RESEARCH ARTICLE SPECIAL ISSUE

RSP Science Hub

International Research Journal on Advanced Science Hub
2582-4376
Vol. 05, Issue 05S May
www.rspsciencehub.com

Check for updates

http://dx.doi.org/10.47392/irjash.2023.S011

International Conference on intelligent COMPUting TEchnologies and Research (i-COMPUTER) 2023

Identifying Personalised treatment plan for GBM using Multidimensional Patient Similarity Analytics

Meera Varma[1], *Jereesh A S* [2]
[1]*Assistant Professor, Department of Computer Science and Engineering, Kannur University,India*
[2]*Cochin University of Science and Technology, India*

Email: meeravarma.cev@gmail.com

Article History

Received: 26 February 2023
Accepted: 15 March 2023
Published: 28 May 2023

Keywords:

Glioblastma Multiform;
Machine Learning models;
Clustering;
Patient similarity;
Cancer survival

Abstract

International Agency for Research on Cancer (IACR) reported an increase in the worldwide cancer rate which is now known to be a major impediment to increasing life expectancy. Glioblastoma multiform, further named as astrocytoma, is a fast-growing truculent type of brain tumour that develops in the cerebral hemispheres, mainly in the frontal and temporal lobes of the brain. According to the National Brain Tumor Society, GBM accounts for 49.1 percent of all primary malignant brain tumors. Despite advances in the available treatment options, there is not much improvement in overall patient survival rate and still ranges from 14.6 to 20.5months. Also, some individuals show adverse drug reactions due to their genetic composition, and the condition is called idiosyncrasy. The proposed work aims to find an effective treatment strategy for GBM patients on the basis of their clinical and genomic factors. The work is presented based on Genomic Data Commons (GDC), cBioportal and Cancer Browser dataset. Here we develop different patient cohorts based on the predictive features using K-means++ algorithm. A test patient acquires the treatment pattern of its most similar neighbour using patient similarity analytics. This is a generalized approach that can be applied to any disease class where personal traits have impact on overall survival.

1. Introduction

Brain Cancer has consistently been a leading cause of death worldwide. However, the emergence of the COVID- 19 pandemic is likely to make cancer care more difficult and pose new challenges. They can be either benign or malignant. Every cancer type is unique, so early diagnosis can improve the median survival rate. Glioblastoma Multiforme (GBM) is a primary brain tumor found in adults that's highly malignant and typically leads to just one year of survival post-diagnosis (Krex et al.), (Hanif et al.). As per the 2022 statistics from American Brain Tumor Association, GBM makes up almost half (49.1 percent) of all primary malignant brain tumors. The ratio of GBM prevalence is slightly higher in males compared to females (Hanif et al.). Glioblastoma comes in four variations: classical, neural, proneural and mesenchymal. These subtypes differ based on their genetic irregularities and the unique clinical features of each case (Varma and Jereesh), (W Verhaak et al.). Understanding the importance of personalized medicine and popularity of machine learning techniques in this field, we develop a treatment strategy for GBM patients considering their unique clinical and genomic characteristics. These characteristics may not necessarily rely on a specific method of treatment. Clustering method used to form patient clusters. We used the concept of patient similarity to recognize individuals who resemble a reference patient and use the information from comparable patient's records to generate customised predictions. Comparison of different clustering methods to cluster patients has been presented in the literature, which highlights the importance of selecting an appropriate clustering technique based on the nature and characteristics of the data. Also, we analysed different feature selection methods to generate the predictive feature list and the best method based on accuracy has been recommended.

2. Literature Survey

Glioblastoma Multiforme is most aggressive of all Glioma among the 4 grades. They are collection of tumors that originates within the central nervous system. According to Holland, Eric et al. (Holland and Multiforme) these gliomas are not cured by surgery alone because of its topologically diffuse nature. The standard treatment of GBM has been the same for many decades: surgical resection, radiation and chemotherapy. Even though many treatment approaches like gene therapy, infecting with viral vectors to kill tumor cells have been tested in animals for gliomas, but they seem to have no therapeutic effect in humans. So Machine Learning (ML) models that predict treatment option based on individual characteristics can improve overall the overall chances of survival.

Kunal Malhotra et al. (M et al.) developed a treatment plan for patients with glioblastoma where logistic regression model with forward feature selection method was used to extract 10 predictive features. A binary feature matrix with a target variable is formed from Clinical factors and genomic features and a target variable is formed based on patient survival period.

Kunal Malhotra et al. (Malhotra et al.) redesigned the initial model (M et al.) .Here they used logistic regression and Cox Regression model for prediction. Age, Karnofsky Performance Score (KPS), neo-adjuvant treatment history, MGMT methylation status, GABRA1 and TP53 gene expressions were identified as predominant features. Patients without date of diagnosis, pre-treatment history and missing values for drug duration were excluded (Varma and Jereesh). Greedy forward feature selection is used to extract predominant factors.

The system proposed by Ladha L et al. (Ladha and Deepa) suggested an empirical comparison of forward and backward feature selection methods and their algorithms. The forward selection starts with no variables and builds gradually, whereas backward selection works in reverse direction i.e. starts with complete feature set, and iteratively eliminates the irrelevant features, until the closure condition is met.

Among the different supervised ML algorithms used with forward and backward feature selection, the best performance is achieved when Support Vector machine(SVM) is used with Sequential Backward Selection(SBS) to extract the predictive features. The identified predictive features included Gender, vital status, neoadjuvant treatment history, MGMT gene methylation status, and EGFR, NEFL, PDGFRA, RELB and TNFRSF1A gene expressions (Varma and Jereesh).

When compared different variants of K-means clustering algorithm like x-means, global K-means and efficient k-means over colon and leukaemia datasets, initial choice of cluster centres plays a crucial role in determining quality of clusters (Kumar, Wasan, and Krishan). They found K-Means++ outperforms others due to its ability to select better cluster centres.

According to Shirkhorshidi et al. (Shirkhorshidi, Aghabozorgi, and Wah) similarity measures are main components of distance-based clustering algorithms. Commonly used distance metrics are Minkowski distance, Average distance, Euclidean distance, Chord distance, Manhattans distance, Jaccard index, Mahalanobis distance, Cosine Similarity and Pearson Correlation. Euclidean distance measure is widely used for numerical data.

The system developed by Panahiazar et al. (Panahiazar et al.) to recommend treatment pattern for Congestive heart failure(failure(CHF) patients by considering information's like lab results, age, gender, race, blood pressure readings, BMI, echocardiogram measurements and 26 co-morbid conditions collected from Electronic Health Record(EHR) data. They used patient similarity analytics to predict medication .Patient cohorts were formed using two techniques. In the first method, they used K-means and hierarchical clustering algorithm and in the second method, a supervised clustering approach was carried out. Finally, Mahalanobis distance is used to compute patient-cluster similarity.

Chen et al. (Chen, Su, and Chang) developed a model to suggest a treatment system for diabetes patient which followed a case based reasoning approach along with ontology. The system used lifestyle related information to form diabetes care ontology. The system was more focused on the clinical factors of the patient and no genetic information was considered. Since type 2 diabetes is linked with family history along with environmental factors, genetic information also needs to be taken into account. We need to create a rich CBR database to in order to identify similar patients. But it is hard to maintain such ontology.

A model that predicts treatment plans for GBM using clinical, biomedical and imaging data was created. The model utilizes the fuzzy C-means clustering algorithm and Wrapper feature selection method (Ershadi, Rise, and Niaki). But, the fuzzy C means algorithm takes longer computational time compared to other clustering algorithms.

(Ogbuabor and N) compared DBSCAN and K-means clustering algorithms on healthcare dataset and evaluated their performance based on silhouette score, clustering accuracy, and computational efficiency. They found Kmeans outperformed DBSCAN with a Silhouette score of 0.97. So it is advisable to use Kmeans or any of its advanced versions to create patient clusters and Euclidean distance as similarity measure.

3. Methodology

The primary objective of this work is to identify personalised treatment plan for Glioblastoma. The approach used was Multidimensional Patient Similarity Analytics of Glioblastoma patients based on their clinical and genomic profile. Figure 1 shows proposed model architecture. We collected data on patients with GBM and prepared it for analysis. We used SBS to identify 8 key factors. Based on their similarities in clinical and genomic characteristics, we grouped patients into clusters. When a patient is tested, they will receive treatment based on the treatment plan of the most similar one within their cluster. The system consists of 5 stages: 1) Data Collection 2) Data Standardization and Pre-processing 3) To find the predictive clinical and genomic feature. 4) Develop patient cohorts based on the predictive features. 5) Patient similarity assessment.

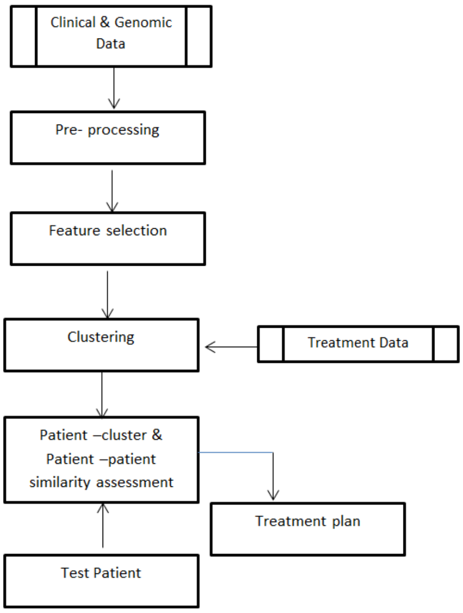

FIGURE 1. Proposed Model Architecture

3.1. Data Collection

We collected sample data of about 300 GBM diagnosed patients from GDC cBioportal (Cerami et al.) and Cancer Browser (Cline et al.). Clinical data consists of demographic information about the patients and valuable indicators regarding condition of patient. Genomic data were taken as genetics plays a pivotal role in drug responses which include Copy number variation of genes, mRNA expression levels and MGMT methylation status. Treatment data consist of sequence of drugs or therapies prescribed.

3.2. Data Standardization and Pre-processing

We removed samples with less than 50 percent data and standardized drug names. Some field like additional chemotherapy consist of values ’completed’ and ’not applicable/ not known’ replaced with binary values 1 and 0 respectively. We converted Beta and M-values to methylation status (Du et al.). Records with missing either start or end date of drug is removed.

3.3. Determine predictive clinical and genomic features

The data included both numeric and categorical data types. After data cleaning, a binary feature matrix was formed with a patient features. When SBS used for feature selection, 78 percent accuracy were obtained (Varma and Jereesh).

3.4. Develop patient cohorts based on the predictive features

The patients are categorized to different patient cohorts according to selected predictive features using a k-means++ clustering algorithm. k-means++ is an advanced version of k-means with better seeding. In order to find optimal value for k, Silhouette method was used (Rousseeuw), (Shahapure and Nicholas). Whenever a new patient comes, Euclidean distance between new sample and each cluster centroid was measured. The patient was allocated to the cluster with minimum distance with it and re-clusters each time.

3.5. Patient Similarity Assessment

The treatment features were sequence of drugs/radiation prescribed to patients. Patient-cluster and patient-patient similarity was estimated by using Euclidean distance. Patient -cluster distance is measured and a patient is assigned to cluster C with minimum distance. Then test patient similarity with all the other patients with survival as 1 (i.e. patient with median survival rate equal to more than 10 months) belonging to that particular cluster C was measured using Euclidean distance. The test patient adopted the treatment pattern of most similar patient.

4. Results

We have analysed about 235 patient samples diagnosed with GBM. 205 samples were used for training and 30 samples were used for testing. SBS is used to select the most predictive features, which were then given as input to an SVM classifier. Initially, there were 8 clinical and 29 genomic features, but after SBS was applied only 3 clinical and 5 genomic features remained as input for the SVM classifier. The outcomes showed a cross-validation accuracy of about 78 percent with both a 3-fold and 5-fold approach. Table 1 shows the 8 predictive features and their biological role. Patient samples were divided into different cohorts based on the predictive features. Frequently occurring treat-

TABLE 1. Predictive features and their biological role.

Predictive Feature	Role
Gender	Male/Female. Female patients with GBM have a higher cancer specific survival (CSS) after surgery (Tian et al.).
Vital Status	Living-last follow-up > 365 days and Living last follow-up < 365days.
History of neoadjuvant treatment	Yes/No - Patients receiving neoadjuvant treatment were found to have longer survival rate.
MGMT gene Methylation status	LM/M/HM: Abrasion in this region led to the loss of MGMT protein expression, which in turn reduces the strength to repair DNA damage (Rivera et al.), (Hegi et al.).
EGFR gene expression	mutation of EGFR called EGFRvIII was observed which enhance the tumor growth, migration, angiogenesis and metastatic spread its over expression led to decreased survival (Hatanpaa et al.), (Saadeh, Mahfouz, and I Assi), (Alentorn et al.).
PDGFRA gene expression	PDGFRA abnormalities were associated with GBM Proneural subtype (W Verhaak et al.). PDGFRA over expression have a negative impact on overall survival rate(OS) and progression free survival rate (PFS) (Alentorn et al.).
RELB gene expression	Patients with GBM mesenchymal subtype have increased RELB expression levels resulting in a shorter OS[28].
TNFRSF1A gene expression	Linked with immune cell infiltration of GBM and its high expression results in low survival in GBM patients (Wang et al.).

TABLE 2. Evaluation Results

Samples	Number of Samples
correctly predicted	22
incorrectly predicted	8

ment patterns within the minimum distance cluster were extracted from samples with a positive survival. The test patient will acquire the treatment pattern with largest frequency within the minimum distance cluster. If all the samples with a positive survival within a cluster have same frequency, then find Euclidean distance between the test patient and each of the candidate samples. Finally, the test patient will acquire the treatment pattern of most similar candidate patient. Out of the 30 samples used for testing, 22 samples were correctly predicted. We used prediction accuracy as a measure of performance. This is calculated by dividing the total number of correctly predicted instances by the overall number of instances (Table 2).

5. Comparison with Existing System

Heart failure therapy recommendation model developed by Panahiazar et al. (Panahiazar et al.) considered some patient specific variable as predictive features and patient clusters were formed using k-means and hierarchical clustering methods. Mahalanobis distance is used as the similarity measure. They obtained an accuracy of 71 percent with k-means clustering method and 73 percent with the hierarchical clustering method. The proposed method used k-means++ clustering method to form patient cohorts and Euclidean distance as similarity measure. We obtained an accuracy of 73.33 percent. Malhothra et al. (M et al.), (Malhotra et al.) developed a system to predict treatment plan for GBM patients and they used KPS score, gender, age, mRNA expression levels of some genes such as TP53,PIK3R1,NF1,EGFR and so on as predictive features. Recent study conducted by Wang et al. (Wang et al.) in 2022 identified that the TNFRSF1A gene expression levels in GBM cells is very high and have an impact on survival of GBM patients. According to the study conducted by Zeng et al.[28], patients with GBM mesenchymal subtype have increased RELB expression levels resulting in a shorter OS. We considered the expression levels of TNFRSF1A and RELB in our predictive feature set so that our system can better predict a optimal treatment plan. We used advanced K-means to form patient cluster which provide better convergence.

6. Conclusion and Future Scope

GBM, also referred as grade IV astrocytoma, is the most aggressive class of brain tumor which spreads rapidly with an average survival of nearly 10- 15 months. A major challenge in treating this fast growing cancer is to choose an ideal treatment strategy for patients after standard line of treatment. We identified the predominant clinical and genomic factors using SBS. Patients were divided into different cohorts using k-means++ algorithms. While using any variants of k-means algorithm, finding an optimal value for k is difficult. The best k-value is selected using Silhouette method. A patient similarity approach is used to extract a clinical and genomically similar patient from the study patient. We recommend a treatment pattern based on the treatments adopted by most similar patient. The proposed approach is generic and if a strong data set is available, it can be applied to any area of the disease in which clinical and genetic factors affect the survival rates. Due to the lack of sufficient data related to dosage of drugs, it is excluded from the study and can give better result if dosage of drugs is included. The accuracy of predictions heavily depends on the quality and size of the dataset used. Using a large enough dataset can result in better performance and more accurate predictions. We have limited the input features to clinical and genomic information. However, Data related to tissue analysis, imaging scans and disease trends along with information on proteins can contribute to better survival rate.

7. Acknowledgement

We are extremely grateful to Dr. Bill Wysocki, Scientific Support Analyst, and Centre for Data Intensive Science, University of Chicago, for the constant support provided in collecting necessary data for development of the system. We are indebted to Cancer Browser and cBioportal support team for the timely reply regarding dataset related queries.

References

Alentorn, A, et al. “Prevalence, clinico-pathological value, and co-occurrence of PDGFRA abnormalities in diffuse gliomas”. *Neuro-Oncology* 14.11 (2012): 1393–1403.

Cerami, Ethan, et al. “The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data”. *Cancer Discovery* 2.5 (2012): 401–404.

Cline, Melissa S, et al. “Exploring TCGA Pan-Cancer Data at the UCSC Cancer Genomics Browser”. *Scientific Reports* 3.1 (2013): 8–8.

Du, Pan, et al. “Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis”. *BMC Bioinformatics* 11.1 (2010): 587–587.

Ershadi, Mohammad Mahdi, Zeinab Rahimi Rise, and Seyed Taghi Akhavan Niaki. “A hierarchical machine learning model based on Glioblastoma patients’ clinical, biomedical, and image data to analyze their treatment plans”. *Computers in Biology and Medicine* 150 (2022): 106159–106159.

Hanif, F, et al. “Glioblastoma Multiforme: A Review of its Epidemiology and Pathogenesis through Clinical Presentation and Treatment. Asian Pac J Cancer Prev”. 18 (2017): 5563115–5563115.

Hatanpaa, Kimmo J, et al. “Epidermal Growth Factor Receptor in Glioma: Signal Transduction, Neuropathology, Imaging, and Radioresistance”. *Neoplasia* 12.9 (2010): 675–684.

Hegi, Monika E, et al. “¡i¿MGMT¡/i¿Gene Silencing and Benefit from Temozolomide in Glioblastoma”. *New England Journal of Medicine* 352.10 (2005): 997–1003.

Krex, D, et al. “Long-term survival with glioblastoma multiforme”. *Brain* 130.10 (2007): 2596–2606.

Kumar, Parvesh, Siri Wasan, and Krishan. “Comparative analysis of kmean based algorithms”. *International Journal of Computer Science and Network Security* 10.4 (2010): 314–318.

Ladha and T Deepa. “Feature selection methods and algorithms”. *International journal on computer science and engineering* 3.5 (2011): 1787–1797.

M, Kunal, et al. “Constraint based temporal event sequence mining for Glioblastoma survival prediction”. *Journal of Biomedical Informatics* 61 (2016): 267–275.

Malhotra, Kunal, et al. “Jimeng Constraint based temporal event sequence mining for Glioblastoma survival prediction”. *Journal of biomedical informatics* 61 (2016): 267–275.

Ogbuabor, Godwin and Ugwoke F. N. “Clustering Algorithm for a Healthcare Dataset Using Silhouette Score Value”. *International Journal of Computer Science and Information Technology* 10.2 (2018): 27–37.

Rivera, A L, et al. “MGMT promoter methylation is predictive of response to radiotherapy and prognostic in the absence of adjuvant alkylating chemotherapy for glioblastoma”. *Neuro-Oncology* 12.2 (2010): 116–121.

Rousseeuw, Peter J. “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis”. *Journal of Computational and Applied Mathematics* 20 (1987): 53–65.

Saadeh, Fadi S, Rami Mahfouz, and Hazem I Assi. “EGFR as a clinical marker in glioblastomas and other gliomas”. *The International Journal of Biological Markers* 33.1 (2018): 22–32.

Shahapure, Ketan Rajshekhar and Charles Nicholas. “Cluster Quality Analysis Using Silhouette Score”. *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)* (2020): 747–748.

Shirkhorshidi, Ali Seyed, Saeed Aghabozorgi, and Teh Ying Wah. “A Comparison Study on Similarity and Dissimilarity Measures in Clustering Continuous Data”. *PLOS ONE* 10.12 (2015): e0144059–e0144059.

Tian, Minjie, et al. “Impact of gender on the survival of patients with glioblastoma”. *Bioscience Reports* 38.6 (2018).

Varma, Meera and A S Jereesh. “Identifying predominant clinical and genomic features for glioblastoma multiforme using sequential backward selection”. *2017 International Conference on Circuit ,Power and Computing Technologies (ICCPCT)* (2017).

W Verhaak, Roel G, et al. “Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in PDGFRA, IDH1, EGFR, and NF1”. *Cancer Cell* 17.1 (2010): 98–110.

Wang, Xianggang, et al. “C1R, CCL2, and TNFRSF1A Genes in Coronavirus Disease-COVID-19 Pathway Serve as Novel Molecular Biomarkers of GBM Prognosis and Immune Infiltration”. *Disease Markers* 2022 (2022): 1–14.

Embargo period: The article has no embargo period.

To cite this Article: Varma, Meera, and Jereesh A S . “**Identifying Personalised treatment plan for GBM using Multidimensional Patient Similarity Analytics**.” International Research Journal on Advanced Science Hub 05.05S May (2023): 81–87. http://dx.doi.org/10.47392/irjash.2023.S011