from Wikimedia Commons. Available online: <https://commons.wikimedia.org/w/index.php?search=tumor&title=Special:MediaSearch&go=Go&type=image> (accessed on 10 June 2024).
74. Cheng, X.; An, J.; Lou, J.; Gu, Q.; Ding, W.; Droby, G.N.; Wang, Y.; Wang, C.; Gao, Y.; Anand, J.R.; et al. Trans-lesion synthesis and mismatch repair pathway crosstalk defines chemoresistance and hypermutation mechanisms in glioblastoma. *Nat. Commun.* **2024**, *15*, 1957. [CrossRef] [PubMed]
75. Ernst, M.E.; Baugh, E.H.; Thomas, A.; Bier, L.; Lippa, N.; Stong, N.; Mulhern, M.S.; Kushary, S.; Akman, C.I.; Heinzen, E.L.; et al. CSNK2B: A broad spectrum of neurodevelopmental disability and epilepsy severity. *Epilepsia* **2021**, *62*, e103–e109. [CrossRef] [PubMed]
76. Borgo, C.; D’Amore, C.; Sarno, S.; Salvi, M.; Ruzzene, M. Protein kinase CK2: A potential therapeutic target for diverse human diseases. *Signal Transduct. Target. Ther.* **2021**, *6*, 183. [CrossRef] [PubMed]
77. Stanzani, E.; Pedrosa, L.; Bourmeau, G.; Anezo, O.; Noguera-Castells, A.; Esteve-Codina, A.; Passoni, L.; Matteoli, M.; de la Iglesia, N.; Seano, G.; et al. Dual Role of Integrin Alpha-6 in Glioblastoma: Supporting Stemness in Proneural Stem-Like Cells While Inducing Radioresistance in Mesenchymal Stem-Like Cells. *Cancers* **2021**, *13*, 3055. [CrossRef] [PubMed]
78. Zhou, B.; Gao, S. Pan-Cancer Analysis of FURIN as a Potential Prognostic and Immunological Biomarker. *Front. Mol. Biosci.* **2021**, *8*, 648402. [CrossRef] [PubMed]
79. Ehlers, L.; Kuppe, A.; Damerau, A.; Wilantri, S.; Kirchner, M.; Mertins, P.; Strehl, C.; Buttgerit, F.; Gaber, T. Surface AMP deaminase 2 as a novel regulator modifying extracellular adenine nucleotide metabolism. *FASEB J.* **2021**, *35*, e21684. [CrossRef] [PubMed]
80. Gao, Q.Z.; Qin, Y.; Wang, W.J.; Fei, B.J.; Han, W.F.; Jin, J.Q.; Gao, X. Overexpression of AMPD2 indicates poor prognosis in colorectal cancer patients via the Notch3 signaling pathway. *World J. Clin. Cases* **2020**, *8*, 3197–3208. [CrossRef] [PubMed]
81. Tse, R.T.; Ding, X.; Wong, C.Y.; Cheng, C.K.; Chiu, P.K.; Ng, C.F. The Association between Spermidine/Spermine N(1)-Acetyltransferase (SSAT) and Human Malignancies. *Int. J. Mol. Sci.* **2022**, *23*, 5926. [CrossRef]
82. Brett-Morris, A.; Wright, B.M.; Seo, Y.; Pasupuleti, V.; Zhang, J.; Lu, J.; Spina, R.; Bar, E.E.; Gujrati, M.; Schur, R.; et al. The polyamine catabolic enzyme SAT1 modulates tumorigenesis and radiation response in GBM. *Cancer Res.* **2014**, *74*, 6925–6934. [CrossRef]
83. Akizu, N.; Cantagrel, V.; Schroth, J.; Cai, N.; Vaux, K.; McCloskey, D.; Naviaux, R.K.; Van Vleet, J.; Fenstermaker, A.G.; Silhavy, J.L.; et al. AMPD2 regulates GTP synthesis and is mutated in a potentially treatable neurodegenerative brainstem disorder. *Cell* **2013**, *154*, 505–517. [CrossRef] [PubMed]
84. Alkan, Z.; Duong, F.L.; Hawkes, W.C. Selenoprotein W controls epidermal growth factor receptor surface expression, activation and degradation via receptor ubiquitination. *Biochim. Biophys. Acta (BBA)-Mol. Cell Res.* **2015**, *1853*, 1087–1095. [CrossRef]
85. Wang, Z.; Li, S.; Xu, F.; Fu, J.; Sun, J.; Gan, X.; Yang, C.; Mao, Z. ncRNAs-mediated high expression of TIMM8A correlates with poor prognosis and act as an oncogene in breast cancer. *Cancer Cell Int.* **2022**, *22*, 177. [CrossRef] [PubMed]
86. He, Y.; Dössing, K.B.V.; Sloth, A.B.; He, X.; Rossing, M.; Kjaer, A. Quantitative Evaluation of Stem-like Markers of Human Glioblastoma Using Single-Cell RNA Sequencing Datasets. *Cancers* **2023**, *15*, 1557. [CrossRef]
87. Bourgonje, A.M.; Verrijp, K.; Schepens, J.T.; Navis, A.C.; Piepers, J.A.; Palmen, C.B.; van den Eijnden, M.; Hooft van Huijsduijnen, R.; Wesseling, P.; Leenders, W.P.; et al. Comprehensive protein tyrosine phosphatase mRNA profiling identifies new regulators in the progression of glioma. *Acta Neuropathol. Commun.* **2016**, *4*, 96. [CrossRef] [PubMed]
88. Tasci, E.; Shah, Y.; Jagasia, S.; Zhuge, Y.; Shephard, J.; Johnson, M.O.; Elemento, O.; Joyce, T.; Chappidi, S.; Cooley Zgela, T.; et al. MGMT ProFWise: Unlocking a New Application for Combined Feature Selection and the Rank-Based Weighting Method to Link MGMT Methylation Status to Serum Protein Expression in Patients with Glioblastoma. *Int. J. Mol. Sci.* **2024**, *25*, 4082. [CrossRef]
89. Krauze, A.V.; Sierk, M.; Nguyen, T.; Chen, Q.; Yan, C.; Hu, Y.; Jiang, W.; Tasci, E.; Zgela, T.C.; Sproull, M.; et al. Glioblastoma survival is associated with distinct proteomic alteration signatures post chemoradiation in a large-scale proteomic panel. *Front. Oncol.* **2023**, *13*, 1127645. [CrossRef]

90. Krauze, A.V.; Zhao, Y.; Li, M.C.; Shih, J.; Jiang, W.; Tasci, E.; Cooley Zgela, T.; Sproull, M.; Mackey, M.; Shankavaram, U.; et al. Revisiting Concurrent Radiation Therapy, Temozolomide, and the Histone Deacetylase Inhibitor Valproic Acid for Patients with Glioblastoma-Proteomic Alteration and Comparison Analysis with the Standard-of-Care Chemoradiation. *Biomolecules* **2023**, *13*, 1499. [CrossRef]
91. Amousey. Human-brain-vector.svg. 21 December 2022, Adapted with Permission from Wikimedia Commons. Available online: <https://commons.wikimedia.org/wiki/File:Human-brain-vector.svg> (accessed on 10 June 2024).
92. Ebrambot. OpenMoji-color 1F489.svg. 24 May 2018, Adapted with Permission from Wikimedia Commons. Available online: https://commons.wikimedia.org/wiki/File:OpenMoji-color_1F489.svg (accessed on 10 June 2024).
93. Fagerving, A. Equipment–Blood Test (8570)–Smart-Servier.png. 31 January 2023, Adapted with Permission from Wiki-media Commons. Available online: [https://commons.wikimedia.org/wiki/File:Equipment_-_Blood_test_\(8570\)_--_Smart-Servier.png](https://commons.wikimedia.org/wiki/File:Equipment_-_Blood_test_(8570)_--_Smart-Servier.png) (accessed on 10 June 2024).

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.