Received: 12 December 2018 | Revised: 9 October 2019 | Accepted: 11 October 2019

DOI: 10.1002/cnr2.1226

REVIEW

Cancer Reports WILEY

Machine learning approaches to study glioblastoma: A review of the last decade of applications

Jessica Valdebenito[1] | Felipe Medina[1,2]

[1]Programa de Bioestadística, Escuela de Salud Pública, Universidad de Chile, Santiago, Chile

[2]Instituto de Estadística, Facultad de Ciencias, Universidad de Valparaíso, Valparaíso, Chile

Correspondence
Felipe Medina, Escuela de Salud Pública, Programa de Bioestadística, Av Independencia 939, Santiago, Chile.
Email: f.medina@uchile.cl

Funding information
Comisión Nacional de Investigación Científica y Tecnológica, Grant/Award Number: CONICYT Ph.D. fellowship 21151523

Abstract

Background: Glioblastoma (GB, formally glioblastoma multiforme) is a malignant type of brain cancer that currently has no cure and is characterized by being highly heterogeneous with high rates of re-incidence and therapy resistance. Thus, it is urgent to characterize the mechanisms of GB pathogenesis to help researchers identify novel therapeutic targets to cure this devastating disease. Recently, a promising approach to identifying novel therapeutic targets is the integration of tumor omics data with clinical information using machine learning (ML) techniques.

Recent findings: ML has become a valuable addition to the researcher's toolbox, thanks to its flexibility, multidimensional approach, and a growing community of users. The goal of this review is to introduce basic concepts and applications of ML for studying GB to clinicians and practitioners who are new to data science. ML applications include exploring large data sets, finding new relevant patterns, predicting outcomes, or merely understanding associations of the complex molecular networks presented within the tumor. Here, we review ML applications published between 2008 and 2018 and discuss ML strategies intending to identify new potential therapeutic targets to improve the management and treatment of GB.

Conclusions: ML applications to study GB vary in purpose and complexity, with positive results. In GB studies, ML is often used to analyze high-dimensional datasets with prediction or classification as a primary goal. Despite the strengths of ML techniques, they are not fail-safe and methodological issues can occur in GB studies that use them. This is why researchers need to be aware of these issues when planning and appraising studies that apply ML to the study of GB.

KEYWORDS

cancer, glioblastoma, machine learning, tumor, tunneling nanotubes

1 | INTRODUCTION

Glioblastoma (GB, formally glioblastoma multiforme) is one of the most aggressive types of brain cancer because of its rapid progression, poor response to treatment, and limited survival rate.[1] The current treatment includes surgical resection of the tumor if possible, followed by radiotherapy accompanied with temozolomide (TMZ) regimen[1-6]; however, tumor reappearance and resistance/adaptation to treatment are high. The poor prognosis has been associated with multiple factors including tumor heterogeneity, poor immune response, infiltration of the tumor into "healthy" tissue, generation of cancer stem cells, and the fast adaptation of the tumor to aggressive treatment. Therefore, there is an urgent need for novel therapeutic targets and drugs to cure or slow down this devastating disease.

Despite the significant accumulation of clinical, pathological, and omics data gathered by high screening platforms of thousands of

samples from healthy and GB individuals, in retrospective as well as ongoing cohorts, no new standard treatment has been approved.[2,7] We quote Micheel et al to define omics as "scientific disciplines comprising the study of related sets of biological molecules. Examples of omics disciplines include genomics, transcriptomics, proteomics, metabolomics, and epigenomics."[8] In general, the analysis of these large data sets is difficult for several reasons. First, the integration of data from different sources suffers from intrinsic heterogeneity at different levels, including biological sample type and recollection protocol, tumor development, and patients' genetic and environmental factors. Second, depending on the type of analysis, result interpretation may require additional adjustments, like false discovery rate corrections in genome-wide association studies.[9,10] Third, most of these analyses focus on association studies to identify DNA mutations or upregulated and downregulated genes present in the more abundant cell subpopulation, excluding those that belong to small cell populations such as GB stem cells (GBSC) or subcellular structures like tunneling nanotubes (TNTs), which are essential for tumor colonization, growth, and adaptation to treatment.[11-13]

The use of computer-intensive methods in data analysis such as machine learning (ML) is growing and has become a science on its own. ML corresponds to the study and development of algorithms and models for automated learning from large and multidimensional data. Currently, most biological studies involving the analysis of large data sets contemplate the use of computational methods such as ML.[14] The use of these techniques has multiple benefits, including exploiting complex relationships in high dimensional settings.[15] Evidence of the increasing popularity of ML-based strategies in cancer research is reflected in the increasing number of publications associated with the terms "cancer" and "machine learning" from 2008 to late 2018 (Figure 1).

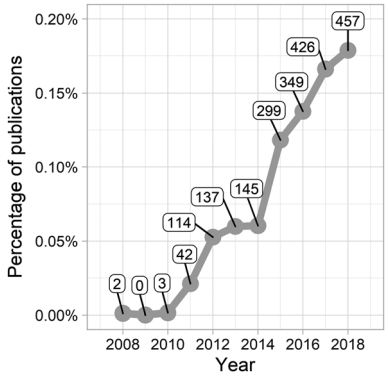

FIGURE 1 The increasing popularity of ML-based strategies in cancer research, as seen in PubMed from 2008 to 2018. The gray line represents the percentage of publications in PubMed's "cancer" subject that includes the MeSH term "machine learning", while the numbers shown in boxes correspond to the actual number of such publications. The information for doing this plot was downloaded on 26 April 2019 from PubMed's search results using the "Results by year" option. The search terms used were: cancer [sb] (for the number of publications in "cancer" subject) and "Machine Learning" [Mesh] AND cancer [sb] (for the number of publications in "cancer" subject that include the MeSH term "machine learning")

ML enables researchers to perform complex analyses of large databases to predict or to understand the pathogenesis of several diseases, including cancer.[16-18] As defined by Beam and Kohane, ML can be seen as a "continuum between fully human-guided *vs.* fully machine-guided data analysis" based on the level of specification of the assumptions built into the model.[19] Briefly, if the analysis considers a fully machine-guided study, the model would be able to learn a particular task with little or no human instruction. This facilitates the analysis of large data sets involving several variables, including the integration of different types of data for precision medicine applications.[20] In contrast, in a fully human-guided analysis, the model would consider variable specification, distribution, and relationships, among other considerations, as is the case when developing a prognostic model.[21]

Unlike traditional inference analysis, which focuses on parameter estimation based on sample data, ML algorithms use sample data either to predict an outcome based on a set of predictor variables (ie, a supervised learning problem) or to find patterns in the input data (ie, unsupervised learning problem). Another important difference is that most ML methods can be applied to examine high-dimensional data, such as when the number of variables surpasses the number of subjects or observations.[22]

Even though both traditional statistical inference and ML can be affected by study design, ML has the key advantage of using algorithms that can learn from the data while making minimal assumptions about the data-generating process.[22] For instance, most traditional statistical methods assume relationships between the outcome and the predictors that are either linear or not very complex (eg, quadratic, exponential, or logarithmic). In contrast, thanks to the use of efficient optimization algorithms, ML techniques can be more flexible models, which can exploit the complex relationships between outcomes and predictors to learn more about the underlying relationships in high-dimensional settings.

Here, we will describe different ML techniques that have been used over the last decade to study GB pathogenesis, and the positive and the negative aspects of the analyses. We will also discuss their limitations as well as suggest potential improvements based on new methodologies. Subsequently, validation practices for ML methods and applications for the examination of targeted cell populations and subcellular structures such as TNTs are challenged. Finally, figures to describe the more common ML techniques used in GB research for researchers starting in these ML techniques are provided and explained.

2 | SUPERVISED MACHINE LEARNING TECHNIQUES

Currently, most of the multiple clinical outputs of GB pathology, treatment, and clinical evolution have been studied using traditional statistical models based on pathology, clinical, molecular, imaging, and omics analyses.[9,23-28] Despite the success, these models are limited to exploring a restricted number of predictors simultaneously.

Conversely, computational models based on ML have been created to achieve these goals in high-dimension spaces, and therefore allow exploring a whole set of predictors simultaneously (see description below and differences in Figures 2–8).

Supervised models attempt to predict or classify an outcome of interest (the response variable) using the information from a pool of predictors (ie, explanatory variables). An example of response variable could be the primary outcome of a clinical trial. The response can be either quantitative (eg, survival time) or categorical with two or more possible outcomes (eg, good vs poor prognosis or the subtype of tumor). Then, supervised ML problems are classified into either prediction or classification problems if the response is quantitative or categorical, respectively. However, in basic and clinical sciences, researchers and clinicians have been more interested in the second one; that is, to classify patients in terms of prognosis or response to treatment.

2.1 | Support vector classifiers and support vector machines

Among the most commonly used ML methods for GB classification and clinical outcomes are support vector classifiers (SVCs). Starting with two groups of observations as input, SVCs search for a mathematical linear function in the high-dimensional data space that can separate both groups (Figure 2A-C). For example, SVCs can be used to predict tumor behavior (outcome), eg, proneural vs multiforme tumor subtype, based on omics data (input). The potential disadvantage of SVCs is that the outcome classes may not be linearly separable. In such cases, SVCs use a soft margin—which is why SVCs are also referred to as soft margin classifiers—that prevents model overfitting and subsequent loss of generalizability at the cost of risking misclassification of some observations as indicated in the examples shown in Figure 2D-F.[17] Thus, for this kind of classification, it is important that the data can be separated by a linear boundary; otherwise alternative methods need to be considered.

An example of SVC application to predict survival and molecular subtype in patients with GB was described by Macyszyn et al.[29] The authors fitted six SVC models using preoperative MRI data from a retrospective cohort to predict two different outcomes: OS (categorized) and the GB molecular multiclass outcome (neural, proneural, classical, and mesenchymal, according to one of their references). The authors demonstrated that SVCs could be extended for multiclass classification using separate SVCs; each one was used to discriminate among each of the four GB molecular subtypes and the remaining three, using the majority vote as the final prediction. Moreover, the authors combined the results from two SVCs to predict OS; one to distinguish between patients surviving less/more than 6 months, and another

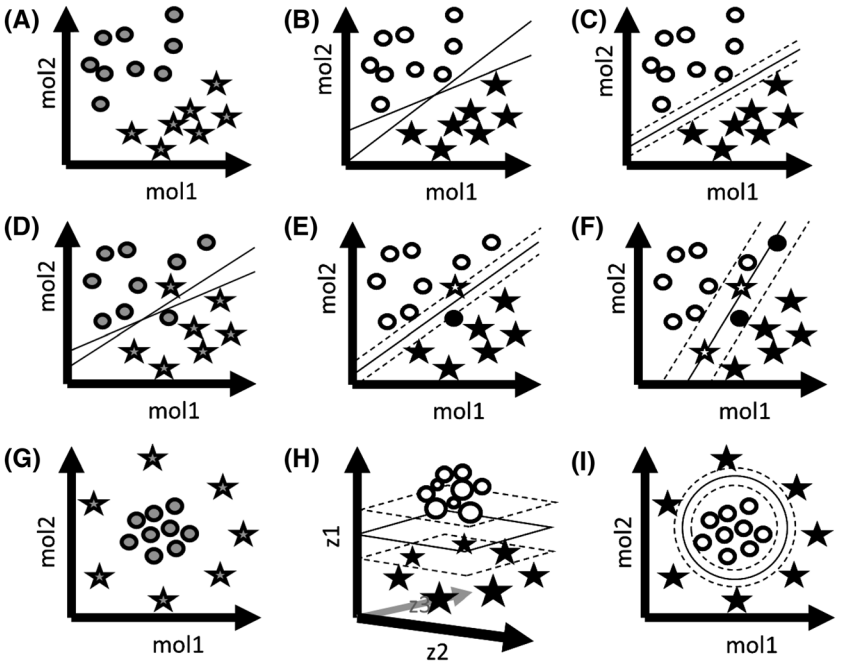

FIGURE 2 A hypothetical example of sample subtype classification using support vector classifiers (A-C), soft margin classifiers (D-F), and support vector machines (G-I). **A**, data from two different subtype samples, represented by circles and stars projected onto a plane made of two molecular features (mol1 and mol2). **B**, possible linear separation of both groups of subtype samples by two different hyperplanes (lines). Samples classified as circles are shown in white, whereas samples classified as stars are shown in black. **C**, the optimal hyperplane is the linear combination of molecular features (solid line) that maximizes the margin between itself and its closest observations (dashed lines). **D**, as in A, but now both groups of subtype samples cannot be perfectly separated by hyperplanes. **E**, allowing some classification error makes it possible to separate both groups by a hyperplane (solid lines) that maximizes the margin (dashed lines). **F**, the amount of classification error that is allowed affects the shape of the linear boundary that separates both groups. **G**, like in A, but now a linear boundary will not work well for separating both groups. **H**, the same samples but projected onto a possible expansion of the original feature space by using a kernel function. In this new space, both tumor subtypes groups can be separated by a hyperplane (solid tilted plane) which maximizes the margin between both groups (dashed tilted planes). **I**, in this example, hyperplanes from the expanded feature space look like ellipses in the original feature space

between less/more than 18 months, to generate a single "survival prediction index" that was ultimately used to classify patients into long, medium or short-term survivors. Using a separate prospective cohort, they obtained accuracies or percentages of correct predictions close to 80% for survival and 76% for GB molecular class. Overall, this example shows the flexibility of SVCs to combine its results to reach good predictions in multiclass settings, even though originally SVCs distinguish between two classes only.[29]

Although SVCs are sometimes referred to as support vector machines (SVM, discussed below), we prefer to make a distinction between them, as they achieve classification differently. SVMs (also known as kernel SVMs) become useful when the outcome cannot be separated by a linear boundary. SVMs are an extension of SVCs that use a kernel function, which is a mathematical transformation applied to the nonlinear/nonseparable data, accommodating data in an enlarged feature space where the separation of outcome classes is possible and translates into a nonlinear boundary in the original space (Figure 2G-I). The technicalities of this method are beyond the scope of this review, but further details can be found in Hastie et al.[18]

For example, Korfiatis et al[30] used SVMs with a particular type of kernel named Gaussian Radial Basis Function to predict already known *MGMT* gene methylation status in a sample of tumors based on potential biomarkers of MRI texture features (correlation, energy, entropy, and local intensity of *T2*-weighted images). The best SVM classifier, based on the four MRI texture features mentioned above, had a maximum area under the receiver-operating characteristic (ROC) curve of 0.85, with sensitivity and specificity at the optimal threshold of the ROC curve of 0.803 and 0.813, respectively. Using the model, SVMs could predict *MGMT* methylation status in preoperative GB tumors using MRI texture features and therefore assess prognosis in a noninvasive way.[30]

A different example of SVM application in brain tumor classification is the work of Metsis et al,[31] where the authors used SVMs to classify different brain tumor types (GB, anaplastic astrocytoma, meningioma, schwannoma, and adenocarcinoma), with the main objective of identifying potential tumor biomarkers based on two independent and heterogeneous data sets: magnetic resonance spectroscopic (MRS) metabolite and microarray gene expression information. After feature selection of MRS metabolite and microarray gene expression, tumor classification accuracy showed to be improved from 61.7% (gene expression data only) and 78.7% (MRS data only) to 87.2% when both types of data were used in the SVM model, although the results should be considered carefully because of some unclear qualitative aspects of their samples discussed below.

Finally, SVM-based classification can also be achieved by first using an SVM regression model to predict a quantitative response that is ultimately used for classification. As an example, Mao et al[32] used the regression approach to classify mutation types in GB and ovarian cancer based on structural and evolutionary features used by over 10 already known algorithms (eg, CHASM, SIFT, and MutationAssessor) that discriminate deleterious mutations from nondeleterious ones. To address the imbalanced number of driver and passenger mutations in the training set, the authors used a weighted version of SVMs, implemented in an annotation tool that they named CanDra. The use of weights into data allows unbalanced classes to be equally represented during the classification process; minority classes get higher weights, while majority classes get lower weights. Subsequently, the resulting outcome value was used to classify every mutation into one of three categories: driver (if its score greater than the 90th percentile), passenger (if less than the 10th percentile), or as a no call. More details about their implementation are available at https://bioinformatics.mdanderson.org/main/CanDrA. This method achieved values of the area under the receiver operating characteristic curve (AUC) of 0.911 and 0.941 in two independent validation datasets, which compared favorably with those obtained by other similar implementations (eg, CHASM with AUC of 0.890 and MutationTastor with AUC of 0.923). Thus, this implementation can achieve a good classification of cancer-type-specific driver mutations using a continuous outcome.[32]

2.2 | Decision trees and random forest classifiers

A different supervised learning technique for tumor classification is a decision tree (DT), which is an ordered set of splitting rules that sequentially segments the predictor space into simple regions. The goal of this ML technique is to obtain homogeneous regions with every new segmentation (see example in Figure 3).[17,33,34] Finding the best DT among all the possible ones is, generally, a computationally prohibitive problem; thus, a heuristic approach becomes necessary. A DT is usually allowed to grow, accrue splitting rules, and then it is pruned to prevent overfitting. Every observation within a region is then classified according to the most frequent class in that region, segmenting the data into simpler homogeneous subsets.

An example of DTs applied to the study of GB can be found in the work of Gollapalli et al,[35] who studied alterations in human serum proteome of healthy and GB patients to identify potential biomarkers for GB pathogenesis. The authors selected two sets of proteins after the analysis of 2DE data (five proteins) and 2D-DIGE data (19 proteins), and used each of these as the input for different DTs. The resulting DTs achieved 93.75% accuracy in the test group (eight healthy and eight GB patients for 2DE, 14 and 14 for 2D-DIGE). Interestingly, the authors found similar results when compared with other classification methods, such as SVMs and Naïve Bayes, indicating that DTs can discriminate healthy from GB patients as accurate as other known methods.[35]

Random forest classifiers (RFCs) are considered the next step in DT-based models (Figure 4). RFCs are an ensemble method, ie, a method that bases its results on the decisions obtained from a collection of methods that work on altered or perturbed versions of the original data set. In RFCs, the collection corresponds to a set of DTs that usually are not allowed to grow much (eg, no more than three splitting rules). Every DT is grown using a bootstrapped sample obtained by sampling with replacement from the original sample. Then, several DTs are grown to build the forest, and then cast a

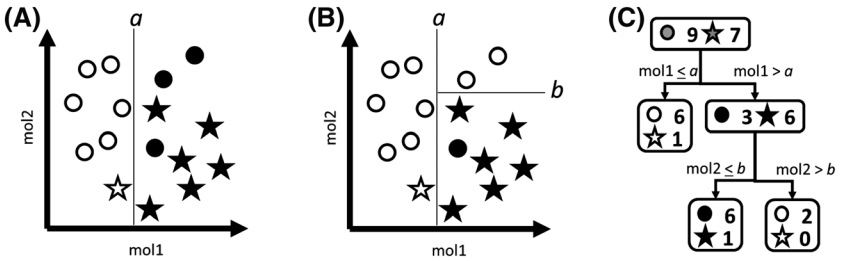

FIGURE 3 A hypothetical example of sample subtype classification using decision trees. A and B, data from two different subtype samples, represented by circles and stars, projected onto a plane of two molecular features (mol1 and mol2). **A**, a solid line at mol1 = *a* divides the feature space into two, and samples are classified according to the majority class in each of the resulting partitions. Subtype samples classified as circles are shown in white, whereas samples classified as stars are shown in black. **B**, another division is done within one of the previous partitions, and samples are classified according to the majority class of their group. **C**, a decision tree representation of B. The starting group is called the root node, while groups that are not further partitioned are called terminal nodes or leaves. The process of dividing groups by partitioning the feature space is called splitting, and it is generally done by considering one feature at the time to produce homogeneous nodes

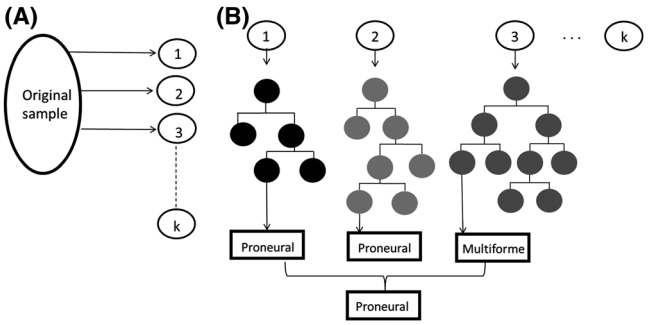

FIGURE 4 A hypothetical example of sample subtype classification using random forest classifiers. **A**, the original data is sampled with replacement *k* (hundreds or thousands) times, generating a diverse set of resamples. **B**, each resample is used to train a decision tree, creating a random forest. Classification of a subtype sample can then be done by the most voted prediction among the decision trees of the random forest

majority vote, which the RFC uses to predict classes. Usually, the growth of the forest is stopped when the results stabilize.

Several studies have used RFCs for cancer analysis.[30,36-39] As a representative example, Chang et al[37] used RFCs to identify potential biomarkers of carcinogenesis using imaging data as input to predict bevacizumab response in recurrent GB patients. The RFC was trained to predict patient OS based on pretherapy and post-therapy MRI data from a retrospective cohort of patients. The RFC was trained using a set of 84 patients and then tested in 42 patients. The results accuracy was variable depending on the input used to predict OS. Despite some sample limitations, their model, including both pre- and post-therapy features, showed hazard ratios of 5.10 and 3.64 in the training and testing cohorts, respectively. This proved the high reproducibility of their results and their utility in assisting with the clinical decision in patients with recurrent GB.[37]

In a different study, Kickingereder et al[36] used RFCs to classify molecular characteristics like methylation patterns and copy number variations in patients with newly diagnosed GB, using multiparametric and multiregional MRI features as input. The RFC method allowed the authors to identify and predict the *EGFR* amplification status (amplified vs nonamplified) and the *RTK II* glioblastoma subgroup ("classic" vs all other GB subgroups) with moderate accuracy (63% and 61%, respectively). However, these overall results were not sufficiently strong for reliable and clinical prediction of molecular features.[36]

Another example of RFC applied in GB research can be found in the work of Korfiatis et al.[30] These authors used a sample of GB tumors with already known *MGMT* methylation status to train an RFC model to classify such a state based on MRI-derived texture features. On the basis of a stratified fivefold crossvalidation, RFC showed an average AUC of 0.756 or 0.840, depending on whether the input corresponded to *T1* postcontrast or *T2* images, respectively. Their results exemplify how supervised ML techniques like RFC can be used to obtain noninvasive imaging correlate for *MGMT* methylation status in preoperative GB tumors.[30]

2.3 | Deep neural networks

Deep neural networks (DNNs) are a different kind of supervised method used in the study of GB, which allows working with nonbinary data. Inspired by biological neural networks, DNNs consist of a set of hidden interconnected layers of neurons or units that, just as neural networks do, attempt to accomplish a specific task, eg, pattern recognition (Figure 5). Once a neuron receives input from a source, these observations are multiplied by weights that shape the future activation function that generates the output. DNNs have the capability of learning throughout the process of net communication, and good training to adapt and improve its predictions.[16,40] Backpropagation is

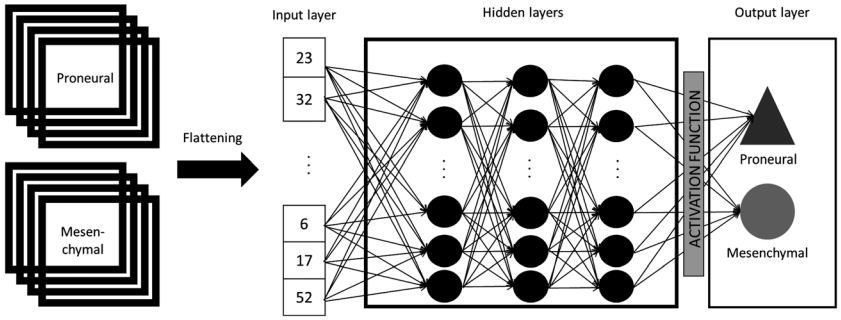

FIGURE 5 A hypothetical example of tumor subtype classification using deep neural networks. Data contain labeled T2-weighted MRI images of proneural and mesenchymal subtypes of the tumor. This fully connected multilayered neural network has one input layer, three hidden layers, and one output layer with two possible outcomes. Each neuron (black dots) is multiplied by its weight plus bias. Information on the weights is not shared by the neurons as in convolutional neural networks. Later the sum of these multiplications is passed to an activation function that defines the output of that neuron given a set of inputs. The learning process will find which weights and biases minimize the out-of-sample error measure

a common deep learning technique for training DNNs and allows searching for the best scheme of weights that will cause the neural network to have the lowest error for a training set because of the minimization of the total loss function.[16] The loss function reflects the performance of the neural network, meaning how well this model achieves the desired results.[17,18] When the expected result for a training set is not obtained, a potential solution is to modify the weights in backpropagation. This relatively recent application of neural networks is becoming more common in brain cancer research studies.[41,42] For instance, Mohsen et al[42] showed how DNNs could be used to classify brain MRIs into normal, or three types of malignant brain tumors (glioblastoma, sarcoma, and metastatic bronchogenic carcinoma). Their fitted model achieved an average rate of correct classification of 96.97% and an average AUC of 0.984, with these averages computed using sevenfold crossvalidation (a validation scheme that we describe later in this review).

A recent application of DNNs using magnetic resonance data from BRATS 2013 and BRATS 2015 is the one from Hussain et al,[41] who used convolution technology on artificial neural networks for tumor segmentation into three categories: enhancing tumor, core tumor (necrosis, nonenhancing, and enhancing tumor), and all tumor classes. This technology groups artificial neurons into overlapping regions based on how close these input neurons are to each other.[16] In the case of image analysis, group artificial neurons must respect the location of pixels and their proximity to emulate how biological eyes process images. In contrast to DNNs, convolutional neural networks (CNNs) do not consider every possible combination of weights but share some of them, helping computational resources when dealing with complex structures.

For model assessment, Hussain et al[41] evaluated the accuracy of their CNN-based algorithm using pixel-by-pixel sensitivity and specificity (correctly classified tumor and normal labels, respectively) and dice score. For a given label, the dice score is twice the number of pixels in which both predicted, and manually segmented labels coincide, divided by the sum of the total number of pixels that were predicted plus the ones manually segmented. Therefore, the dice score measures the overlapping predicted output image with the manually segmented labels. The authors obtained high dice scores for the complete (0.86), core (0.87), and enhancing (0.90) tumor labels. However, the most important result of this study was that CNNs outperformed the "state-of-the-art" techniques in the analysis of BRATS 2013 and BRATS 2015 datasets.[43] For instance, for BRATS 2013 database, the best implementation was one that used concatenated features for RFC, which achieved dice scores lower than Hussain et al.[41] CNN implementation for labeling core (0.78 vs 0.89, respectively) and enhancing tumor regions (0.74 vs 0.92, respectively), had comparable results for labeling the whole tumor region (0.87). In the case of BRATS 2015, one of the best previous methods was an implementation of deep CNN[44] that focused on global information in contrast to the patch-based approach of Hussain et al. Comparing these two approaches to analyze BRATS 2015 data, the latter approach showed a higher dice score for the enhancing label (0.90 vs 0.72, respectively), similar dice scores for core (0.87 vs 0.85, respectively), and lower for complete (0.86 vs 0.92, respectively). Thus, deep learning shows us a complex and interesting branch of ML with successful applications on image analysis for cancer research pattern recognition.

3 | UNSUPERVISED MACHINE LEARNING TECHNIQUES

Another approach to learn from data is unsupervised learning, a set of ML techniques that are mostly focused on the data structure, like the relationships between variables, rather than predicting a particular outcome, with the prospect of discovering patterns that lead them to new hypotheses. In contrast to supervised learning, this method of learning lacks supervision as there is no main outcome to predict or classify; hence the name "unsupervised." Overall, unsupervised ML techniques focus on learning about the structure of the data, producing results that may be used as the input of further analyses as we describe below.

3.1 | Principal component analysis

A popular unsupervised ML technique for analyzing quantitative features is principal components analysis (PCA). This technique organizes a large set of correlated features into a smaller number of

representative variables called principal components (PCs) that together explain most of the variability originally observed in the sample. PCs are constructed as linear combinations of the original features such that they contain most of the variability of the data and are linearly uncorrelated (see example in Figure 6).[45] An example of PCA application is the work of Akbari et al[46] where PCA was used to preprocess the time series contrast-enhanced MRI data down to a few PCs that could capture most of the information on the temporal dynamics of blood perfusion. PCA, in this case, served as a good data reduction technique as it captured more than 99% of the variance in the perfusion signal, so it quantified almost all the subtleties that these time curves store. Moreover, each of these PCs conveyed different characteristics of the perfusion signal as inferred from a visual exploration of them as a function of time.

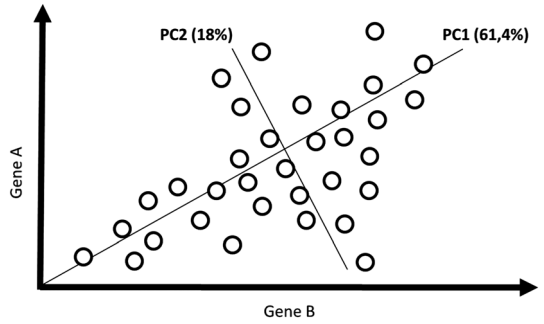

FIGURE 6 A hypothetical example of gene expression data exploration using principal components analysis. Each dot represents a subtype sample (e.g., tumor subtype) against its expression levels for two genes. PCA captures two principal components (PC1 and PC2) containing most of the variability in the data as indicated by the lines. PC1 accounts for most of the variability of the data (61,4%), while PC2 becomes the second to contain most of the variability (18%) of the explored genes

Another use of PCA is to perform data visualization. The main idea is that, if the first few PCs explain most of the variability observed in the data, then the projection of the data onto the first few PCs should result in a low-dimensional representation that captures most of the original information. As an example, Madhavan et al[47] used PCA to assist real-time data exploration and biological hypothesis generation in the analysis of the Repository of Molecular Brain Neoplasia Data (Rembrandt). PCA can be used to ordinate gene expression data in two- or three-dimensional graphs that plot each sample as a point at coordinates defined by the PCs, which in this case are low-dimensional summary representations of the gene expression data. Further visualization options, like filtering, sample coloring, and selection, help the user to explore data in ways that would not be practical to do gene-wise. Nowadays, this use of PCA can be found in the Georgetown Database of Cancer or G-DOC platform (Georgetown University, https://gdoc.georgetown.edu/gdoc/), which encompasses a diverse collection of cancer datasets and analysis tools to enable the integrative analysis of multiple data types to understand their disease mechanisms.

3.2 | Clustering methods

If the objective of an analysis is to find groups of instances (eg, individuals) or clusters within the data, a collection of unsupervised ML techniques known as clustering methods can be used. Most of these techniques focus on finding discrete homogeneous groups of individuals such that members within each cluster are similar to each other by their features, as well as members from different clusters are dissimilar to each other[48] (see Figure 7). Normally, a hierarchical algorithm output corresponds to hierarchical relationships between observations in the dataset, such that at each level of the hierarchy, clusters within the same group are more alike than those in different groups (see example in Figure 7H). Starting with every instance in a

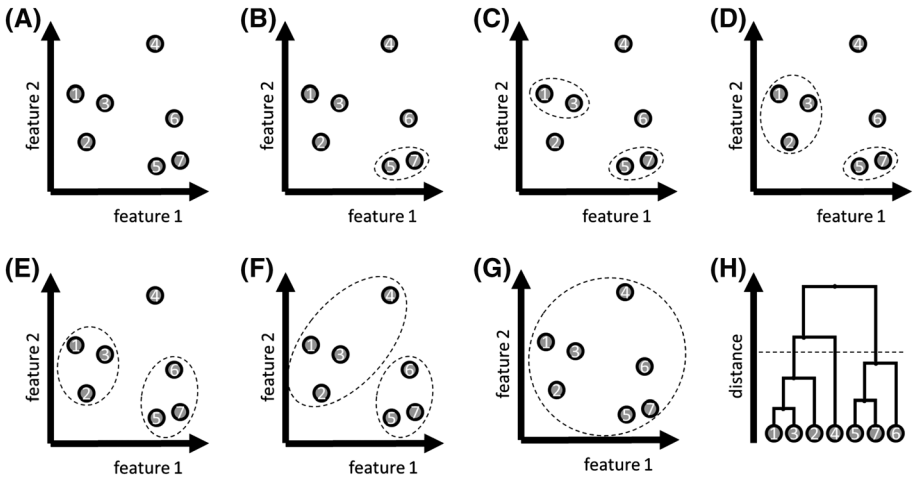

FIGURE 7 A hypothetical example of subtype data clustering using a hierarchical agglomerative algorithm. **A**, data from seven tumor samples projected onto a plane made of two features. **B–G**, samples, and groups of samples are sequentially grouped, joining those who are closer in terms of a metric or distance measure and a linkage criterion. In this example, the distance measure is the Euclidean distance while the linkage criterion is based on linking clusters with the closest centroids first. **H**, dendrogram representation of the clustering procedure. If the distance represented by the dashed line is used as a threshold for deciding the number of clusters, we would then identify three clusters: one made of samples 1, 2 and 3, another made of samples 5, 6 and 7, and sample 4 in its own (singleton) cluster

different cluster, this algorithm iteratively combines the least dissimilar pair of groups according to a metric (a function describing the distance between pairs of observations) and a linkage criterion (a rule for computing distances between pairs of groups).[48]

An example of a hierarchical clustering (HC) technique applied to GB research is seen in the work of Bredel et al,[49] where they applied a two-way complete linkage clustering based on Pearson correlation as the distance metric to discover both GB samples and landscape genes that showed similar gene alteration patterns. "Two-way" means that the algorithm was applied to a group individuals based on features, and also to group features based on individuals; "complete linkage" means that distances between groups are computed as the maximum pairwise distance observed between the instances from one group and the instances from the other group, using Pearson correlation coefficient as a metric between instances. The authors found two subgroups of GB samples that showed distinct profiles of chromosomal alterations. These results led them to generate and test further hypotheses, showing once again that unsupervised ML techniques can be used for data exploration and hypotheses generation.[49]

3.3 | Association rules mining

Another technique for finding structural patterns in an unsupervised way is the association rules mining (ARM). Originally popular in market research, ARM can identify frequent association rules (ARs) of the form "if antecedents then consequents" from a set of lists containing the items that are candidates to be antecedents and consequents.[18] Although ARM is not as popular in cancer research as the other techniques, Cremaschi et al[50] used ARM to assist molecular-based cancer research. In the study, the authors gathered information from different cancer-related datasets found in the GEO Datasets Archive (http://www.ncbi.nlm.nih.gov/gds); subsequently they obtained the pairwise comparisons between the tumor and normal tissue samples from those data sets, and finally obtained a list of differentially expressed probes of long noncoding RNAs (lncRNAs) that were seen in each of those pairwise comparisons. Then they used these lists as input for the ARM algorithm and identified 102 nonredundant ARs, from which they selected the one with the highest number of differentially expressed lncRNAs. This last AR was found to be present in comparisons from a subset of datasets of different kinds of human brain tumors and had 13 lncRNAs, of which 10 were found to be consistently up or downregulated in another gene expression database obtained from the ArrayExpress archive (https://www.ebi.ac.uk/arrayexpress/). These 10 lncRNAs were further investigated with PCA to assess if they could help to distinguish between brain tumor and normal samples, finding that this was the case. A further comodulation analysis followed by a gene enrichment analysis was done and used to find that biological processes specific to the nervous system could be compromised. This is a clear example of how unsupervised ML techniques can be used to assist molecular-based cancer research.[50]

3.4 | Deep learning models

Deep learning models (DLMs) can also be used to learn about the hierarchical structure of data in an unsupervised learning setup. For this, a DLM is composed of multiple layers of latent variables (ie, hidden layers). Each layer learns from the original data, and an alternative representation of it, varying the degree of complexity within each layer. In GB research, Young et al[51] applied a DLM to cancer gene expression data hypothesizing that a hidden layer was likely to represent the activation state of the signaling systems that regulate transcriptomic activities in cancer cells. These authors found that the number of hidden elements in the first layer agreed with the number of human transcription factors and, therefore, favored the interpretation of this layer as a representation of the transcription factors used by cancer cells. Then, using six similar DLMs, they clustered GB samples based on gene expression, identifying six clusters of samples that showed differences in patient survival. Further investigation of the associations of these clusters with differentially expressed genes and mutations led the authors to find many genes with functions or mutations relevant to cancer. This example shows that it is possible to study cellular signaling pathways without the need to concentrate on a handful of (hypothesis-driven selected) molecular agents, but instead, studying all the molecular agents that are measurable at once. By using unsupervised DLM on the whole transcriptomic mixture of expressed genes, regulated by active pathways, Young et al were able to show how the hierarchical decomposition ability of this method could reveal subtypes (clusters) of GB that encoded clinically relevant information.[51]

To conclude this section, it is noteworthy to mention that unsupervised or semi-supervised methods (ie, having a portion of the data already labeled) tend to be less expensive regarding the computational resources and time they demand, whereas using supervised methods for predicting or classifying data can take a longer time to run.

4 | DISCUSSION OF MODEL ASSESSMENT, VALIDATION, AND LIMITATIONS

As indicated in Figure 9, an important step after training an ML model is assessment and validation. To that end, different measures of performance and strategies may be used. Besides reviewing some basic aspects of model assessment and validation, the focus of this section is to discuss several aspects that need consideration as they can pose significant limitations that cannot be overcome by ML techniques and need the researcher's attention and evaluation.

4.1 | Appropriate use of machine learning nomenclature

Because of the current increase in popularity and use of ML methods in cancer research, it is important that researchers in health sciences

become knowledgeable of the appropriate use of the ML nomenclature to avoid misunderstandings and promote clear and reproducible research.[52] Misunderstandings can happen because of misconceptions, like inaccurate definitions or misinterpretation of the results achieved. For instance, clustering techniques used for finding a priori unlabeled groups are unsupervised learning techniques, and thus it would be wrong to present them as supervised, even if a posteriori the group output is used to define classes within the data. Another source of potential misunderstanding is interpreting the results of a methodology beyond their intended scope. For example, crossvalidation does not guarantee that results are generalizable to other samples; in fact, generalizability must be sought by doing replication studies with the same target population.

4.2 | Is there a best technique to learn from the data?

After explaining different ML techniques, the readers may have asked themselves if there is one method that surpasses the rest. For instance, we can argue that DNN is better than SVMs because the first can work directly with complex image data structures, whereas the second needs the user to handcraft input features (like the preprocessed features as described for Akbari et al[46] work). In this respect, DNNs are considered better because they learn directly from the data in a hierarchical manner, as its hidden layers "do" the preprocessing.[16]

Although the flexibility to learn from highly complex data structures is a desired property, sometimes restrictive but simpler models can perform sufficiently well and have easier or more straightforward interpretations regarding the model. Interpretability is then another desirable characteristic of model assessment. Ultimately, it must be remembered that if we intend to predict an outcome, our efforts should concentrate on choosing the most appropriate variables and data sets to achieve the best prediction results for our ML model. Conversely, if instead of predicting an outcome the goal is to assess an explicit association based on a predefined biological hypothesis, then the use of an explanatory model and traditional statistical inference may be of more interest for finding the model that fits best to our theoretical hypotheses.[53]

4.3 | Model assessment and model validation

Although we did not mention it, models fitted through ML techniques are assessed in different terms depending on whether they are supervised or unsupervised, and whether the outcomes are continuous or categorical. For instance, classifiers like SVCs, SVMs, DTs, and RFCs are assessed regarding quantities that describe their classification quality, very similar to how medical screening tests are assessed. Therefore, classifiers obtained by using different inputs, or even different ML techniques, can be compared regarding their sensitivity, specificity, AUC, and misclassification rates as described.[17,18,54]

Of course, model assessment based on the very same dataset that was used to train the model is likely to result in overoptimistic results (ie, overestimate the classifier accuracy). To prevent this problem, different model validation strategies are required.[17,18,55] Considering the examples described above, the most popular approaches for model validation are: (a) testing the model in a dataset that is independent of the training dataset and (b) performing a subsampling validation scheme known as crossvalidation. A testing set independent of the training set may be obtained by splitting the original sample or by sampling the same study population from which the training set was sampled, although this increases the costs of the study in most cases.

Crossvalidation is a general algorithm that tries to balance the benefits of splitting the dataset in training and testing sets to have a better estimation of the accuracy of the trained model, with the benefits of gaining fitting precision by using the whole sample to train the model. For this, crossvalidation consists of dividing the dataset into k subsets or folds (with k usually 5 or 10), leaving each subset out at a time and using the rest to train a model that is tested with the set that was left out at that time (Figure 8). This crossvalidation produces k sets of estimates of model accuracy, which gives the researcher a better idea of the accuracy of the method. An alternative to this method is to use classical goodness of fit measures, such as the Akaike information criterion (AIC), Bayesian information criterion (BIC), Mallow's Cp or adjusted R^2 as described.[17,53] For the sake of reproducibility and comparability across studies, we suggest that researchers use more than one measure for assessing their models, including (if possible) the ones that have already been used in past studies.

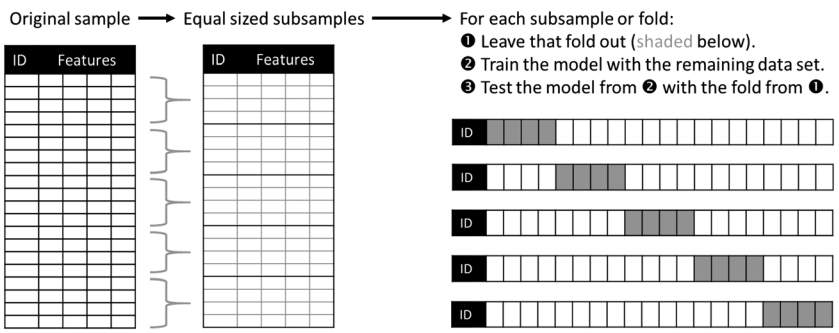

FIGURE 8 A 5-fold cross-validation toy example. A simple way to assess how well the results from a model will generalize to independent data is to create equal-sized subsamples or folds from the original data set and use each of these folds to test a statistical model that has been trained with the data that was not part of that respective fold. By the end of this procedure, there will be as many model assessments as folds, which helps to assess how well the applied model will generalize to new data

4.4 | Preventing overfitting

If the investigators are working with parametric models in a high-dimensional setting, overfitting can easily happen because of the number of parameters being estimated. If left unaddressed, this problem can result in high estimated variances for the coefficients of the model. Moreover, overfitting hinders the external validity or generalizability of the model results, resulting in poor predictions, classification, or consistency of the discovered structures when the model is used in new and independent data sets. To address this point, it is always recommended to address overfitting, for which simple strategies like splitting the dataset or crossvalidation can be a good starting point as described.[17,18,53]

It must be noted that not all ML techniques are equally prone to the issue of overfitting. For instance, although DTs can be affected by overfitting, random forests as an ensemble of DTs are less affected by this problem. Alternatively, some techniques may be regularized to prevent overfitting. For example, in neural networks, regularization usually involves modifying the weights of the neural network as the training is carried out. In DNNs, dropout is another commonly used regularization technique, which consists of removing neurons as the training progresses to prevent the network from becoming overly dependent on any single neuron.[16]

4.5 | Dealing with missing data

Having to deal with missing data can occur when working with large databases, for example, missing information on mutation sites or clinical information of a patient because of a variety of reasons. This is typical of high-dimensional space, either biological or clinical data sets, where we can either ignore or impute the missing data. The main recommendation is to investigate the missing data mechanism, which could be missing at random (MAR), missing completely at random (MCAR), or missing not at random (MNAR). The worst case is when data is MNAR because it cannot be controlled by the researcher, the true population is misrepresented, and thus future predictions based on the sample are not reliable. For this reason, under the MNAR mechanism, we do not recommend imputation, but to describe how the data was collected and the reasons for missingness of the data.

When observations are MCAR, omission or imputation of data could be a viable alternative. If we consider MCAR as a random sample of the data, omitting those values will not generate bias, but we must be careful because reducing the sample size can reduce accuracy and precision. Imputation can be performed in a classical procedure, which is considering the mean or median of nonmissing observations for that variable. A different approach could also refer to regression models for imputation, which consider predicting a value for the missing observation by using a regression or ML model as described by Ishwaran et al on the adaptative tree imputation.[56]

Regarding MCRA and binary data, we agree with Hastie et al[18] that one way to diagnose that missingness is not happening entirely at random is to include an indicator of missingness as a predictor and check if this helps to predict the outcome. Furthermore, different approaches can be used to deal with this problem. For example, in Mao et al.,[32] when selecting the best features for their SVM model, they also applied an algorithm to deal with missingness called the *k*-nearest neighbor algorithm, which replaces missing values using the information from the *k* nearest mutations in the same gene being studied.

Another example regarding MAR mechanism is found when working with RNA-seq data, which typically has observed zeros due to technical (ie, batch effects) and biological variation. One way to tackle the problem of an incomplete data matrix is to use an imputation method that considers the distribution of counts in RNA-seq data. For example, some proposed methods that comply with this are the Poisson mixture model, Bayesian Poisson regression, and zero-inflated Poisson.[57] There are many options to consider for missing data imputation and as Baghfalaki et al[57] suggest, using the one which outperforms the others in a training-testing approach can be considered as a suitable approach to select one.

4.6 | How to deal with representation in low cell populations?

In general, profiles of small populations of cells, such as cancer stem cells, are masked because of their low abundance in the bulk of the tumor that is composed mostly of proliferating cells and immune cells. Both conventional and ML-based statistical methods cannot resolve this problem because the signals of interest are weakened because of the measurement limitations of the experimental design. This problem is a direct consequence of searching for associations between survival and gene expression data derived from whole blood samples instead of, for example, cancer stem cells.[9] The sample cells of interest are diluted in a larger population of cells.

For instance, in RNA-seq data, the low relative abundance of cancer stem cells can result in missing read counts for genes that are differentially expressed in the underrepresented cell subpopulation. If this is the case, imputation methods could be used as an alternative approach to "rescue" some of the signals. Using this imputed data matrix, ML methods will enable the processing of all the available information simultaneously to generate a fuller conceptual framework. Most of the development of these methods is required and urgent for targeted exploratory identification of novel biomarkers of disease.

For example, in GB and other types of tumors, a rare subpopulation of cancer stem cells or subcellular structures such as TNTs has been described.[11-13] In the area of GBSC, this cell population is characterized by their self-renewing proliferation properties, resistance to treatment, and differentiation into multiple cell types. Moreover, these cells give rise to new tumors after therapeutic eradication of the primary tumor. Thus, it is essential to understand their properties and to identify their contribution to tumor development, treatment resistance, and survival. To study this small population, it is better to

consider the hierarchical structure of this subpopulation of cells while designing the study. Thus, using exclusive data coming from the bulk of the tumor is not enough, and single-cell data need to be considered. The adaptation of the ML methods to work with an omic hierarchical structure is a research area of interest in GB treatment and is still in development.

4.7 | How to deal with class imbalance?

Another issue related to representativeness is class imbalance. This problem arises when the classes to be predicted have relative frequencies that are too uneven, making the training model biased towards favoring the majority class. The latter occurs because the trained model seems to be more accurate, even though it fails to detect the rarer instances that are of interest to detect in the first place. Mao et al[32] faced this issue when trying to predict driver mutations in cancer, as these happen in a low proportion compared with passenger mutations. To tackle this problem, the authors used a weighted version of SVMs.

A different approach for the class imbalance problem is to use resampling strategies to produce balance.[58] Undersampling, for example, consists in extracting a smaller random sample from the majority class, such that this one is used instead of the original majority class. Another resampling strategy is oversampling, which means adding instances from the lesser class through duplication or slightly altered copies. Although these strategies should be repeated to assess the stability of the results, each time that they are used, they risk losing information (when undersampling) or overfitting the data (when oversampling).

4.8 | Sampling

Despite the efforts of gathering information for GB and subsequent analyses, sometimes results can be misunderstood in light of the study design used to obtain the sample subjects. In the field of health sciences, dealing with convenience samples or independent databases becomes a relatively regular practice considering the limitations encountered by clinicians (ie, accruing patients of low-prevalence diseases), which makes them prioritize availability and accessibility to patients over sample representativeness of the population. In brief, it is important to acknowledge the limitations of each of these databases when reporting our results.

Using nonprobabilistic samples can lead to biased and nongeneralizable results, so it quickly becomes imperative that researchers clearly state what their target population, study population, and sampling method are before concluding or making inferences from their results. A good start could be a clear description of the study design and sampling process that will generate the input data of ML techniques, including the strong and weak areas of these.

Another reason to consider an initial and appropriate sampling procedure is to assure independence between observations in our data; that is, every individual contributes independent pieces of information to the learning process. Dependence among individuals can happen if, for instance, we ignore the hierarchy of our data (eg, measures of subjects in different clusters or entities). Then, if our sample is indeed highly correlated, we can expect our training and testing sets to be highly correlated too. This redundancy in information can jeopardize the generality of the results because of the training and testing sets not being representative of the target population. For instance, we realized that one of the cited manuscripts[31] did not fully describe nor consider the samples' hierarchical aspect when running the analyses, having more than one biological sample per subject for some of the individuals, and thus the risk of data redundancy because of pseudoreplication could be a potential problem.[31]

4.9 | Additional ML techniques

The reader must be aware that only a small sample of the available spectrum of ML methods have been described here. Moreover, most of these techniques can be modified to change the type of learning problem or analysis. Therefore, there is a vast array of options of statistical models waiting to be applied in the field of brain cancer research. For instance, random forests can be modified to predict quantitative outcomes that are subject to censoring data (eg, OS), resulting in a supervised learning technique known as random survival forests that allow learning about how features are associated with patients' survival time.[56]

As for unsupervised learning, there are several extensions of PCA and a myriad of clustering algorithms. An example of an extension of PCA is functional PCA (fPCA),[59] which is applied to functional data (eg, spectra and time series) to find PC functions that can be used to describe the data better. Like in PCA, results from fPCA can be used for exploratory data analysis or as a preprocessing step with an output that is used as input in a supervised learning method. For example, if fPCA were to be applied to MRS metabolite data from Metsis et al,[31] we would be interested in checking whether different brain tumor types show different (averaged) values of the PC functions (ie, different patterns).

As mentioned before, there are many clustering algorithms, most of which can be classified as either hierarchical or partitioning. Partitioning methods, unlike HC, search for a predefined number of groups or partitions in the data set. One of the most popular algorithms for doing this is the *K*-means algorithm, which will assign each observation to one of the predefined numbers (*K*) of groups. For this, *K*-means iterates between two steps: firstly, the means of the currently assigned clusters are recomputed, so the total cluster variance is minimized; secondly, observations are reassigned to their currently closest cluster mean so that the total cluster variance is minimized. The algorithm is stopped once cluster assignments stabilize or do not change.

To assess the validity of its results, the stability of the output of the *K*-means algorithm is checked for several different starting

random cluster assignments (random seeds), as well as for a varying number of *K* (within a sensible range of values). An example of the application of *K*-means can be found in the work of Sturm et al,[60] where the authors used this technique to discover clusters of GB samples based on DNA methylation, finding six distinct biological subgroups that were characterized using further analyses.

4.10 | Input variables justification

Considering the large quantity of data that ML models require, scientists interested in predicting prognosis or understanding associations explaining the biology and tumor behavior may think that using all the available information can help them to comply with the data size requirement. Nevertheless, it is pertinent to always explain the inclusion of every variable in the input of any ML technique, as this exercise can save time, resources, and lead to better performance.

As an example of this last point, Wang and Liu[61] recently used a weighted version for random survival forest, where they showed how prediction accuracy could be improved by skewing the model to choose highly connected genes on gene expression data. By using these topologically relevant genes and signatures, they were able to obtain robust prognostic values with high biological relevance to the development of GB and esophageal squamous cell carcinoma.[61]

The answers to questions like "Can this feature be measured in a clinical setup?" or "Is it plausible that this feature is associated with the biological process?" can help to justify the inclusion or exclusion of variables. Two additional reasons for justifying input variables are to prevent turning ML techniques into data dredging tools and the fact that ML is not immune to the principle of "garbage in, garbage out."

5 | CONCLUDING REMARKS

So far, by using omics, researchers have been able to get closer to the promise of precision medicine,[14,62-64] especially now that ML is becoming more and more popular and well-established in the clinical arena. Moreover, ML consisting of flexible techniques that readily allow the use of independent datasets[32,41] or combined information from different technologies as input[31] makes us acknowledge and address dependency limitations. To help researchers keep in mind what we have discussed throughout this review, we have provided a flowchart with general guidelines for using ML techniques in cancer research (Figure 9).

Our manuscript focuses on GB because of its highly lethal outcome and complex heterogeneity as a model for other cancers, which we believe needs to be addressed through modern statistical ML methods. Using such methods will allow us to reach the goal of learning about GB oncogenesis, elucidating the relationships between targeted molecular structures, like TNTs and tumor development. Finally, new developments in the ML arena are necessary to identify therapies or drugs to eradicate GB.

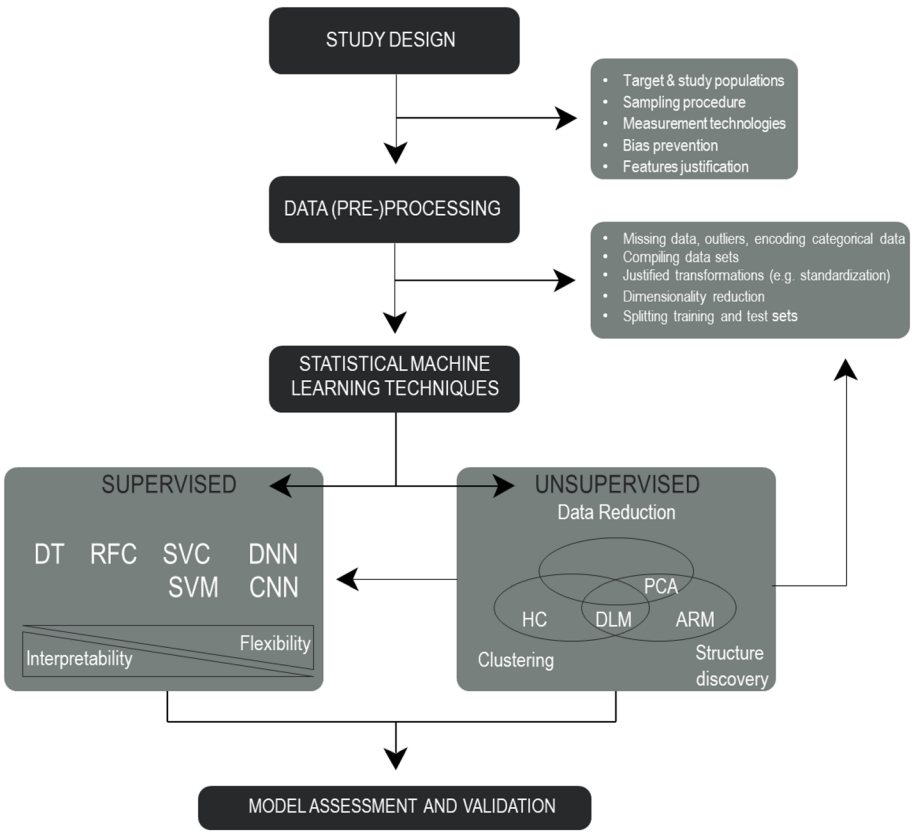

FIGURE 9 A simplified roadmap to the use of machine learning (ML) techniques in cancer research. Major stages are shown in dark-gray boxes, while further details and summaries are shown in light-gray boxes. Supervised ML models differ on their flexibility to fit the data and the level of interpretability that their results exhibit. Unsupervised ML methods can be used for: data reduction, clustering, and structure discovery. Only the ML techniques reviewed in this work are included in this figure

AUTHORS' CONTRIBUTIONS

All authors had full access to the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. Conceptualization, J.V., F.M.; Methodology, F.M.; Investigation, J.V., F.M.; Formal Analysis, J.V., F.M.; Resources, J.V., F.M.; Writing - Original Draft, J.V., F.M; Writing - Reviewand Editing, J.V., F.M.; Visualization, J.V., F.M; Supervision, F.M; Funding Acquisition, F.M.

ACKNOWLEDGMENTS

F.M. thankfully acknowledges funding from Comisión Nacional de Investigación Científica y Tecnológica, CONICYT Ph.D. fellowship 21151523. The authors want to thank Andrés Iturriaga, Ph.D., Professor at the School of Public Health of Universidad de Chile, for his valuable suggestions to this review, and Francisca Kong for technical assistance.

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

ORCID

Jessica Valdebenito https://orcid.org/0000-0002-6854-0463
Felipe Medina https://orcid.org/0000-0003-4486-9237

REFERENCES

1. Thakkar JP, Dolecek TA, Horbinski C, Ostrom QT, Lightner DD, Barnholtz-Sloan JS, et al. Epidemiologic and molecular prognostic review of glioblastoma. *Cancer Epidemiol Biomarkers Prev*. 2014;23(10):1985–1996. https://www.ncbi.nlm.nih.gov/pubmed/25053711
2. Stupp R, Mason WP, Van Den Bent MJ, Weller M, Fisher B, Taphoorn MJ, Belanger K, Brandes AA, Marosi C, Bogdahn U, Curschmann J. Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. *N Engl J Med*. 2005;10(352):987–996. 00
3. Davis ME. Glioblastoma: Overview of Disease and Treatment. *Clin J Oncol Nurs*. 2016;20(5):1-14. 0
4. Nam JY, de Groot JF. Treatment of Glioblastoma. *J Oncol Pract*. 2017;13(10):629–638. https://www.ncbi.nlm.nih.gov/pubmed/29020535
5. Hanif F, Muzaffar K, Perveen K, Malhi SM, Simjee SU. Glioblastoma Multiforme: A Review of its Epidemiology and Pathogenesis through Clinical Presentation and Treatment. *Asian Pacific J Cancer Prev*. 2017;18:1-9. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5563115/
6. Urbańska K, Sokołowska J, Szmidt M, Sysa P. Glioblastoma multiforme - An overview. *Contem Oncol (Pozn)*. 2014; 18(5):307–312. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4248049/
7. Ghosh D, Nandi S, Bhattacharjee S. Combination therapy to checkmate Glioblastoma: clinical challenges and advances. *Clin Transl Med*. 2018;7:33. https://doi.org/10.1186/s40169-018-0211-8
8. Micheel C, Nass S, Omenn G. *Evolution of Translational Omics: Lessons Learned and the Path Forward*. Washington, DC: The National Academies Press; 2012;354 https://www.ncbi.nlm.nih.gov/pubmed/24872966
9. Colman H, Zhang L, Sulman EP, McDonald JM, Shooshtari NL, Rivera A, et al. A multigene predictor of outcome in glioblastoma. *Neuro Oncol*. 2010;12(1):49–57. https://www.ncbi.nlm.nih.gov/pubmed/20150367
10. Sham PC, Purcell SM. Statistical power and significance testing in large-scale genetic studies. *Nat Rev Genet*. 2014;15(5):335-346. https://doi.org/10.1038/nrg3706
11. Osswald M, von Deimling A, Weil S, et al. Brain tumour cells interconnect to a functional and resistant network. *Nature*. 2015;528(7580):93-98. https://www.nature.com/articles/nature16071
12. Lathia JD, Mack SC, Mulkearns-Hubert EE, Valentim CLL, Rich JN. Cancer stem cells in glioblastoma. *Genes Dev*. 2015;29:1203-1217. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4495393/
13. Bhat KPL, Salazar KL, Balasubramaniyan V, Wani K, Heathcock L, Hollingsworth F, et al. The transcriptional coactivator TAZ regulates mesenchymal differentiation in malignant glioma. *Genes Dev*. 2011;25(24):2594–2609. https://www.ncbi.nlm.nih.gov/pubmed/22190458
14. Cao H, Wang F, Li XJ. Future Strategies on Glioma Research: From Big Data to the Clinic. *Genomics, Proteomics Bioinforma*. 2017;15(4):263–265. https://doi.org/10.1016/j.gpb.2017.07.001
15. Bzdok D, Krzywinski M, Altman N. Machine learning: a primer. *Nat Methods*. 2017;14(12):1119–1120. https://doi.org/10.1038/nmeth.4526
16. Heaton J. *Artificial Intelligence for Humans, Volumen 3: Deep Learning and Neural Networks*. Heaton T, editor. Chesterfield, MO: Heaton Research, Inc; 2015. 268 p. https://www.heatonresearch.com/book/
17. James G, Witten D, Hastie T, Tibshirani R. *An Introduction to Statistical Learning with applications in R. Casella G, Fienberg S, Olkin I, editors. Springer*. New York: Springer US; 2017. 426 p. https://www.springer.com/gp/book/9781461471370
18. Hastie T, Tibshirani R, Friedman J. *The elements of statistical learning. Data Mining, Inference, and Prediction. Second ed*. New York, NY: Springer US; 2017. 764 p. https://web.stanford.edu/~hastie/ElemStatLearn/
19. Beam AL, Kohane IS. Big Data and Machine Learning in Health Care. *J Am Med Assoc*. 2018;319(13):1317–1318. https://www.ncbi.nlm.nih.gov/pubmed/29532063
20. Gligorijević V, Malod-Dognin N, Pržulj N. Integrative Methods for Analyzing Big Data in Precision Medicine. *Proteomics*. 2016;16(5):741-758.
21. Royston P, Moons KGM, Altman DG, Vergouwe Y. Prognosis and prognostic research: Developing a prognostic model. *BMJ*. 2009;16(5):1373–1377. https://doi.org/10.1002/pmic.201500396
22. Bzdok D, Altman N, Krzywinski M. Statistics versus machine learning. *Nat Methods*. 2018;15(4):233–234. https://doi.org/10.1038/nmeth.4642
23. Burton EC, Lamborn KR, Feuerstein BG, Prados M, Scott J, Forsyth P, et al. Genetic Aberrations Defined by Comparative Genomic Hybridization Distinguish Long-Term from Typical Survivors of Glioblastoma. *Cancer Res*. 2002;62(21):6205–6210. https://cancerres.aacrjournals.org/content/62/21/6205
24. Trépant AL, Bouchart C, Rorive S, Sauvage S, Decaestecker C, Demetter P, et al. Identification of OLIG2 as the most specific glioblastoma stem cell marker starting from comparative analysis of data from similar DNA chip microarray platforms. *Tumor Biol*. 2015;36(3):1943–1953. https://www.ncbi.nlm.nih.gov/pubmed/25384509
25. Brennan CW, Verhaak RGW, McKenna A, Campos B, Noushmehr H, Salama SR, et al. The Somatic Genomic Landscape of Glioblastoma. *Cell*. 2013;155(2):462–477. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3910500/
26. Tanwar MK, Gilbert MR, Holland EC. Gene Expression Microarray Analysis Reveals YKL-40 to Be a Potential Serum Marker for Malignant Character in Human Glioma. *Cancer Res*. 2002;62(15):4364–4368. https://www.ncbi.nlm.nih.gov/pubmed/12154041
27. Verhaak RGW, Hoadley KA, Purdom E, Wang V, Qi Y, Wilkerson MD, et al. An Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in

PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell*. 2010;17(1):98–110. https://www.ncbi.nlm.nih.gov/pubmed/20129251
28. Phillips HS, Kharbanda S, Chen R, Forrest WF, Soriano RH, Wu TD, et al. Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis. *Cancer Cell*. 2006;9(3):157–173. https://www.ncbi.nlm.nih.gov/pubmed/16530701
29. Macyszyn L, Akbari H, Pisapia JM, Da X, Attiah M, Pigrish V, et al. Imaging patterns predict patient survival and molecular subtype in glioblastoma via machine learning techniques. *Neuro Oncol*. 2016;18(3):417–425. https://www.ncbi.nlm.nih.gov/pubmed/26188015
30. Korfiatis P, Kline TL, Coufalova L, et al. MRI texture features as biomarkers to predict MGMT methylation status in glioblastomas. *Med Phys*. 2016;43(6):2835-2844. https://doi.org/10.1118/1.4948668
31. Metsis V, Huang H, Andronesi OC, Makedon F, Tzika A. Heterogeneous data fusion for brain tumor classification. *Oncol Rep*. 2012;28(4):1413–1416. https://www.ncbi.nlm.nih.gov/pubmed/22842996
32. Mao Y, Chen H, Liang H, Meric-Bernstam F, Mills GB, Chen K. CanDrA: Cancer-specific driver missense mutation annotation with optimized features. *PLoS One*. 2013;8(10):e77945. https://www.ncbi.nlm.nih.gov/pubmed/24205039
33. Krzywinski M, Altman N. Classification and regression trees. *Nat Methods*. 2017;14(8):757–759. nature.com/articles/nmeth.4370
34. Kingsford C, Salzberg SL. What are decision trees? *Nat Biotechnol*. 2008;26(9):1011-1012. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2701298/
35. Gollapalli, Kishore, Ray, Sandipan, Srivastava R et al. Investigation of serum proteome alterations in human glioblastoma multiforme. *Proteomics*. 2012;12:2378–2390. https://www.ncbi.nlm.nih.gov/pubmed/22684992
36. Kickingereder P, Bonekamp D, Nowosielski M, Kratz A, Sill M, Burth S, et al. Radiogenomics of glioblastoma: Machine Learning–based Classification of Molecular Characteristics by Using Multiparametric and Multiregional MR Imaging Features. *Radiology*. 2016;281(3):907–918. https://pubs.rsna.org/doi/10.1148/radiol.2016161382
37. Chang K, Zhang B, Guo X, Zong M, Rahman R, Sanchez D, et al. Multimodal imaging patterns predict survival in recurrent glioblastoma patients treated with bevacizumab. *Neuro Oncol*. 2016;18(12):1680–1687. https://www.ncbi.nlm.nih.gov/pubmed/27257279
38. Tustison NJ, Shrinidhi KL, Wintermark M, Durst CR, Kandel BM, Gee JC, et al. Optimal Symmetric Multimodal Templates and Concatenated Random Forests for Supervised Brain Tumor Segmentation (Simplified) with ANTsR. *Neuroinformatics*. 2015;13(2):209–225. https://www.ncbi.nlm.nih.gov/pubmed/25433513
39. Glinsky G V., Higashiyama T, Glinskii AB. Classification of Human Breast Cancer Using Gene Expression Profiling as a Component of the Survival Predictor Algorithm. *Clin Cancer Res*. 2004;10(7):2272–2283. https://www.ncbi.nlm.nih.gov/pubmed/15073102
40. Goodfellow I, Bengio Y, Courville A. Deep Learning. MIT Press; 2016. http://www.deeplearningbook.org
41. Hussain S, Anwar SM, Majid M. Segmentation of glioma tumors in brain using deep convolutional neural network. *Neurocomputing*. 2017;282:248–261. https://doi.org/10.1016/j.neucom.2017.12.032
42. Mohsen H, El-Dahshan E-SA, El-Horbaty E-SM, Salem A-BM. Classification using Deep Learning Neural Networks for Brain Tumors. *Futur Comput Informatics J*. 2017;3(1):68-71. http://linkinghub.elsevier.com/retrieve/pii/S2314728817300636
43. Menze BH, Jakab A, Bauer S, Kalpathy-Cramer J, Farahani K, Kirby J, et al. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Trans Med Imaging*. 2015;34(10):1993–2024. https://www.ncbi.nlm.nih.gov/pubmed/25494501
44. Tseng KL, Lin YL, Hsu W, Huang CY. Joint sequence learning and cross-modality convolution for 3D biomedical segmentation. In: 30th IEEE Conference on Computer Vision and Pattern Recognition. 2017. p. 8. https://arxiv.org/abs/1704.07754
45. Lever J, Krzywinski M, Altman N. Principal component analysis. *Nat Methods*. 2017;14(7):641-642. https://doi.org/10.1038/nmeth.4346
46. Akbari H, Macyszyn L, Da X, Bilello M, Wolf RL, Martinez-Lage M, et al. Imaging surrogates of infiltration obtained via multiparametric imaging pattern analysis predict subsequent location of recurrence of glioblastoma. *Neurosurgery*. 2016;78(4):572–580. https://www.ncbi.nlm.nih.gov/pubmed/26813856
47. Madhavan S, Zenklusen J-C, Kotliarov Y, Sahni H, Fine HA, Buetow K. Rembrandt: Helping Personalized Medicine Become a Reality through Integrative Translational Research. *Mol Cancer Res*. 2009;7(2):157–167. http://mcr.aacrjournals.org/cgi/doi/10.1158/1541-7786.MCR-08-0435
48. Altman N, Krzywinski M. Clustering. *Nat Methods*. 2017;14(6):545-546. https://www.nature.com/articles/nmeth.4299
49. Bredel M, Scholtens DM, Harsh GR, et al. A network model of a cooperative genetic landscape in brain tumors. *J Am Med Assoc*. 2009;302(3):261-275. https://jamanetwork.com/journals/jama/fullarticle/184262
50. Cremaschi P, Carriero R, Astrologo S, Colì C, Lisa A, Parolo S, et al. An Association Rule Mining Approach to Discover lncRNAs Expression Patterns in Cancer Datasets. *Biomed Res Int*. 2015; Article ID 146250. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4530207/
51. Young JD, Cai C, Lu X. Unsupervised deep learning reveals prognostically relevant subtypes of glioblastoma. *BMC Bioinformatics*. 2017;18(Suppl 11):381. https://www.ncbi.nlm.nih.gov/pubmed/28984190
52. Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. *Lancet*. 2019;393(10181):1577-1579. https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(19)30037-6/fulltext
53. Shmueli G. To Explain or To Predict? *Stat Sci*. 2010;25(3):289-310. https://www.stat.berkeley.edu/~aldous/157/Papers/shmueli.pdf
54. Lever J, Krzywinski M, Altman N. Classification evaluation. *Nat Methods*. 2016;13(10):890-890. https://www.nature.com/articles/nmeth.3945
55. Lever J, Krzywinski M, Altman N. Model selection and overfitting. *Nat Methods*. 2016;13(9):703-704. https://doi.org/10.1038/nmeth.3968
56. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. *Ann Appl Stat*. 2008;2(3):841-860. https://arxiv.org/pdf/0811.1645.pdf
57. Baghfalaki T, Ganjali M, Berridge D. Missing Value Imputation for RNA-Sequencing Data Using Statistical Models: A Comparative Study. *J Stat Theory Appl*. 2016;15(3):221–236. https://www.atlantis-press.com/journals/jsta/25862105
58. Rahman MM, Davis DN. Addressing the Class Imbalance Problem in Medical Datasets. *Int J Mach Learn Comput*. 2013;3(2):224–228. https://doi.org/10.7763/IJMLC.2013.V3.307
59. Viviani R, Grön G, Spitzer M. Functional principal component analysis of fMRI data. *Hum Brain Mapp*. 2005;24(2):109–129. https://doi.org/10.1002/hbm.20074
60. Sturm D, Witt H, Hovestadt V, Khuong-Quang DA, Jones DTW, Konermann C, et al. Hotspot Mutations in H3F3A and IDH1 Define Distinct Epigenetic and Biological Subgroups of Glioblastoma. *Cancer Cell*. 2012;22(4):425–437. https://www.ncbi.nlm.nih.gov/pubmed/23079654
61. Wang W, Liu W. Integration of gene interaction information into a reweighted random survival forest approach for accurate survival prediction and survival biomarker discovery. *Sci Rep*. 2018;8(1):1-14. https://doi.org/10.1038/s41598-018-31497-0
62. Rudie JD, Rauschecker AM, Bryan RN, Davatzikos C, Mohan S. Emerging Applications of Artificial Intelligence in Neuro-Oncology. *Radiology*. 2019;290(3):607-618. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6389268/

63. Tandel GS, Biswas M, Kakde OG, et al. A review on a deep learning perspective in brain cancer classification. *Cancers (Basel)*. 2019; 11: 111. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6356431/
64. Shah P, Kendall F, Khozin S, et al. Artificial intelligence and machine learning in clinical development: a translational perspective. *npj Digit Med*. 2019; 2:69. https://doi.org/10.1038/s41746-019-0148-3

How to cite this article: Valdebenito J, Medina F. Machine learning approaches to study glioblastoma: A review of the last decade of applications. *Cancer Reports*. 2019;2:e1226. https://doi.org/10.1002/cnr2.1226