San Jose State University
SJSU ScholarWorks

Faculty Research, Scholarly, and Creative Activity

11-1-2023

An interactive web-based tool for predicting and exploring brain cancer survivability

Gopal Nath
Murray State University Murray

Austin Coursey
Vanderbilt University

Yang Li
Montclair State University

Srikanth Prabhu
Manipal Institute of Technology

Harish Garg
Thapar Institute of Engineering & Technology

See next page for additional authors

Follow this and additional works at: https://scholarworks.sjsu.edu/faculty_rsca

Recommended Citation

Gopal Nath, Austin Coursey, Yang Li, Srikanth Prabhu, Harish Garg, Shaymal C. Halder, and Saptarshi Sengupta. "An interactive web-based tool for predicting and exploring brain cancer survivability" *Healthcare Analytics* (2023). https://doi.org/10.1016/j.health.2022.100132

This Article is brought to you for free and open access by SJSU ScholarWorks. It has been accepted for inclusion in Faculty Research, Scholarly, and Creative Activity by an authorized administrator of SJSU ScholarWorks. For more information, please contact scholarworks@sjsu.edu.

Authors

Gopal Nath, Austin Coursey, Yang Li, Srikanth Prabhu, Harish Garg, Shaymal C. Halder, and Saptarshi Sengupta

This article is available at SJSU ScholarWorks: https://scholarworks.sjsu.edu/faculty_rsca/2195

Healthcare Analytics 3 (2023) 100132

Contents lists available at ScienceDirect

Healthcare Analytics

journal homepage: www.elsevier.com/locate/health

An interactive web-based tool for predicting and exploring brain cancer survivability

Gopal Nath [a,*], Austin Coursey [b], Yang Li [c], Srikanth Prabhu [d], Harish Garg [e,f,g], Shaymal C. Halder [h], Saptarshi Sengupta [i]

[a] Murray State University, Department of Mathematics & Statistics, Murray, KY 42071, USA
[b] Vanderbilt University, Department of Computer Science, Nashville, TN 37235, USA
[c] Montclair State University, Feliciano School of Business, Montclair, NJ 07043, USA
[d] Manipal Institute of Technology, Department of Computer Science and Engineering, Manipal, India
[e] School of Mathematics, Thapar Institute of Engineering & Technology (Deemed University), Patiala 147004, Punjab, India
[f] Department of Mathematics, Graphic Era Deemed to be University, Dehradun, 248002, Uttarakhand, India
[g] Applied Science Research Center, Applied Science Private University, Amman, 11931, Jordan
[h] Grand Valley State University, The Department of Statistics, Allendale, MI 49401, USA
[i] San Jose State University, Department of Computer Science, San Jose, CA 95192, USA

ARTICLE INFO

Keywords:
machine learning
data mining
brain cancer
Cook's distance
cross-validation
predictive analytics

ABSTRACT

Brain cancer is one of the most deadly cancers, with a very low survival rate. By understanding the factors that lead to cancer spreading, practitioners can concentrate their efforts on providing the most effective treatment, and they can modify the treatment plan as necessary. Also, knowing the likelihood of a patient's survival over a specified time period can enable them to make informed decisions about adjusting their routines, future investments, and other health-related decisions. The use of data-driven models in cancer research has gained increased popularity over the past several decades. Moreover, there is still much uncertainty surrounding the factors that contribute to survival of cancer, making it difficult to develop a model. The existing literature on brain cancer contains a variety of machine learning models. However, many of them lack a high degree of accuracy, and, in medical research, accuracy is essential to the proper guidance of treatment decisions. Therefore, we have proposed a framework comprising multiple phases of classical statistics and machine learning methods to find a parsimonious model with a high degree of accuracy (98.9%) for predicting brain cancer survivability. Furthermore, we develop a prototype web-based interactive tool to facilitate the practical implementation of the proposed model and provide a deeper understanding of how a particular factor affects survival when other factors remain unchanged. By integrating this tool into healthcare settings, medical professionals can rapidly detect potentially vulnerable patients, and it can also be useful in determining the most effective treatment plan.

1. Introduction

Brain cancers, a term used to describe various forms of cancerous tumors growing in the brain and spinal cord, are often fatal due to their invading nature and the tendency to resist traditional surgical treatments [1]. Brain cancers are one of the most lethal cancers and are responsible for significant morbidity and mortality in the United States [2]. According to the Surveillance, Epidemiology, and End Results Program, the annual incidence of brain cancers is 6.3 cases per 100,000 people. There will be 25,050 new cases and 18,280 deaths from brain cancers in 2022 [3]. Despite the rarity of brain cancers, they have a significant economic and social impact on individuals, families, and the community [4]. Furthermore, brain cancers have a great impact on the healthcare system because they have an inherently disabling effect on patients by preventing them from being able to function independently [5]. There have been few significant advances in prevention, early detection, and treatment of such illnesses over the past four decades. There was only a marginal improvement in the 5-year survival rate of glioblastoma patients, a major brain cancer, from 4 to 7% [2]. An accurate understanding of cancer survival after the first diagnosis is essential for both doctors and patients. This enables the medical practitioner to make better treatment planning decisions [6]. Additionally, the patient may be able to consider possible lifestyle changes, determine treatment options and make important financial decisions [7]. A wide range of techniques has become increasingly

* Corresponding author.
E-mail address: gnath@murraystate.edu (G. Nath).

https://doi.org/10.1016/j.health.2022.100132
Received 6 July 2022; Received in revised form 17 December 2022; Accepted 17 December 2022

2772-4425/© 2022 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

common in recent years to predict cancer survival after the first diagnosis [7–24]. Researchers have pioneered machine learning algorithms to predict the survival rates of patients with brain cancer using magnetic resonance imaging (MR) [8–12]. Some research focuses on identifying essential features, primarily from groups of genes, that can assist in discovering the causes of cancer [13–16,25–29]. However, in recent years, researchers have focused their attention on identifying demographic and clinical features associated with cancer and predicting survival using various supervised models [7,24,30–39]. To identify the limitations of the existing literature, we summarize these studies in the following subsections.

1.1. Research based on image processing

The analysis of medical images can help physicians reduce their workload and identify the type and class of brain tumors. In conjunction with machine learning algorithms, detecting the tumor region in Magnetic Resonance(MR) brain images can furnish physicians with detailed information about the location and size of the tumor so that appropriate treatment can be administered. Clustering algorithms are among the most widely applied algorithms in the field of image segmentation. For example, Srinivas et al. used a Fuzzy C-means clustering algorithm along with K-means clustering for segmentation of the MR image of the brain [40]. Vithanuvarthanan et al. employed a hybrid self-organized map along with the Fuzzy k-means algorithms to accurately identify tumors and segment brain tissue regions [41]. Several other image segmentation algorithms are designed through the use of Artificial Neural Networks (ANN) [42–45]. The models are vital for identifying cancers and hence, may facilitate the development of more effective treatment methods. However, none of the above models were used to predict cancer survival, which is vital information for patients and their families so that they can adjust their living styles and make informed financial decisions.

However, authors have pioneered this field recently by predicting the survival rates of patients with brain cancer based on MRI images using machine learning algorithms [8–12]. Chato et al. for example, utilized the BraTS2017 dataset, which consists of 163 samples with four sequences of MRI brain images, the overall survival period in days, and the age of the patients [8]. The authors applied multiple machine learning algorithms and found that they achieved the best accuracy using pre-trained AlexNet and trained by Linear Discriminants. Due to the noise present in the MRI images, histogram features attained an accuracy of 68.5%. Another research study aims to evaluate the survival rate for glioma (one of the major brain tumors) using different MRI techniques. A decision tree and cross-validation techniques were used to calculate survival estimations based on 4524 radiomic features derived from segmented tumor regions. Three different CNN architectures were used to improve the performance of the network in tumor segmentation. This study achieved an accuracy of 61.0% in identifying short-, mid-, and long-term survivors [12]. In all of the studies above, the accuracy scores were found to be very low (maximum accuracy of 68.5%). Additionally, according to Kickingereder et al. the relationships between MR images and underlying tumor characteristics did not appear strong enough for the generation of reliable and clinically meaningful classification models through machine learning [46]. Therefore, the above MRI-based models may not be reliable in any real healthcare setting.

1.2. Research based on gene expression

Comparing genes expressed in normal tissue and diseased tissue may provide physicians with better insights into the pathology of cancer and assist in making decisions. In order to assess the molecular and physiological mechanisms of disease, it is crucial to examine gene expression patterns for characteristics associated with clinical behavior, thereby providing alternative approaches to understanding the molecular and physiological mechanisms of disease [25–27]. A genome-scale gene expression profile is useful for the identification of intertumor heterogeneity and homogeneity [28,29]. Recent studies have been conducted to identify glioma subtypes associated with certain molecular characteristics [13–16]. As genetic data is better fitted for unsupervised methods like clustering [7], the majority of studies [47,48] in this field have utilized data sets containing various genetic profiles associated with specific cancers. An unsupervised algorithm was used in a recent study by Li et al. to predict the prognostic features and prognostic groups for gliomas [48]. The authors employed two unsupervised machine learning approaches to look at genome-wide gene expression profiles for 159 different gliomas, ultimately resulting in a model for glioma classification that relied exclusively on molecular data. Information provided by the above study is essential for identifying cancer subtypes and was not intended to predict cancer survival. However, these models, despite their value for analyzing data, may not be suitable to predict survival or make therapeutic decisions as genetic factors are unable to provide information on whether cancer has spread or if a specific treatment is required [7].

1.3. Research based on clinical data

Numerous models have been developed to address the above concerns based on commonly available clinical features such as grade, age, cancer size, node size, laterality, surgery, etc. Existing studies have used statistical and machine learning algorithms to predict the survival of cancer patients based on clinical data. The following subsections provide a brief overview of the literature review based on the clinical data.

1.3.1. Traditional statistical analysis

For survival analysis, Cox proportional hazards models and Kaplan–Meier methods are among the most frequently used statistical methods for clinical data.

Cox proportional hazards models: Existing studies that use clinical data tend to use Cox proportional hazards regression models to determine how covariates affect disease-specific survival [17–22]. For example, Rosenberg et al. explored 72,367 breast cancer patients aged between 21–91 years from the Surveillance, Epidemiology, and End Results (SEER) database from 1973 to 1998 and analyzed the effect of patient and tumor characteristics on survival using a proportional hazards model [18]. The authors found that larger tumor sizes and higher tumor grades negatively correlated with survival, and the year of diagnosis was positively associated with survival. Patients' age and stage of disease violated the proportional hazards assumption since distant diseases had a much lower short-term survival rate than one would predict from a proportional hazards model. In spite of the fact that the Cox proportional hazard regression model can be utilized to extract valuable information from clinical data, there are few studies that have verified the validity of the proportional hazards assumptions. It has been noted that the Cox proportional hazard regression model critically relies on the assumption that hazards between comparison groups are in a constant proportion over time [18]. Additionally, Cox proportional hazard regression models assume that each predictor variable is a linear factor, ignoring any nonlinear effects on outcome variables. Since tumor development and changes depend on a variety of factors, it is unlikely that models can accurately predict patient outcomes for cancer patients [49].

Kaplan–Meier method: Further, the Kaplan–Meier method [20,50–52] is used to estimate overall survival (OS) over a fixed period of time, which is a very common way of analyzing clinical data. Using the SEER database, Fang et al. obtained data on bladder cancer cases between 2010 and 2015 [50]. The Kaplan–Meier survival analysis and nomogram analyses were used to visually and effectively predict the 3- and 5-year overall survival of patients with bladder cancer. The

authors found that based on all Kaplan–Meier survival curves, there are no significant differences in the survival rates of blacks and American Indians/Native Alaskans, but these two races have lower survival rates than whites and Asian and Pacific Islanders. In a recent study, Liu et al. analyzed data on brainstem gliomas diagnosed between 2004 and 2016 from the SEER database that comprised 3387 cases of brainstem gliomas [51]. Using Kaplan–Meier curves and the Cox proportional hazards model, the author concluded that patients with tumors of less than 3 cm in diameter had a better chance of survival, and surgery was effective in improving overall survival rates. Radiotherapy and chemotherapy did not appear to improve overall survival. Despite the fact that traditional statistical methods are better suited for investigating the relationship between covariates and end-point events, some factors were not included in the model due to a lack of statistical significance [53]. While the Kaplan–Meier method is the most popular method for survival analysis, it only examines the effects of one factor at a time and is therefore unsuitable for multivariate analysis [54].

1.3.2. Machine learning algorithms

For traditional statistical techniques, we have seen that there are many assumptions that must be followed in order to apply the method. However, machine learning techniques typically rely on fewer assumptions and provide superior and more reliable results [30]. Existing studies have increasingly utilized clinical data to develop supervised machine learning algorithms, for example, artificial neural networks (ANNs) [7,31–33], support vector machines (SVMs) [33,36,55,56], logistic regression (LR) [7,33,36,55,56], K-nearest neighbors (KNNs) [55], decision tree (DT) [33,36], etc., for cancer survival prediction. The researchers employed a variety of machine learning algorithms to predict the 1-year [7,34–36], 2-year [36], 5-year [7,24,37,38,57–59], 10-year [7,23,24] and 15-year [24,60] survival of different cancer types based on clinical data.

While there has been considerable research on machine learning, relatively few articles have discussed machine learning algorithms for modeling and validating brain cancer based on clinical data. A majority of the articles analyzed traditional statistical analysis and MRI images to predict patient survival or compare survival rates among brain cancer patients; however, these models have some limitations (as discussed in previous subsections). Samara et al. were amongst the very few authors to use the SEER dataset for glioblastoma, one of the most common forms of brain cancer, and to develop a prognostication system based on ensemble learning to predict short-, intermediate-, and long-term survival [36]. The authors of this study conclude that *age, CS tumor size, county, month of diagnosis, RX Sum, primary site, and laterality* are the most significant features for glioblastoma cancer survival. Even though the top model achieved an AUC score of 94%, the authors used some features, such as the *insurance Recode*, which may not be useful to practitioners exploring factors directly linked to cancer prognosis. In an earlier study, Senders et al. designed an online calculator to assess survival in patients with glioblastoma by combining classical statistics and machine learning algorithms [61]. The authors included *insurance* as one of the features; additionally, the authors did not report performance metrics as part of the model validation process. Thus, the above models may not be helpful to practitioners in exploring factors directly associated with brain cancer and predicting patient survivability.

1.4. Contribution of our study

As discussed in the previous sections, the prevailing literature focusing on predicting the survival of brain cancer patients has limitations. In light of these deficiencies, the present study has adopted clinical data and developed a machine learning model with a high degree of accuracy for brain cancer survivability. Therefore, to find the most parsimonious models for brain cancer survival, we have taken the following steps.

Using statistical methods to eliminate inconsistent observations: Based on the discussion in the preceding sections, we have chosen to use clinical data (SEER) and machine learning for this study as they offer a number of advantages. Since medical data is collected from various sources, such as images, interviews, physician notes, etc. [39], it is possible to have inconsistent information due to errors in data entry. The authors of studies that use SEER data derived from the National Institutes of Health (NIH) tend to remove inconsistent or redundant observations manually or with expert knowledge without performing statistical analysis [38,62–66]. Delen et al. for example, removed patients with tumors that measured greater than 200 mm for predicting breast cancer survival [38]. The authors of a recent study by Gupta et al. eliminated outliers or inconsistent records without performing any statistical analysis and then utilized restricted Boltzmann machines, deep autoencoders, and convolutional neural networks (CNNs) to analyze the postoperative survival of breast cancer patients [63]. We have, therefore, identified outliers or inconsistent observations and removed them during preprocessing of the data, but not manually; we instead employed a statistical technique proposed by Cook [67] in 1977. According to our knowledge, the technique has never been applied in this field. There are many researchers throughout the world who are studying Surveillance, Epidemiology, and End Results(SEER) data resources extensively. In search of PubMed (www.ncbi.nlm.nih.gov/pubmed) using the keywords Surveillance, Epidemiology, and End Results between 1973 and 2015, 40,031 citations were found, which demonstrates the research productivity generated by this program [68]. Therefore, our proposed preprocessing statistical method can be implemented in other types of cancer for SEER data.

Features selection, balancing techniques, cross-validation and classification models: By employing the Least Absolute Shrinkage and Selection Operator (LASSO) [69] an embedded technique [70], we extract the features that contribute significantly to the model's predictive power, subsequently yielding highly parsimonious models. A significant difference in survival classes results in imbalance problems in most survival datasets [7]. Thus, to solve the imbalance problem, two sampling techniques, random under-sampling (RUS) [71] and synthetic minority over-sampling (SMOTE) [72] were used to increase the sensitivity of the classification models. Furthermore, to reduce the bias estimation of the performance metrics [73], we have performed 5-fold stratified cross validation [74]. Finally, we used Random Forests (RF) [75] and Artificial Neural Networks (ANN) [76] on the training dataset and validated our results using testing dataset with various performance metrics. As a result of implementing all of the techniques, the model is capable of predicting brain cancer survival with an accuracy of 98.9% and an AUC score of 97.2%. To our best knowledge, a classification algorithm with such a high degree of accuracy has not been reported in extant studies on brain cancer based on clinical data.

A web-based predictive tool: Finally, by integrating the best performing model, we have developed a web-based interactive tool that facilitates the application of the proposed model and allows a deeper understanding of how a particular factor influences survival when other features remain constant.

As a result of developing the algorithm with a high level of accuracy and web-based prediction tools, practitioners will be able to gain valuable insight from clinical data that could help them estimate a patient's chance of surviving this deadly disease. With this system, physicians can create treatment plans that are tailored to the specific needs of the patient, rather than relying on personal experience, anecdotes, or aggregate risk assessments [77]. Additionally, the web-based tool automatically incorporates the machine learning algorithm, offering a simple user interface to get information from the model. Therefore, doctors can make more informed decisions without needing any knowledge of machine learning.

Following is an outline of the remainder of this paper. Section 2 describes the data preprocessing, cleaning, balancing, and variable selection techniques and the classification models employed in the

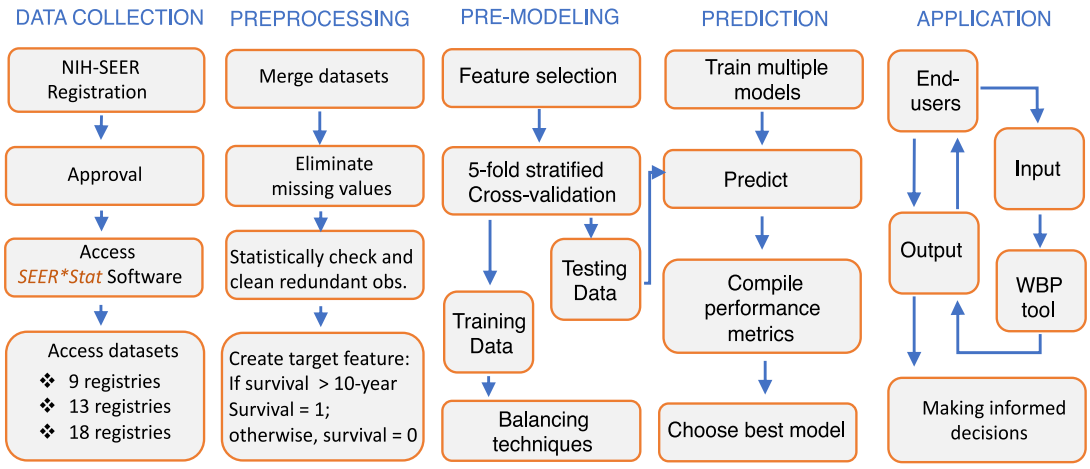

Fig. 1. An overview of the study methodology.

current study. Results and a web-based prediction tool are presented in Section 3. The limitations of the current study are discussed in Section 4. The paper concludes with a summary, conclusions, and guidelines for future research directions, which are presented in Section 5.

2. Materials and methods

As discussed previously, the objective of this study is to propose a comprehensive, data-driven method that will provide better predictions of cancer survival over a specific period of time. We will cover all the methods and procedures with short descriptions for the proposed study in the following subsections.

2.1. An overview of the study design

An extensive data-driven procedure comprising five consecutive stages was employed in the present work, as illustrated in Fig. 1. The first stage was to finish registration. Upon receiving approval, we were able to download three sets of raw data pertaining to patients with brain cancer through the use of *SEER*Stat* software. In the second stage, three datasets were merged, missing values were eliminated, a regression diagnostic test was performed to clean the redundant observations, and a target feature was created. In stage three, significant features were selected through LASSO, then testing and training datasets were created using five-fold-stratified cross-validation, and balancing techniques were applied to the training dataset. In the fourth stage, we trained multiple machine learning models, validated their performance on test data, and then chose the best model based on its performance. Finally, a web-based decision support tool was developed to facilitate the practical implementation of the most appropriate model that can be used by practitioners without requiring any previous knowledge of machine learning algorithms. The following subsections provide additional details about some key steps in the research methodology.

2.1.1. Data source, access, and collection

After submitting the request form and receiving approval, data for this study was obtained from the Surveillance, Epidemiology, and End Results (SEER) program of the National Cancer Institute using SEER*Stat software. Three datasets of clinical and histological records of patients with brain tumors and other central nervous system tumors were collected from 9, 13, and 18 registries between 1975 and 2018 [78].

2.1.2. Data pre-processing

Three datasets have been combined into one dataset and the missing values and unnecessary features that are not relevant to brain cancer have been removed. It has been determined that a number of features are not available for a particular year, and these are indicated by "Blank(s)"; we have removed the rows of data that contain at least one "Blank(s)".

2.1.3. Cook's distance to detect redundant observations

In regression analysis, Cook's distance is used to identify influential outliers in a set of independent features [79]. The Cook distance was first proposed by American statistician Dennis Cook in 1977 [67]. Cook's distance, D, is generally calculated after removing the ith data point from the model and recompiling the regression estimate. It indicates, in essence, how much the values of the outputs in the regression model change when the ith observation is eliminated [80]. We used a regression model in which *survival months* (before transforming into binary) is a dependent feature and independent features are presented in Table 1. Detailed information about the features documentation can be found at https://seer.cancer.gov/analysis/. A regression diagnostic is used to determine which outliers are influential after a model fitting process for 90,390 records of cancer patients has been performed. The results of the regression diagnostic are shown in Fig. 2. Fig. 2(A) illustrates that some cook's distances are significantly large, which indicates the presence of influential outliers. Furthermore, Fig. 2(B) is a bubble plot, where the x-axis represents the fitted values, the y-axis represents the standardized residuals, and the proportional size of each bubble represents the Cook's distance. We can see that some bubbles are larger than others, confirming the presence of influential outliers in the data. Even though the scale location plot in Fig. 2(C) shows some minor deviations from the homogeneity of variance, Fig. 2(D) shows that the points broadly fall along the reference line (red dashed line). Therefore, overall the assumption of normality was met for the regression model. We have used the recommended threshold value of 4/n [80,81] for Cook's distance, which implies that observations with Cook's distance higher than 4/n are considered to be influential outliers and have been eliminated from the dataset. As a result of applying the Cook distance, approximately 9% of the 90,390 records were removed. It is recommended to consult a physician before determining an appropriate threshold value and discarding outliers, since removing outliers may result in the loss of information. For simplicity, the method of eliminating redundant observations using Cook's distance threshold value is referred to as CDE.

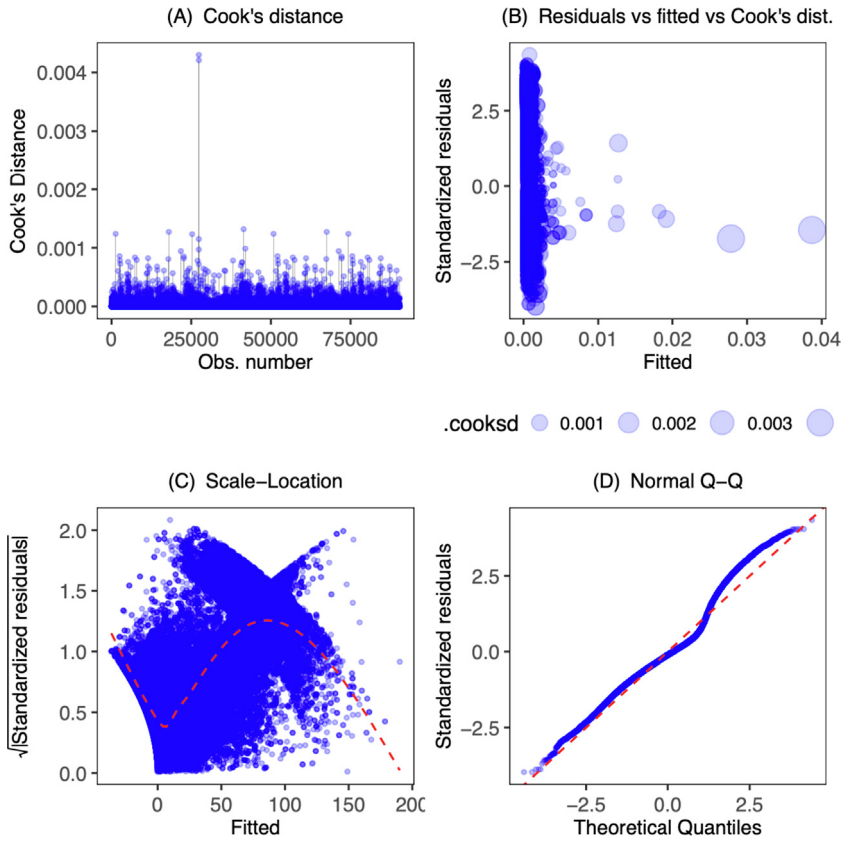

Fig. 2. Diagnostic plots for analyzing redundant or influential observations.

Table 1
Feature selection using LASSO.

Feature	Feature coefficients	Selected feature
Age	−2.31	✓
CS version input original	−1.103	✓
RX Summ-Surg Prim	0.418	✓
Year of diagnosis	−0.340	✓
ICD-O-3 Hist behav	−0.308	✓
CS extension	−0.260	✓
First malignant primary indicator	0.240	✓
Grade	−0.171	✓
CS version input current	0.032	✓
Primary site	0.029	✓
SS-seq	−0.001	✓
Laterality	−0.001	✓
CS tumor size	0	
ICD-O-3 Hist behav malignant	0	
Combined summary stage	0	
Regional nodes	0	
Histologic ICD-O-3	0	
Total number of malignant tumors	0	
Total number of benign tumors	0	
CS site specific factor 1	0	

2.1.4. Formulate the target feature

After performing all pre-processing steps as outlined above for data cleaning, we are left with 83,599 observations and 20 independent features, and a dependent feature, *survival months*. The main purpose of this study is to predict the 10-year survival of brain cancer patients. Therefore, a patient who has a value higher than 120 for *survival months* is considered to survive. Otherwise, we consider them to not survive. Hence, the target feature is derived using the following piece-wise function

$$\text{Survival 10-year} = \begin{cases} 0 & \text{if } survival\ months \leq 120 \\ 1 & otherwise \end{cases} \tag{1}$$

2.1.5. Feature selection

Feature selection techniques are primarily used to identify the key feature contributing to the forecasting target from a number of features [82]. Moreover, it enhances the overall predictive power of the classifier [83] and significantly reduces the computational cost [84]. In addition, having fewer features makes models easier to understand and reduces their complexity, which makes them more useful to practitioners and decision makers in healthcare settings [7]. Three effective methods currently are available for selecting features: filter, wrapper, and embedded techniques [70]. We selected the Least Absolute Shrinkage and Selection Operator (LASSO, an embedded technique) for this study due to its simplicity and because it outperformed other techniques in the preliminary analysis phase. LASSO was originally proposed by Robert Tibshirani [85]. For linear regression with a standardized independent feature x_{ab} and response values y_a for $a = 1, 2, \ldots, n$ and $b = 1, 2, \ldots, r$, LASSO solves the $L1$-penalized regression problem of computing $\gamma = \{\gamma_b\}$ by minimizing the following equation [85]

$$\sum_{a=1}^{n}\left(y_a - \sum_{b} x_{ab}\gamma_b\right)^2 + \alpha \sum_{b=1}^{r} |\gamma_b| \tag{2}$$

where n and r are the numbers of observations and independent features, respectively. It is equivalent to minimizing the sum of squares with a restriction of the form $\sum |\gamma_b| \leq t$.

2.1.6. Stratified 5-fold cross-validation

There are several problems that occur during the training of a classifier, including overfitting, in which the error on the training set is driven to a very small value, but the error increases when new data is presented. Thus, cross-validation is used to prevent overfitting [86]. The main objective of cross-validation in machine learning is to increase the predictive power of machine learning algorithms for unseen data. The K-fold cross-validation technique is applied due to

its simplicity and ease of use in training and validating the model. The K-fold partition may be purely random; however, some folds will have distinct distributions [87]. Therefore, we applied 5-fold stratified cross-validation techniques by randomly splitting the dataset into 5 equal folds, each fold containing the same class distribution [88].

2.1.7. Balancing techniques

A dataset that has uneven amounts or distributions of data among classes is imbalanced. Often, a classifier that is optimizing for a high performance metric will make biased predictions towards classes with more available data points, a majority class. This causes classes with less available data points, the minority classes, to be underpredicted [89]. In survival problems, such as the one being explored in this paper, a significant difference in the survival classes can result in a problem of inbalance in the dataset [7]. To deal with these imbalance problems, we perform two methods of sampling: random under-sampling (RUS) [71] and the synthetic minority over-sampling technique (SMOTE) [72]. We perform these two sampling methods on the training data for our models. This helps our models learn the differences in each class equally and ensures that the test and validation data remains untouched, as close to real world data as possible.

Random under-sampling is a simple method of balancing classes. In RUS, a random data point from the majority class is selected. This data point is then removed from the dataset. This continues until the majority class has the same amount of data as the minority class [71]. A clear potential downside of this method is we are reducing the amount of data available for training.

The synthetic minority over-sampling technique is a technique that increases the amount of data in the minority classes. It does this by creating synthetic data for the minority classes. The synthetic data is generated by joining a random minority class data point with its k nearest neighbors. This creates a line segment joining the point to each of its neighbors. After randomly selecting which neighbor to generate data from, SMOTE randomly picks a point along the line segment joining the two neighbors as a new minority class data point [72]. By repeating this process, SMOTE creates new data for the minority class. This data is synthetic and therefore has the potential to incorrectly represent the true minority training data, a potential downside of SMOTE.

2.1.8. Random forest

Random Forests are a type of machine learning model that can be used for both classification and regression problems [90]. Random Forests are composed of many Decision Trees that have aspects of randomness in them [75]. Decision Trees are prediction models that recursively partition the space of the data they are classifying into smaller regions. While doing so, they build a tree with decisions at each node. The leaves of the tree are the prediction category associated with the decisions leading to that terminal leaf [91]. Random Forests typically make a prediction by training many Decision Trees and taking the most voted class as the final prediction, improving the predictive capabilities over a single Decision Tree [75].

2.1.9. Artificial neural networks

An artificial neural network is a representation of the human brain [92]. It consists of neurons that can be activated. These neurons are connected to other neurons in layers, creating a neural network. An artificial neural network typically consists of an input layer, one or more hidden layers, and an output layer. The connections between the neurons across layers are weighted, and each neuron has a bias [93]. Each neuron contains an activation function that takes the weighted sum of the previously connected nodes' outputs as and the current bias as its input. The output of the activation function is the output of a neuron. The amount that an artificial neural network is incorrect can be quantified through a loss function. This loss function can be minimized using an algorithm called backpropagation [94]. Backpropagation updates the weights and biases from the error of the network, enabling the artificial neural network to approximate unknown nonlinear functions based on data, i.e. learn [93]. Artificial neural networks are the foundation for deep learning and have had large impacts in fields such as image recognition, natural language processing, medicine, and more [95].

2.1.10. Performance metrics

In order to understand how well our machine learning models are performing and compare performance among models, we calculate the accuracy, F1-score, and AUC score of our models.

Accuracy: Accuracy is an essential metric that demonstrates how correct a models predictions are. It can be defined as follows for a binary classification problem [96].

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

In the above equation, TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives. In problems where classes are imbalanced, such as in the dataset we are looking at in this paper, accuracy can misrepresent the success of the model [96]. Consider a situation where a minority class label belongs to 5% of the data. In that case, a model with 95% accuracy could theoretically always be incorrect about the predictions for that class. In order to mitigate this issue, we can introduce precision, recall, and the F1-score.

Precision: Precision can quantify how well a model can predict positive values and is defined as follows [96].

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4}$$

Recall: Recall quantifies the proportion of positive outcomes the model correctly identifies and is defined as follows [96].

$$\text{Recall} = \frac{TP}{TP + FN} \tag{5}$$

F1-Score: The F-measure considers both the precision and recall, and is defined as follows [97].

$$F - \text{score} = (1 + \alpha^2)\frac{P \times R}{\alpha^2(P + R)} \tag{6}$$

In the above equation, P is precision, R is recall, and α represents their balance. When $\alpha = 1$, the F-measure can be called an F1-score [98].

ROCs and the AUC score: The receiver operating characteristic (ROC) curve is a popular method of showing the tradeoff between the recall and the false positive rate. Typically, it is plotted on a graph with the false positive rate on the x axis and the recall on the y axis [99]. The top left point of the curve is the best tradeoff of the two. By finding the area under the ROC curve, the area under curve (AUC) metric can be calculated. This will be a value between 0 and 1 that quantifies the performance of the model in terms of its ROC curve. The higher the AUC, the better the classification model is.

3. Results

3.1. Feature selection and classification results

To perform feature selection, we have applied several well-known feature selection techniques (Random Forest importance, Forward selection, etc.). LASSO proved to be more effective in our experimental study than other techniques. Therefore, we have included only LASSO in this study for simplicity. Based on their absolute value, we have presented the relative importance of each of the features from most important to least important in Table 1. We observe that only twelve (indicated by a checkmark) of the twenty features are

Table 2
Performance metrics for stratified 5-fold cross-validation with RF and ANN.

Model	Balancing technique	Confusion matrix		Accuracy	F1-score	AUC score
CDE + RF	SMOTE	15273	121	0.989	0.930	0.972
		66	1259			
	RUS	14634	760	0.954	0.775	0.973
		8	1317			
RF	SMOTE	15684	245	0.975	0.901	0.948
		189	1960			
	RUS	14700	1229	0.929	0.766	0.949
		52	2097			
CDE + ANN	SMOTE	14529	865	0.936	0.681	0.899
		193	1132			
	RUS	13695	1699	0.895	0.591	0.924
		54	1271			
ANN	SMOTE	14347	1582	0.869	0.536	0.768
		782	1367			
	RUS	12922	3007	0.810	0.501	0.807
		424	1725			

selected by LASSO. Again, the coefficient [100] for *Age* is negative, which is the highest, suggesting that brain cancer survival is negatively associated with age. It is consistent with the findings of Fisher et al. [101], who demonstrated that the survival probability for brain cancers (Glioma and glioma subtypes, including glioblastoma) is lower for older individuals.

Furthermore, based on our comprehensive experimental analysis, we found that RF and ANN classification models outperformed other classification models (such as more complicated neural network architectures such as LSTMs). Therefore, in this study, we have considered only these two models for the sake of simplicity. We evaluated the models using the performance metrics (accuracy, F1-score, and AUC score) described above. The results were achieved using five-fold stratified cross-validation for each model. The final results were derived by averaging the results across all folds and are presented in Table 2. Our study showed that RF and ANN with CDE outperform RF and ANN independently, which should help the reader understand what contributions are being made by applying CDE. However, the CDE with RF outperformed the CDE with ANN in both cases of SMOTE and RUS. We presented the detailed prediction results of the validation data sets in the form of confusion matrices [38] in Table 2. It will assist in selecting the appropriate balancing techniques. Although the AUC of the models based on RUS is marginally higher than SMOTE, when it comes to the F1-score, SMOTE outperformed RUS in all scenarios. Furthermore, the confusion matrix also indicates that the number of false positives and false negatives with SMOTE is similar; however, with RUS, the number of false positives and false negatives differs significantly. Thus, SMOTE outperformed RUS on an overall basis across all scenarios. This result is not surprising when a dataset contains many features. To train the model effectively, the number of observations should also increase exponentially as the number of features increases [102]. Hence, rather than decreasing the number of observations (as in RUS), increasing the number of observations (as in SMOTE) might improve the model's predictive power, as there is no data loss [7].

Therefore, we propose to use RF combined with CDE and SMOTE, which achieved a classification accuracy of 0.989 with an F1-score of 0.930 and an AUC score of 0.972 for the prediction of brain cancer survival (10-year) on SEER data. We have used a 10-year survival to conduct this study, as many studies have employed 10-year survival [7,23,24] in the extant literature on cancer. In addition, Kim et al. concluded in their study that it takes at least five years for a specific cancer patient's record to be marked as survived or dead [103]. It should be noted that there is no universally accepted best method to use when making predictions. To find the best model for each scenario, it is necessary to conduct trial and error experiments [104].

3.2. Interactive web-based prediction tool

Implementing a machine-learning algorithm for predicting cancer survivability can be quite complex and time-consuming, as it requires an in-depth knowledge of mathematics, statistics, and machine learning, and requires programming proficiency. Therefore, developing an interactive web-based tool that incorporates our proposed high degrees of accuracy model for examining cancer survivability can be a helpful medium between physicians and machine learning algorithms. Thus, we have developed a prototype of a web-based interactive tool to benefit the readers of this study. Accordingly, RF was selected for the prototype based on the results of the performance metrics explored in the previous subsection. This prototype was developed using the Streamlit environment, an open-source Python platform for developing web-based interactive applications. Upon gathering patient information, the end-user can input the data. By clicking on the *Predict* button, the tool will calculate and display the patient's survival probability. Based on providing all information for the hypothetical patient (named as A), the tool predicts that the 10-year brain cancer survival chance is 70.2%; the screenshot of the results is presented in Fig. 3. Note that the classifier in RF is designed to make a binary prediction in each tree by taking a majority vote in the terminal node of the tree. Based on their findings, Malley et al. [105] concluded that RF algorithms are valid machine learning methods for estimating individual probabilities for binary responses. Moreover, this tool is not only useful for predicting survival probability, but it is also useful for practitioners to perform scenario analysis. For example, suppose we increase the age of our hypothetical patient (A) from 8 to 13, given that the other attributes remain constant. In that case, the 10-year survival probability of the patient decreases from 70.2% to 67.6%. This scenario analysis for a particular feature may lead practitioners to more efficient and effective treatments or may allow them to adjust their treatment plans for cancer patients. It is important to note that the prototype was developed for demonstration purposes only and was developed based on the first 20,000 records. Therefore, this prototype cannot be used for medical consultation, diagnosis or treatment.

4. Limitations

Assessment of limitations in a research project can serve as a base for further enhancements or improvements to the study. There are a number of limitations to the study, some of which we highlight here. First, prognostic factors such as Karnofsky performance status at the time of diagnosis and other comorbidities were not included [106].

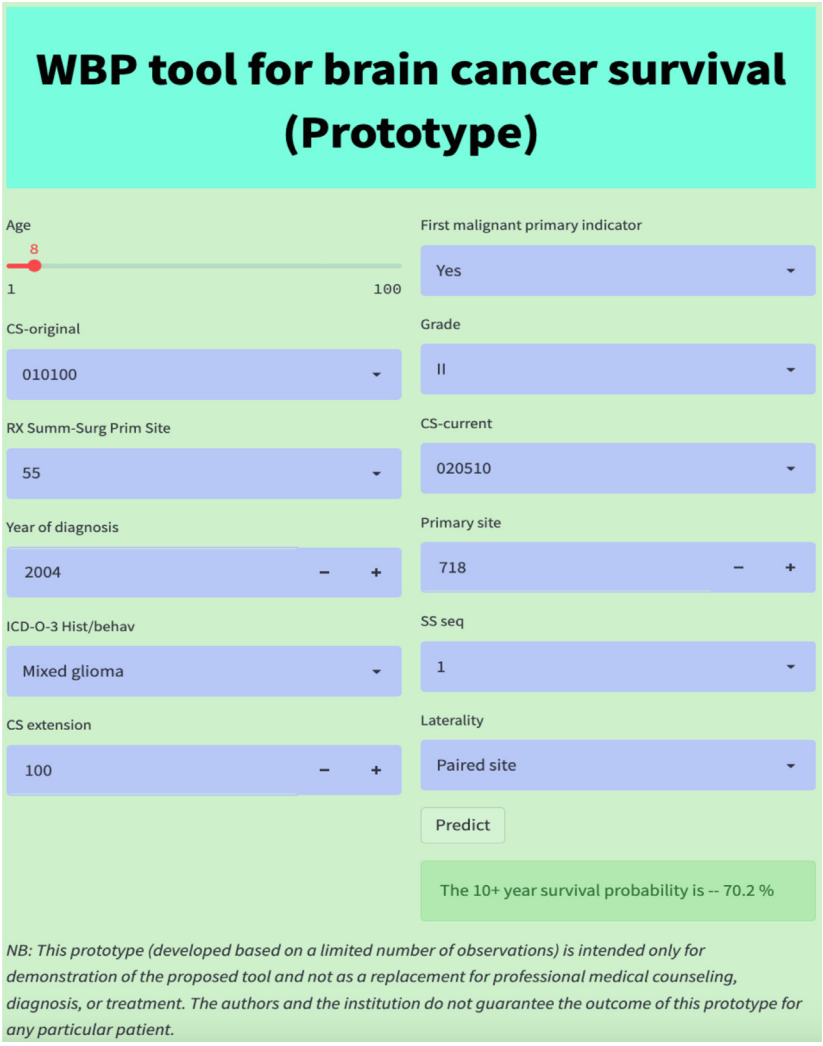

Fig. 3. A screenshot of the prototype tool. You can find a detailed description and related information about the features at http://seer.cancer.gov/seerstat/variables/seer/ajcc-stage. The tool predicts that the 10-year survival probability is 70.2% based on the information of the patient. The link for this tool is available at https://share.streamlit.io/gopalnath1926/brain_cancer/main/app_BCNS.py.

Second, the SEER dataset has no record regarding disease-free progression or corresponding treatment choices. Furthermore, all patients included in the study underwent both initial chemotherapy and radiotherapy [36]. Third, even though the SEER dataset for this study is from 1973 to 2018, some important features have been recorded after 2003 (*CS version input original, CS tumor size, Combined summary stage*, etc.). We, therefore, removed a large number of observations before 2003 that contained missing observations. Moreover, with larger sample sizes, prediction accuracy increases [107]. It is likely since smaller training sets tend to be more heterogeneous [108], and some models are more robust when the sample size is small [109]. Fourth, several essential new features have been added to the SEER dataset recently. For example, $HER2$ (included after 2010) represents a gene type that can contribute to cancer development, and $AJCC-7\ T/N/M$ contains information regarding tumors, lymph nodes, and metastasis of the cancer tumor [110]. However, they were not included in the study because the data was collected over a limited time, and it usually takes at least five years after diagnosis for a cancer patient's record to be marked as survival or death [103]. Fifth, the collection, coding, and quality assessment of SEER data may require up to three years. For instance, SEER data for 2019 cases will be available for analysis in 2022. It is unlikely that this unexpected delay will significantly affect most studies, but it may play a role in particular experiments [111]. It is important to note that, even though data mining can assist in making a diagnosis or prescribing a treatment, it cannot replace a physician's intuitive judgment and interpretive ability [112]. Finally, the findings that are obtained through the use of machine learning in this study should be evaluated by medical specialists to determine if they are valid and applicable.

5. Overview, conclusions, and future research plan

This paper presents a comprehensive analysis with a novel framework comprising multiple phases aiming to predict brain cancer survivability with high accuracy. The proposed methodology was developed employing three independent sets of SEER datasets (9, 13, and 18 registries) from 1973 to 2018. Having obtained a large dataset from the SEER program, we underwent a long process of transforming the data to use it. As medical data is obtained from many sources, such as images, interviews, physician notes, etc., there is a possibility of data entry errors leading to some redundant or inconsistent observations. We have therefore applied statistical diagnostic techniques (Cook's distances) to check and clean the data rather than relying solely on personal assumptions. We have then transformed the target variable from continuous to binary. If a patient survives ten years from the time of initial diagnosis of brain cancer, the survival value is 1; otherwise,

the survival value is 0. We have applied two balancing techniques (SMOTE and RUS) to address the imbalance issues induced by the binary classification (8% survival, 92% death) of the target feature. We have used LASSO to select the most significant factors related to brain cancer survival and subsequently deployed two popular machine learning models, RB and ANN. After performing all of the steps above, the most parsimonious model was developed, as shown in Table 2, where the top model achieved a score of 0.989, 0.93, and 0.972 for accuracy, F1, and AUC, respectively. Thus, the present study contributes significantly to the existing literature on brain cancer by predicting brain cancer survivability with a high degree of accuracy.

Furthermore, we have developed a prototype for the purpose of proving the concept. Our prototype automatically incorporates our chosen machine learning model, so it is ready to be used by doctors who have no prior knowledge of machine learning. Upon making minor adjustments and consulting with a medical practitioner, this tool can be very useful for identifying high-risk patients. Additionally, the web-based interactive tool can be used not only for survival prediction but also for scenario analysis, as discussed previously, without an extensive understanding of machine learning algorithms. Medical practitioners may utilize this tool to adapt patients' treatment plans quickly.

In this study, we used data mining techniques to develop a parsimonious prognosis model with high accuracy without collaboration and guidance from medical specialists. Furthermore, the trends that are identified by using data mining approaches may not necessarily be associated with the cancer prognosis or may not be relevant to healthcare professionals. Thus, an in-depth evaluation by a medical professional in this field is necessary to apply the proposed method in a health care setting.

A research extension of this study could be to validate our comprehensive techniques with the help of medical experts on other types of cancer and to compare with the current literature to see whether there is an improvement in the model accuracy.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share the data.

References

[1] C. Liu, H. Zong, Developmental origins of brain tumors, Curr. Opin. Neurobiol. 22 (5) (2012) 844–849, http://dx.doi.org/10.1016/j.conb.2012.04.012.

[2] K.D. Miller, Q.T. Ostrom, C. Kruchko, N. Patil, T. Tihan, G. Cioffi, H.E. Fuchs, K.A. Waite, A. Jemal, R.L. Siegel, J.S. Barnholtz-Sloan, Brain and other central nervous system tumor statistics, 2021, CA: Cancer J. Clin. 71 (5) (2021) 381–406, http://dx.doi.org/10.3322/caac.21693.

[3] Cancer stat facts: Brain and other nervous system cancer, 2022, https://seer.cancer.gov/statfacts/html/brain.html, Accessed: 2022-06-01.

[4] E. Connell, N. Bartlett, J. Harvey, L. Moon, M. Short, B. Davis, Brain and Other Central Nervous System Cancers, Australian Institute of Health and Welfare, 2017.

[5] A.P. Patel, J.L. Fisher, E. Nichols, F. Abd-Allah, J. Abdela, A. Abdelalim, et al., Global, regional, and national burden of brain and other CNS cancer, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016, Lancet Neurol. 18 (4) (2019) 376–393, http://dx.doi.org/10.1016/s1474-4422(18)30468-x.

[6] J.F. Desforges, W.L. McGuire, G.M. Clark, Prognostic factors and treatment decisions in axillary-node-negative breast cancer, N. Engl. J. Med. 326 (26) (1992) 1756–1761, http://dx.doi.org/10.1056/nejm199206253262607.

[7] S. Simsek, U. Kursuncu, E. Kibis, M. AnisAbdellatif, A. Dag, A hybrid data mining approach for identifying the temporal effects of variables associated with breast cancer survival, Expert Syst. Appl. 139 (2020) 112863, http://dx.doi.org/10.1016/j.eswa.2019.112863.

[8] L. Chato, S. Latifi, Machine learning and deep learning techniques to predict overall survival of brain tumor patients using MRI images, in: 2017 IEEE 17th International Conference on Bioinformatics and Bioengineering, BIBE, 2017, pp. 9–14, http://dx.doi.org/10.1109/BIBE.2017.00-86.

[9] A. Rafi, T. Madni, U. Janjua, M. Ali, M. Abid, Multi-level dilated convolutional neural network for brain tumour segmentation and multi-view-based radiomics for overall survival prediction, Int. J. Imaging Syst. Technol. 20 (2021) 1–17, http://dx.doi.org/10.1002/ima.22549.

[10] L. Weninger, O. Rippel, S. Koppers, D. Merhof, Segmentation of brain tumors and patient survival prediction: Methods for the BraTS 2018 challenge, in: A. Crimi, S. Bakas, H. Kuijf, F. Keyvan, M. Reyes, T. van Walsum (Eds.), Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, Springer International Publishing, Cham, 2019, pp. 3–12, http://dx.doi.org/10.1007/978-3-030-11726-9_1.

[11] S. Starke, C. Eckert, A. Zwanenburg, S. Speidel, S. Löck, S. Leger, An integrative analysis of image segmentation and survival of brain tumour patients, in: A. Crimi, S. Bakas (Eds.), Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, Springer International Publishing, Cham, 2020, pp. 368–378, http://dx.doi.org/10.1007/978-3-030-46640-4_35.

[12] L. Sun, S. Zhang, H. Chen, L. Luo, Brain tumor segmentation and survival prediction using multimodal MRI scans with deep learning, Front. Neurosci. 13 (2019) 1–9, http://dx.doi.org/10.3389/fnins.2019.00810.

[13] R. Shai, T. Shi, T.J. Kremen, S. Horvath, L.M. Liau, T.F. Cloughesy, P.S. Mischel, S.F. Nelson, Gene expression profiling identifies molecular subtypes of gliomas, Oncogene 22 (31) (2003) 4918–4923, http://dx.doi.org/10.1038/sj.onc.1206753.

[14] P.S. Mischel, R. Shai, T. Shi, S. Horvath, K.V. Lu, G. Choe, D. Seligson, T.J. Kremen, A. Palotie, L.M. Liau, T.F. Cloughesy, S.F. Nelson, Identification of molecular subtypes of glioblastoma by gene expression profiling, Oncogene 22 (15) (2003) 2361–2373, http://dx.doi.org/10.1038/sj.onc.1206344.

[15] C.L. Nutt, R.A. Betensky, M.A. Brower, T.T. Batchelor, D.N. Louis, A.O. Stemmer-Rachamimov, YKL-40 is a differential diagnostic marker for histologic subtypes of high-grade gliomas, Clin. Cancer Res. 11 (6) (2005) 2258–2264, http://dx.doi.org/10.1158/1078-0432.ccr-04-1601.

[16] Y. Liang, M. Diehn, N. Watson, A.W. Bollen, K.D. Aldape, M.K. Nicholas, K.R. Lamborn, M.S. Berger, D. Botstein, P.O. Brown, M.A. Israel, Gene expression profiling reveals molecularly and clinically distinct subtypes of glioblastoma multiforme, Proc. Natl. Acad. Sci. 102 (16) (2005) 5814–5819, http://dx.doi.org/10.1073/pnas.0402870102.

[17] L. Zhu, X. Sun, W. Bai, Nomograms for predicting cancer-specific and overall survival among patients with endometrial carcinoma: A SEER based study, Front. Oncol. 10 (2020) 1–12, http://dx.doi.org/10.3389/fonc.2020.00269.

[18] J. Rosenberg, Y.L. Chia, S. Plevritis, The effect of age, race, tumor size, tumor grade, and disease stage on invasive ductal breast cancer survival in the U.S. SEER database, Breast Cancer Res. Treat. 89 (1) (2005) 47–54, http://dx.doi.org/10.1007/s10549-004-1470-1.

[19] V.L. Tsikitis, B.C. Wertheim, M.A. Guerrero, Trends of incidence and survival of gastrointestinal neuroendocrine tumors in the United States: A seer analysis, J. Cancer 3 (2012) 292–302, http://dx.doi.org/10.7150/jca.4502.

[20] S.M. Ryu, S.-H. Lee, E.-S. Kim, W. Eoh, Predicting survival of patients with spinal ependymoma using machine learning algorithms with the SEER database, World Neurosurg. 124 (2019) e331–e339, http://dx.doi.org/10.1016/j.wneu.2018.12.091.

[21] L. Li, Z. Yang, Y. Hou, Z. Chen, Moving beyond the Cox proportional hazards model in survival data analysis: a cervical cancer study, BMJ Open 10 (7) (2020) e033965, http://dx.doi.org/10.1136/bmjopen-2019-033965.

[22] R. Huang, Z. Sun, H. Zheng, P. Yan, P. Hu, H. Yin, J. Zhang, T. Meng, Z. Huang, Identifying the prognosis factors and predicting the survival probability in patients with non-metastatic chondrosarcoma from the SEER database, Orthop. Surg. 11 (5) (2019) 801–810, http://dx.doi.org/10.1111/os.12521.

[23] C. Lee, A. Light, A. Alaa, D. Thurtle, M. van der Schaar, V.J. Gnanapragasam, Application of a novel machine learning framework for predicting non-metastatic prostate cancer-specific mortality in men using the Surveillance, Epidemiology, and End Results (SEER) database, Lancet Digit. Health 3 (3) (2021) e158–e165, http://dx.doi.org/10.1016/S2589-7500(20)30314-9.

[24] M. Lundin, J. Lundin, H. Burke, S. Toikkanen, L. Pylkkänen, H. Joensuu, Artificial neural networks applied to survival prediction in breast cancer, Oncology 57 (4) (1999) 281–286, http://dx.doi.org/10.1159/000012061.

[25] M.H. van Vliet, H.M. Horlings, M.J. van de Vijver, M.J.T. Reinders, L.F.A. Wessels, Integration of clinical and gene expression data has a synergetic effect on predicting breast cancer outcome, PLoS ONE 7 (7) (2012) e40358, http://dx.doi.org/10.1371/journal.pone.0040358.

[26] M. Kolasa, R. Wojtyna, R. Długosz, W. Jóźwicki, Application of artificial neural network to predict survival time for patients with bladder cancer, in: E. Kącki, M. Rudnicki, J. Stempczyńska (Eds.), Computers in Medical Activity, Springer Berlin Heidelberg, Berlin, Heidelberg, 2009, pp. 113–122, http://dx.doi.org/10.1007/978-3-642-04462-5_11.

[27] Y. Shimada, F. Sato, K. Shimizu, G. Tsujimoto, K. Tsukada, cDNA microarray analysis of esophageal cancer: discoveries and prospects, Gen. Thorac. Cardiovasc. Surg. 57 (7) (2009) 347–356, http://dx.doi.org/10.1007/s11748-008-0406-9.

[28] H.S. Phillips, S. Kharbanda, R. Chen, W.F. Forrest, R.H. Soriano, T.D. Wu, A. Misra, J.M. Nigro, H. Colman, L. Soroceanu, P.M. Williams, Z. Modrusan, B.G. Feuerstein, K. Aldape, Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis, Cancer Cell 9 (3) (2006) 157–173, http://dx.doi.org/10.1016/j.ccr.2006.02.019.

[29] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, E.S. Lander, Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring, Science 286 (5439) (1999) 531–537, http://dx.doi.org/10.1126/science.286.5439.531.

[30] K. Shameer, K.W. Johnson, B.S. Glicksberg, J.T. Dudley, P.P. Sengupta, Machine learning in cardiovascular medicine: are we there yet? Heart 104 (14) (2018) 1156–1164, http://dx.doi.org/10.1136/heartjnl-2017-311198.

[31] K.C. Koo, K.S. Lee, S. Kim, C. Min, G.R. Min, Y.H. Lee, W.K. Han, K.H. Rha, S.J. Hong, S.C. Yang, B.H. Chung, Long short-term memory artificial neural network model for prediction of prostate cancer survival outcomes according to initial treatment strategy: development of an online decision-making support system, World J. Urol. 38 (10) (2020) 2469–2476, http://dx.doi.org/10.1007/s00345-020-03080-8.

[32] J.A. Bartholomai, H.B. Frieboes, Lung cancer survival prediction via machine learning regression, classification, and statistical techniques, in: 2018 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), IEEE, 2018, http://dx.doi.org/10.1109/isspit.2018.8642753.

[33] C. Guo, J. Wang, Y. Wang, X. Qu, Z. Shi, Y. Meng, J. Qiu, K. Hua, Novel artificial intelligence machine learning approaches to precisely predict survival and site-specific recurrence in cervical cancer: A multi-institutional study, Transl. Oncol. 14 (5) (2021) 101032, http://dx.doi.org/10.1016/j.tranon.2021.101032.

[34] C.-F. Wu, Y.-J. Wu, P.-C. Liang, C.-H. Wu, S.-F. Peng, H.-W. Chiu, Disease-free survival assessment by artificial neural networks for hepatocellular carcinoma patients after radiofrequency ablation, J. Formos. Med. Assoc. 116 (10) (2017) 765–773, http://dx.doi.org/10.1016/j.jfma.2016.12.006.

[35] M.S. Iraji, Prediction of post-operative survival expectancy in thoracic lung cancer surgery with soft computing, J. Appl. Biomed. 15 (2) (2017) 151–159, http://dx.doi.org/10.1016/j.jab.2016.12.001.

[36] K.A. Samara, Z.A. Aghbari, A. Abusafia, GLIMPSE: a glioblastoma prognostication model using ensemble learning—a surveillance, epidemiology, and end results study, Health Inf. Sci. Syst. 9 (1) (2021) 1–13, http://dx.doi.org/10.1007/s13755-020-00134-4.

[37] A.V. Karhade, Q. Thio, P. Ogink, J. Kim, S. Lozano-Calderon, K. Raskin, J.H. Schwab, Development of machine learning algorithms for prediction of 5-year spinal chordoma survival, World Neurosurg. 119 (2018) e842–e847, http://dx.doi.org/10.1016/j.wneu.2018.07.276.

[38] D. Delen, G. Walker, A. Kadam, Predicting breast cancer survivability: a comparison of three data mining methods, Artif. Intell. Med. 34 (2) (2005) 113–127, http://dx.doi.org/10.1016/j.artmed.2004.07.002.

[39] D. Delen, Analysis of cancer data: a data mining approach, Expert Syst. 26 (1) (2009) 100–112, http://dx.doi.org/10.1111/j.1468-0394.2008.00480.x.

[40] B. Srinivas, G.S. Rao, Unsupervised learning algorithms for MRI brain tumor segmentation, in: 2018 Conference on Signal Processing and Communication Engineering Systems, SPACES, 2018, pp. 1181–1184, http://dx.doi.org/10.1109/SPACES.2018.8316341.

[41] G. Vishnuvarthanan, M.P. Rajasekaran, P. Subbaraj, A. Vishnuvarthanan, An unsupervised learning method with a clustering approach for tumor identification and tissue segmentation in magnetic resonance brain images, Appl. Soft Comput. 38 (2016) 190–212, http://dx.doi.org/10.1016/j.asoc.2015.09.016.

[42] P. Sahoo, S. Soltani, A. Wong, A survey of thresholding techniques, Comput. Vis. Graph. Image Process. 41 (2) (1988) 233–260, http://dx.doi.org/10.1016/0734-189X(88)90022-9.

[43] İ. Güler, A. Demirhan, R. Karakış, Interpretation of MR images using self-organizing maps and knowledge-based expert systems, Digit. Signal Process. 19 (4) (2009) 668–677, http://dx.doi.org/10.1016/j.dsp.2008.08.002.

[44] S. Ong, N. Yeo, K. Lee, Y. Venkatesh, D. Cao, Segmentation of color images using a two-stage self-organizing network, Image Vis. Comput. 20 (4) (2002) 279–289, http://dx.doi.org/10.1016/S0262-8856(02)00021-5.

[45] J. Alirezaie, M. Jernigan, C. Nahmias, Automatic segmentation of cerebral MR images using artificial neural networks, IEEE Trans. Nucl. Sci. 45 (4) (1998) 2174–2182, http://dx.doi.org/10.1109/23.708336.

[46] P. Kickingereder, D. Bonekamp, M. Nowosielski, A. Kratz, M. Sill, S. Burth, A. Wick, O. Eidel, H.-P. Schlemmer, A. Radbruch, J. Debus, C. Herold-Mende, A. Unterberg, D. Jones, S. Pfister, W. Wick, A. von Deimling, M. Bendszus, D. Capper, Radiogenomics of glioblastoma: Machine learning–based classification of molecular characteristics by using multiparametric and multiregional MR imaging features, Radiology 281 (3) (2016) 907–918, http://dx.doi.org/10.1148/radiol.2016161382.

[47] J.P. Brunet, P. Tamayo, T.R. Golub, J.P. Mesirov, Metagenes and molecular pattern discovery using matrix factorization, Proc. Natl. Acad. Sci. 101 (12) (2004) 4164–4169, http://dx.doi.org/10.1073/pnas.0308531101.

[48] A. Li, J. Walling, S. Ahn, Y. Kotliarov, Q. Su, M. Quezado, J.C. Oberholtzer, J. Park, J.C. Zenklusen, H.A. Fine, Unsupervised analysis of transcriptomic profiles reveals six glioma subtypes, Cancer Res. 69 (5) (2009) 2091–2099, http://dx.doi.org/10.1158/0008-5472.can-08-2100.

[49] H. Yu, T. Huang, B. Feng, J. Lyu, Deep-learning model for predicting the survival of rectal adenocarcinoma patients based on a surveillance, epidemiology, and end results analysis, BMC Cancer 22 (1) (2022) http://dx.doi.org/10.1186/s12885-022-09217-9.

[50] W. Fang, Z.-Y. Yang, T.-Y. Chen, X.-F. Shen, C. Zhang, Ethnicity and survival in bladder cancer: a population-based study based on the SEER database, J. Transl. Med. 18 (1) (2020) 1–11, http://dx.doi.org/10.1186/s12967-020-02308-w.

[51] H. Liu, X. Qin, L. Zhao, G. Zhao, Y. Wang, Epidemiology and survival of patients with brainstem gliomas: A population-based study using the SEER database, Front. Oncol. 11 (2021) http://dx.doi.org/10.3389/fonc.2021.692097.

[52] H. Sun, H. Ma, G. Hong, H. Sun, J. Wang, Survival improvement in patients with pancreatic cancer by decade: A period analysis of the SEER database, 1981–2010, Sci. Rep. 4 (1) (2014) 1–10, http://dx.doi.org/10.1038/srep06747.

[53] K. Huang, J. Zhang, Y. Yu, Y. Lin, C. Song, The impact of chemotherapy and survival prediction by machine learning in early Elderly Triple Negative Breast Cancer (eTNBC): a population based study from the SEER database, BMC Geriatr. 22 (1) (2022) 1–12, http://dx.doi.org/10.1186/s12877-022-02936-5.

[54] K.J. Jager, P.C. van Dijk, C. Zoccali, F.W. Dekker, The analysis of survival data: the Kaplan–Meier method, Kidney Int. 74 (5) (2008) 560–565, http://dx.doi.org/10.1038/ki.2008.217.

[55] P.J. García-Laencina, P.H. Abreu, M.H. Abreu, N. Afonoso, Missing data imputation on the 5-year survival prediction of breast cancer patients with unknown discrete values, Comput. Biol. Med. 59 (2015) 125–133, http://dx.doi.org/10.1016/j.compbiomed.2015.02.006.

[56] R.O. Alabi, A.A. Mäkitie, M. Pirinen, M. Elmusrati, I. Leivo, A. Almangush, Comparison of nomogram with machine learning techniques for prediction of overall survival in patients with tongue cancer, Int. J. Med. Inform. 145 (2021) 104313, http://dx.doi.org/10.1016/j.ijmedinf.2020.104313.

[57] P.N. Kamalapathy, D.B. Ramkumar, A.V. Karhade, S. Kelly, K. Raskin, J. Schwab, S. Lozano-Calderón, Development of machine learning model algorithm for prediction of 5-year soft tissue myxoid liposarcoma survival, J. Surg. Oncol. 123 (7) (2021) 1610–1617, http://dx.doi.org/10.1002/jso.26398.

[58] H.P. Bhambhvani, A. Zamora, E. Shkolyar, K. Prado, D.R. Greenberg, A.M. Kasman, J. Liao, S. Shah, S. Srinivas, E.C. Skinner, J.B. Shah, Development of robust artificial neural networks for prediction of 5-year survival in bladder cancer, Urol. Oncol.: Semin. Orig. Investig. 39 (3) (2021) 193.e7–193.e12, http://dx.doi.org/10.1016/j.urolonc.2020.05.009.

[59] M.H. Osman, R.H. Mohamed, H.M. Sarhan, E.J. Park, S.H. Baik, K.Y. Lee, J. Kang, Machine learning model for predicting postoperative survival of patients with colorectal cancer, Cancer Res. Treat. 54 (2) (2022) 517–524, http://dx.doi.org/10.4143/crt.2021.206.

[60] P. Tai, E. Yu, V. Vinh-Hung, G. Cserni, G. Vlastos, Survival of patients with metastatic breast cancer: twenty-year data from two SEER registries, BMC Cancer 4 (1) (2004) http://dx.doi.org/10.1186/1471-2407-4-60.

[61] J.T. Senders, P. Staples, A. Mehrtash, D.J. Cote, M.J.B. Taphoorn, D.A. Reardon, W.B. Gormley, T.R. Smith, M.L. Broekman, O. Arnaout, An online calculator for the prediction of survival in glioblastoma patients using classical statistics and machine learning, Neurosurgery 86 (2) (2020) E184–E192, http://dx.doi.org/10.1093/neuros/nyz403.

[62] I. Kaur, M.N. Doja, T. Ahmad, M. Ahmad, A. Hussain, A. Nadeem, A.A.A. El-Latif, An integrated approach for cancer survival prediction using data mining techniques, Comput. Intell. Neurosci. 2021 (2021) 1–14, http://dx.doi.org/10.1155/2021/6342226.

[63] S. Gupta, M.K. Gupta, A comparative analysis of deep learning approaches for predicting breast cancer survivability, Arch. Comput. Methods Eng. 11 (2021) 1–17, http://dx.doi.org/10.1007/s11831-021-09679-3.

[64] M.U. Khan, J.P. Choi, H. Shin, M. Kim, Predicting breast cancer survivability using fuzzy decision trees for personalized healthcare, in: 2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2008, pp. 5148–5151, http://dx.doi.org/10.1109/IEMBS.2008.4650373.

[65] M. Salehi, J. Razmara, S. Lotfi, Development of an ensemble multi-stage machine for prediction of breast cancer survivability, J. AI Data Min. 8 (3) (2020) 371–378, http://dx.doi.org/10.22044/jadm.2020.8406.1978.

[66] S.P. Rajamohana, K. Umamaheswari, K. Karunya, R. Deepika, Analysis of classification algorithms for breast cancer prediction, in: Data Management, Analytics and Innovation, Springer Singapore, 2019, pp. 517–528, http://dx.doi.org/10.1007/978-981-32-9949-8_36.

[67] R.D. Cook, Detection of influential observation in linear regression, Technometrics 19 (1) (1977) 15–18, http://dx.doi.org/10.2307/1268249.

[68] M.A. Duggan, W.F. Anderson, S. Altekruse, L. Penberthy, M.E. Sherman, The surveillance, epidemiology, and end results (SEER) program and pathology, Am. J. Surg. Pathol. 40 (12) (2016) 94–102, http://dx.doi.org/10.1097/pas.0000000000000749.

[69] R. Tibshirani, Regression Shrinkage and selection via the Lasso, J. R. Stat. Soc. Ser. B Stat. Methodol. 58 (1) (1996) 267–288, http://dx.doi.org/10.1111/j.2517-6161.1996.tb02080.x.

[70] C. Chen, Y. Tsai, F. Chang, W. Lin, Ensemble feature selection in medical datasets: Combining filter, wrapper, and embedded feature selection results, Expert Syst. 37 (5) (2020) 1–10, http://dx.doi.org/10.1111/exsy.12553.
[71] Y. Kamei, A. Monden, S. Matsumoto, T. Kakimoto, K. Matsumoto, The effects of over and under sampling on fault-prone module detection, in: First International Symposium on Empirical Software Engineering and Measurement (ESEM 2007), 2007, pp. 196–204, http://dx.doi.org/10.1109/ESEM.2007.28.
[72] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: Synthetic minority over-sampling technique, J. Artificial Intelligence Res. 16 (2002) 321–357, http://dx.doi.org/10.1613/jair.953.
[73] R. Kohavi, A study of cross-validation and bootstrap for accuracy estimation and model selection, in: Proceedings of 14th International Joint Conference on AI, Vol. 2, Morgan Kaufmann, 1995, pp. 1137–1145.
[74] X. Zeng, T.R. Martinez, Distribution-balanced stratified cross-validation for accuracy estimation, J. Exp. Theor. Artif. Intell. 12 (1) (2000) 1–12, http://dx.doi.org/10.1080/095281300146272.
[75] T.K. Ho, Random decision forests, in: Proceedings of 3rd International Conference on Document Analysis and Recognition, Vol. 1, 1995, pp. 278–282, http://dx.doi.org/10.1109/ICDAR.1995.598994.
[76] C. Hervás, A. Garrido, B. Lucena, N. García, E.D. Pedro, Near infrared spectroscopy for classification of iberian pig carcasses using an artificial neural network, J. Near Infrared Spectrosc. 2 (4) (1994) 177–184, http://dx.doi.org/10.1255/jnirs.44.
[77] A. Agrawal, S. Misra, R. Narayanan, L. Polepeddi, A. Choudhary, Lung cancer survival prediction using ensemble data mining on SEER data, Sci. Program. 2012 (2012) 29–42, http://dx.doi.org/10.3233/SPR-2012-0335.
[78] SEER Cancer Statistics Review. Surveillance, Epidemiology, and End Results (SEER) program public-use data (1975—2018). National Cancer Institute, Surveillance Research Program, Cancer Statistics Branch, released April 2003. Based on the November 2020 submission. Diagnosis period 1973—2018, 9, 13 and 18 registries, 2022, www.seer.cancer.gov, Accessed: 2022-03-01.
[79] A. Khan, M.A. Ullah, M. Amin, A.H. Muse, R. Aldallal, M.S. Mohamed, Empirical examination of the Poisson regression residuals for the evaluation of influential points, Math. Probl. Eng. 2022 (2022) 1–9, http://dx.doi.org/10.1155/2022/6995911.
[80] L. Rutkowski, R. Scherer, M. Korytkowski, W. Pedrycz, R. Tadeusiewicz, J.M. Zurada (Eds.), Artificial Intelligence and Soft Computing, Springer International Publishing, 2018, http://dx.doi.org/10.1007/978-3-319-91262-2.
[81] C.M. Salgado, C. Azevedo, H. Proença, S.M. Vieira, Noise versus outliers, in: Secondary Analysis of Electronic Health Records, Springer International Publishing, 2016, pp. 163–183, http://dx.doi.org/10.1007/978-3-319-43742-2_14.
[82] J.-H. Yang, C.-H. Cheng, C.-P. Chan, A time-series water level forecasting model based on imputation and variable selection method, Comput. Intell. Neurosci. 2017 (2017) 1–11, http://dx.doi.org/10.1155/2017/8734214.
[83] E.O. Jessica, M. Hamada, S.I. Yusuf, M. Hassan, The role of linear discriminant analysis for accurate prediction of breast cancer, in: 2021 IEEE 14th International Symposium on Embedded Multicore/Many-Core Systems-on-Chip (MCSoC), IEEE, 2021, http://dx.doi.org/10.1109/mcsoc51149.2021.00057.
[84] M. Alghamdi, M. Al-Mallah, S. Keteyian, C. Brawner, J. Ehrman, S. Sakr, Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford ExercIse Testing (FIT) project, PLOS ONE 12 (7) (2017) e0179805, http://dx.doi.org/10.1371/journal.pone.0179805.
[85] R. Tibshirani, Regression shrinkage and selection via the lasso: a retrospective, J. R. Stat. Soc. Ser. B Stat. Methodol. 73 (3) (2011) 273–282, http://dx.doi.org/10.1111/j.1467-9868.2011.00771.x.
[86] Y. Zhang, L. Wu, S. Wang, Magnetic resonance brain image classification by an improved artificial bee colony algorithm, Prog. Electromagn. Res. 116 (2011) 65–79, http://dx.doi.org/10.2528/pier11031709.
[87] Y. Zhang, S. Wang, G. Ji, A rule-based model for bankruptcy prediction based on an improved genetic ant colony algorithm, Math. Probl. Eng. 2013 (2013) 1–10, http://dx.doi.org/10.1155/2013/753251.
[88] R. May, H. Maier, G. Dandy, Data splitting for artificial neural networks using SOM-based stratified sampling, Neural Netw. 23 (2) (2010) 283–294, http://dx.doi.org/10.1016/j.neunet.2009.11.009.
[89] J.L. Leevy, T.M. Khoshgoftaar, R.A. Bauder, N. Seliya, A survey on addressing high-class imbalance in big data, J. Big Data 5 (1) (2018) http://dx.doi.org/10.1186/s40537-018-0151-6.
[90] L. Breiman, Mach. Learn. 45 (1) (2001) 5–32, http://dx.doi.org/10.1023/a:1010933404324.
[91] B. Efron, T. Hastie, Computer Age Statistical Inference, Cambridge University Press, 2016, http://dx.doi.org/10.1017/cbo9781316576533.
[92] R. Uhrig, Introduction to artificial neural networks, in: Proceedings of IECON '95 - 21st Annual Conference on IEEE Industrial Electronics, IEEE, http://dx.doi.org/10.1109/iecon.1995.483329.
[93] B. Mehlig, Machine Learning with Neural Networks, Cambridge University Press, 2021, http://dx.doi.org/10.1017/9781108860604.
[94] C. Nwankpa, W. Ijomah, A. Gachagan, S. Marshall, Activation functions: Comparison of trends in practice and research for deep learning, 2018, http://dx.doi.org/10.48550/ARXIV.1811.03378.
[95] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444, http://dx.doi.org/10.1038/nature14539.
[96] P. Bruce, A. Bruce, Practical Statistics for Data Scientists, O'Reilly Media, Sebastopol, CA, 2017.
[97] J. Sun, H. Li, H. Fujita, B. Fu, W. Ai, Class-imbalanced dynamic financial distress prediction based on Adaboost-SVM ensemble combined with SMOTE and time weighting, Inf. Fusion 54 (2020) 128–144, http://dx.doi.org/10.1016/j.inffus.2019.07.006.
[98] K. Zhang, H. Su, Y. Dou, Beyond AP: a new evaluation index for multiclass classification task accuracy, Appl. Intell. 51 (10) (2021) 7166–7176, http://dx.doi.org/10.1007/s10489-021-02223-7.
[99] K. Hajian-Tilaki, Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation, Caspian J. Intern. Med. 4 (2) (2013) 627–635.
[100] T. Panagiotidis, T. Stengos, O. Vravosinos, On the determinants of bitcoin returns: A LASSO approach, Finance Res. Lett. 27 (2018) 235–240, http://dx.doi.org/10.1016/j.frl.2018.03.016.
[101] J.L. Fisher, J.A. Schwartzbaum, M. Wrensch, J.L. Wiemels, Epidemiology of brain tumors, Neurol. Clin. 25 (4) (2007) 867–890, http://dx.doi.org/10.1016/j.ncl.2007.07.002.
[102] J. Han, J. Pei, M. Kamber, Data Mining, Southeast Asia Edition, second ed., in: The Morgan Kaufmann Series in Data Management Systems, Morgan Kaufmann, Oxford, England, 2006.
[103] J. Kim, H. Shin, Breast cancer survivability prediction using labeled, unlabeled, and pseudo-labeled patient data, J. Am. Med. Inform. Assoc. 20 (4) (2013) 613–618, http://dx.doi.org/10.1136/amiajnl-2012-001570.
[104] E. Ruiz, F.H. Nieto, A note on linear combination of predictors, Statist. Probab. Lett. 47 (4) (2000) 351–356, http://dx.doi.org/10.1016/s0167-7152(99)00177-7.
[105] J.D. Malley, J. Kruppa, A. Dasgupta, K.G. Malley, A. Ziegler, Probability machines, Methods Inf. Med. 51 (01) (2012) 74–81, http://dx.doi.org/10.3414/me00-01-0052.
[106] K. Barami, L. Lyon, C. Conell, Type 2 diabetes mellitus and glioblastoma multiforme–assessing risk and survival: Results of a large retrospective study and systematic review of the literature, World Neurosurg. 106 (2017) 300–307, http://dx.doi.org/10.1016/j.wneu.2017.06.164.
[107] R.L. Figueroa, Q. Zeng-Treitler, S. Kandula, L.H. Ngo, Predicting sample size required for classification performance, BMC Med. Inform. Decis. Mak. 12 (1) (2012) 1–10, http://dx.doi.org/10.1186/1472-6947-12-8.
[108] H.G. Schnack, R.S. Kahn, Detecting neuroimaging biomarkers for psychiatric disorders: Sample size matters, Front. Psychiatry 7 (2016) 1–12, http://dx.doi.org/10.3389/fpsyt.2016.00050.
[109] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, J. R. Stat. Soc. Ser. B Stat. Methodol. 67 (2) (2005) 301–320, http://dx.doi.org/10.1111/j.1467-9868.2005.00503.x.
[110] A.Z. Dag, Z. Akcam, E. Kibis, S. Simsek, D. Delen, A probabilistic data analytics methodology based on Bayesian Belief network for predicting and understanding breast cancer survival, Knowl.-Based Syst. 242 (2022) 108407, http://dx.doi.org/10.1016/j.knosys.2022.108407.
[111] E. Scosyrev, J. Messing, K. Noyes, P. Veazie, E. Messing, Surveillance Epidemiology and End Results (SEER) program and population-based research in urologic oncology: An overview, Urol. Oncol.: Semin. Orig. Investig. 30 (2) (2012) 126–132, http://dx.doi.org/10.1016/j.urolonc.2009.11.005.
[112] G. Richards, V. Rayward-Smith, P. Sönksen, S. Carey, C. Weng, Data mining for indicators of early mortality in a database of clinical records, Artif. Intell. Med. 22 (3) (2001) 215–231, http://dx.doi.org/10.1016/s0933-3657(00)00110-x.