


# Design and Development of an Intelligent System for Predicting 5-Year Survival in Gastric Cancer

Mohammad Reza Afrash<sup>1</sup>, Mostafa Shanbehzadeh<sup>2</sup> and Hadi Kazemi-Arpanahi<sup>3,4</sup>

<sup>1</sup>Department of Health Information Technology and Management, School of Allied Medical Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran. <sup>2</sup>Department of Health Information Technology, School of Paramedical, Ilam University of Medical Sciences, Ilam, Iran. <sup>3</sup>Department of Health Information Technology, Abadan University of Medical Sciences, Abadan, Iran. <sup>4</sup>Student Research Committee, Abadan University of Medical Sciences, Abadan, Iran.

Clinical Medicine Insights: Oncology  
Volume 16: 1–13  
© The Author(s) 2022  
Article reuse guidelines:  
sagepub.com/journals-permissions  
DOI: 10.1177/11795549221116833



## ABSTRACT

**BACKGROUND:** Gastric cancer remains one of the leading causes of worldwide cancer-specific deaths. Accurately predicting the survival likelihood of gastric cancer patients can inform caregivers to boost patient prognostication and choose the best possible treatment path. This study intends to develop an intelligent system based on machine learning (ML) algorithms for predicting the 5-year survival status in gastric cancer patients.

**METHODS:** A data set that includes the records of 974 gastric cancer patients retrospectively was used. First, the most important predictors were recognized using the Boruta feature selection algorithm. Five classifiers, including J48 decision tree (DT), support vector machine (SVM) with radial basic function (RBF) kernel, bootstrap aggregating (Bagging), hist gradient boosting (HGB), and adaptive boosting (Ada-Boost), were trained for predicting gastric cancer survival. The performance of the used techniques was evaluated with specificity, sensitivity, likelihood ratio, and total accuracy. Finally, the system was developed according to the best model.

**RESULTS:** The stage, position, and size of tumor were selected as the 3 top predictors for gastric cancer survival. Among the 6 selected ML algorithms, the HGB classifier with the mean accuracy, mean specificity, mean sensitivity, mean area under the curve, and mean F1-score of 88.37%, 86.24%, 89.72%, 88.11%, and 89.91%, respectively, gained the best performance.

**CONCLUSIONS:** The ML models can accurately predict the 5-year survival and potentially act as a customized recommender for decision-making in gastric cancer patients. The developed system in our study can improve the quality of treatment, patient safety, and survival rates; it may guide prescribing more personalized medicine.

**KEYWORDS:** Machine learning, data mining, survival, stomach neoplasm

**RECEIVED:** April 11, 2022. **ACCEPTED:** July 13, 2022.

**TYPE:** Original Research Article

**FUNDING:** The author(s) received no financial support for the research, authorship, and/or publication of this article.

**DECLARATION OF CONFLICTING INTERESTS:** The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

**CORRESPONDING AUTHOR:** Hadi Kazemi-Arpanahi, Department of Health Information Technology, Abadan University of Medical Sciences, Education, Research and Student Vice-Chancellor, Airport square next to Ayatollah Jami International Airport, PO Box: 517, Abadan, Khuzestan 6313833177, Iran. Email: h.kazemi@abadanums.ac.ir

## Introduction

According to global cancer statistics (GLOBOCAN) 2020, gastric cancer ranks fifth for incidence (5.6% of total new cases of cancer, 1089103 people) and fourth for mortality (7.7% of total cancer-related deaths, 768793 deaths) globally. It is the most commonly diagnosed malignancy and the chief cause of cancer-related mortality in several developing countries.<sup>1</sup> Despite the downward trend during the last decades globally, like many other Asian countries, Iran still has constantly increasing incidence and mortality rates of gastric cancer. According to GLOBOCAN 2020, gastric cancer is the second most common cancer in Iran with 13191 (11.2%) new cases of total cancer and is first with 79136 (16.4) deaths of total cancer-related deaths. This rising incidence in Iran is likely due to the recent demographic and epidemiological transitions in its population.<sup>2,3</sup>

This malignancy imposes heavy costs on the health system and patients' families. Therefore, prevention and early screening of gastric cancer should be the main priority of the country's

health system programs.<sup>4</sup> The fundamental issue in patients with gastric cancer, as in many other clinical areas, is the multi-dimensional and ambiguous nature of its diagnosis and treatment processes.<sup>5</sup> The treatment of tumors depends largely on the prognosis judgment that strongly rests on the phase, in which it is detected.<sup>6,7</sup> The 5-year relative survival rate is up to 70% for lesions in the early stages and 4% for lesions in the advanced stages.<sup>6,8,9</sup> Survival often refers to the likelihood, by which a patient will live 60 months after being diagnosed with cancer. This index is commonly used in medical science to evaluate the effects of surgical and treatment plans.<sup>10</sup> Accurately predicting the survival of gastric cancer patients could help clinicians make better decisions about the diagnosis and treatment process, including the choice of treatment methods, treatment schedule, and follow-up visits, which can increase the patients' outcomes and contain economic costs.<sup>11,12</sup> But calculating survival time in gastric cancer patients by using traditional clinical and statistical methods is faced with limitations and challenges as follows:<sup>6,13</sup>



The traditional tumor-node-metastasis (TNM) staging system has been useful in stratifying gastric cancer patients; however, mid-stage patients show a variety of prognostic outcomes and there is a critical need to categorize these patients more carefully.<sup>14</sup> Thus, the TNM system is insufficient due to the large differences in survival outcomes.<sup>15</sup> The gastric cancer treatment outcomes are related to many variables, and it is not possible to predict the survival of the disease by using one factor alone because several factors related to the disease, the patient, and the treatment process can affect the survival of cancer patients.<sup>16,17</sup> Thus, multivariate analysis tools are needed to find patterns and relationships between multiple variables simultaneously. The multivariate analysis allows to predict the effects that a change in one variable will have on other variables. Multivariate analysis can provide a more accurate picture and understanding of data behaviors which are related to each other.<sup>18,19</sup> Multivariate analysis techniques are complex and require a statistical program to perform this analysis. One of the important limitations of multivariate analysis is that it is not always easy for physicians to interpret statistical modeling outputs. In addition, a large sample of data is required to obtain meaningful results for multivariate techniques.<sup>8,20</sup> In the past, researchers have used a variety of survival analysis methods to describe the relationship between response variables and a set of independent variables in various fields of medical science. In this context, conventional survival methods such as Cox proportional hazard modeling are still the most common approach for analyzing the relative importance of the predictive variables in the development of the disease.<sup>21,22</sup> However, when using this model, some basic assumptions such as the proportionality of risks and the independence of variables affecting the risk rate must be considered.<sup>23</sup>

Technical advances in statistics and artificial intelligence (AI) enable computer engineers and health scientists to work closely to improve the prognosis using multifactorial analysis, conventional logistic regression, and Cox analysis.<sup>21,24,25</sup> The accuracy of such predictions is significantly higher than the experimental predictions. In addition, research shows that traditional statistical methods do not provide as accurate analyzes as AI. With the implementation of AI, researchers have recently developed models using AI algorithms to predict and diagnose cancer. These methods currently play an important role in increasing the accuracy of predicting cancer vulnerability, recurrence, and survival.<sup>19,24,26</sup>

Machine learning (ML), as a special concept, is a subset of AI, increasingly used in medicine. This technique is used to build predictive models to extract hidden patterns and uncover unknown correlations from massive historical data. ML has been widely used in improving the prognosis of patients.<sup>27-29</sup> Prognosis is important expertise in clinical practice, especially for physicians who make decisions in complex and ambiguous situations such as caring for cancer patients.<sup>12,30</sup> Past research has shown that ML techniques improve the accuracy of predicting cancer vulnerability, relapse, and survival, 3 facets that

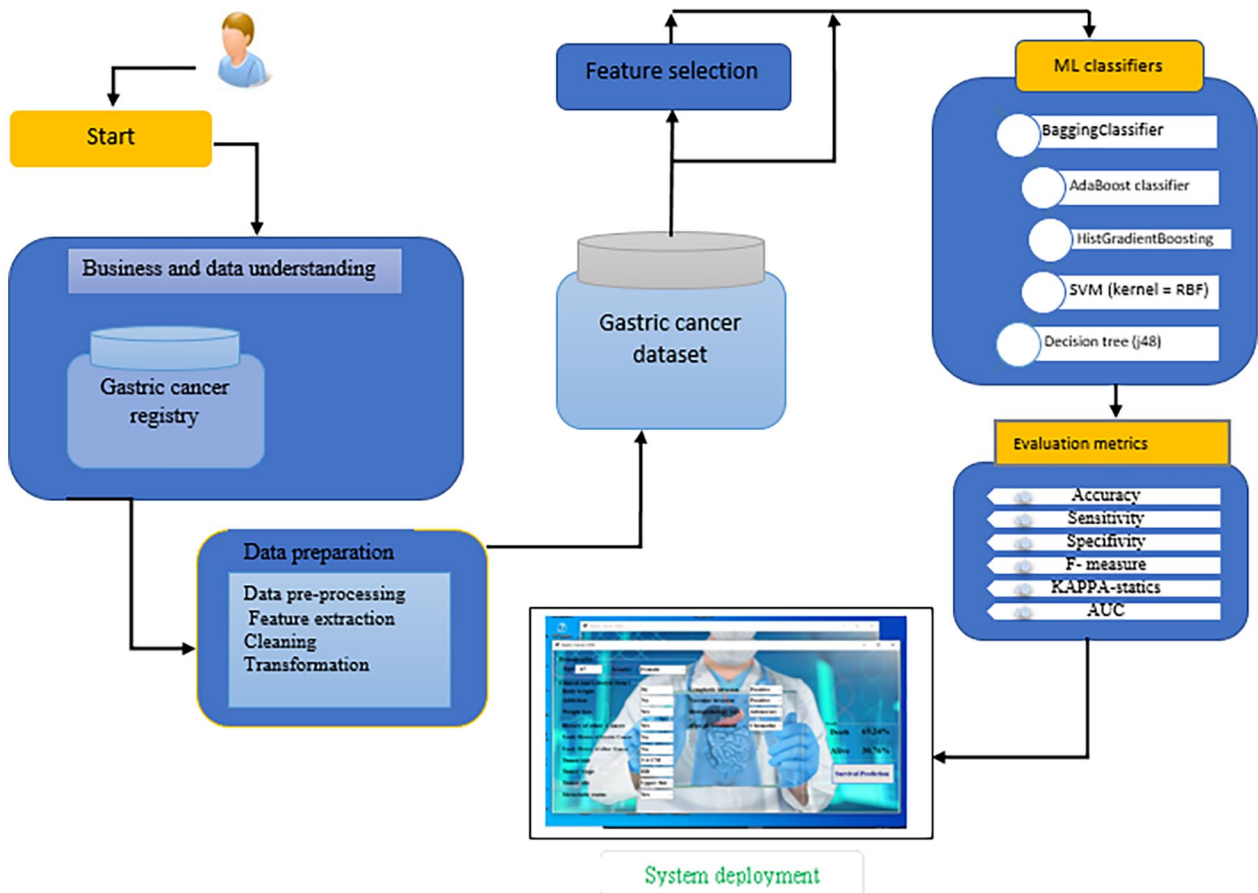
are essential for early detection and prognosis of cancer. ML can provide good results according to the clinical condition of patients.<sup>31-33</sup> By apprehending multifaceted non-linear relationships in the data, the ML technique can increase the prediction performance more than traditional statistical methods. Many studies have applied ML algorithms for predicting cancer survival. Presently, ML can predict breast cancer survivability in the primary stages.<sup>34-36</sup> Das et al<sup>37</sup> and Hauser et al<sup>38</sup> have compared selected ML methods to the survival prognosis of patients with leukemia. They have respectively found that the gradient boosting algorithms (BAs) such as hist gradient boosting (HGB) with area under the curve (AUC) of 0.779 and XGBoost with AUC of 0.87 achieve the highest performance. Okagbue et al,<sup>39</sup> Kaur et al,<sup>40</sup> and Liu et al<sup>41</sup> have assessed the performance of selected ML-based BAs to predict breast cancer survival. Finally in the reviewed studies, the AdaBoost, HGB, and XGBoost classifiers have achieved the best performance with the AUC of 98.3%, 91.1%, and 83%, respectively. Feng et al's<sup>42</sup> experimental results showed that the XGBoost method achieved the accuracy of 91.64%, recall of 91.14%, and AUC of 91.35% for neuroblastoma survival prediction.

Given the high prevalence of gastric cancer in Iran and lack of a reliable study to determine risk factors of the disease survival based on ML methods, our study aims to develop an intelligence system regarding the use of novel ML algorithms for the development and validation of gastric cancer survival prediction. The primary outcome indicator is the accuracy of the different models in predicting a 5-year (60 months or 1825 days) survival rate for gastric cancer to provide a better theoretical basis for the application of ML in survival prediction.

## Methods

### *Study design and setting*

This is a retrospective study using a data set from Ayatollah Taleghani Hospital in the southwest of Khuzestan Province, Iran. Data related to 1220 patients pathologically confirmed gastric cancer were extracted from the electronic medical record (EMR) database after obtaining appropriate approval from Research Ethical Committee, Abadan University of Medical Sciences. The study methodology complied with the cross-industry standard process for data mining (CRISP-DM). The CRISP method determined 6 phases for a data mining project including business understanding, data understanding, data preparation, modeling, evaluation, and deployment. Figure 1 represents the CRISP-DM research methodology. All the prediction models were developed using Python programming language (3.7). J48 decision tree (DT) and support vector machine (SVM) (with RBF kernel) were implemented using Python library scikit-learn (0.23.2), while bootstrap aggregating (Bagging) classifier, HGB, and adaptive boosting (AdaBoost) were implemented using another specific Python library (see Figure 1).



**Figure 1.** The framework of the machine learning method based on CRISP-DM.

AUC indicates area under the curve; CRISP-DM, cross-industry standard process for data mining; SVM, support vector machine; RBF, radial basic function.

### Data understanding

There is a large number of features collected for the patients with gastric cancer in the EMR database. So, we checked the definition of the features included in the data dictionary section of the database to completely understand the data definitions and choice of proper variables. The criteria for identifying the candidate variables related to gastric cancer for survival prediction were based on consulting with experts' oncologists and studying the related literature. Patients were only included in the study if all the following criteria were met: (1) patients who were pathologically diagnosed with gastric cancer; (2) the survival status of patients (alive/dead) was available in their records; (3) in terms of the timeframe, we considered patients diagnosed between 2010 and 2017 so as to have adequate follow-up period (5 years or more) after the diagnosis; (4) age of more than or equal to 18 years<sup>43</sup>; the patients aging under 18 years old should be included in the scope of pediatric exploration; (5) records with missing values of less than 30%.

Accordingly, from 1220 patients' records, 59 records for patients who were aged <18 years old were excluded. In the preprocessing phase, 187 incomplete rows of data (with missing data of greater than 70%) were removed. After these criteria were applied, a total of 974 patients (399 survived and 575 dead within 5 years) remained for additional analysis. Survival at 5 years was selected as the outcome variable. The following

covariates were extracted based on the literature review coupled with experts' opinions from the EMR database, as depicted in Table 1.

### Data preparation

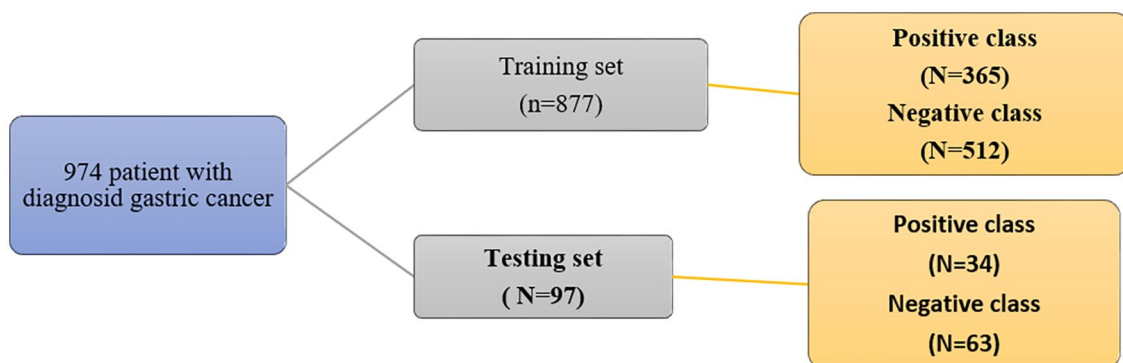
Since the raw data with missing values, noisy data, and outliers or inconsistent data will affect ML algorithms' performance, in our study to improve the performance of prediction models, the preprocessing step was made on the raw data to make it balanced, effective, and noise-free. In this phase, the attribution of missing values means and regression-based techniques were used. The rows with missing values of greater than 70% were removed. The Z-score standardization technique was applied as a data distribution-based data scaling and, for data range-based scaling, the min-max techniques were used. The data set was randomly divided into a training data set ( $n=877$ ) and a testing data set ( $n=79$ ) with the proportion of 9:1. The procedure of our study is shown in Figure 2.

### Feature extraction and feature selection

After the data set cleaning and imputation steps, we can extract relevant and important features. For this purpose, in our study, first, the previous literature was studied to extract the candidate features related to predicting survival in the patients with

**Table 1.** Characteristics of patients with gastric cancer.

NO.	FEATURE NAME	SCALE	VALUE
1	Sex	Nominal	Male—Female
2	Age at diagnosis	Interval	Ranged between 23 and 79
3	Bodyweight	Interval	>>60, <60
4	Weight loss	Nominal	Yes—No
5	Addiction	Nominal	Yes—No
6	History of other cancers	Nominal	Yes—No
7	Family history of gastric cancer	Nominal	Yes—No
8	Family history of other cancers	Nominal	Yes—No
9	Tumor size	Ordinal	<<3 CM, 3-6 CM, >6
10	Tumor stage	Ordinal	IA, IB, IIA, IIB, IIIA, IIIB, IIIC
11	Tumor location	Ordinal	Lower third, middle third, upper third, whole stomach
12	Metastatic status	Nominal	Yes—No
13	Histological type/histology	Ordinal	Rivers, diffuse, complex
14	Lymphatic invasion	Nominal	Positive, negative
15	Vascular invasion	Nominal	Positive, negative
16	Histopathology type	Ordinal	Adenocarcinoma, lymphoma, sarcoma
17	Treatment	Ordinal	Surgery, chemotherapy, surgery + chemotherapy + radiotherapy
18	Outcome class	Nominal	Survived, did not survived

**Figure 2.** Gastric cancer patient inclusion diagram (test and training set).

gastric cancer. Then, we adopted the Boruta feature selection algorithm to select the most important variables and, using these selected features, the performance of the ML algorithms was calculated. In this study, we also tested the performances of different ML predictive models for gastric cancer survival prediction on all and selected features.

#### *Development prediction models and evaluation method*

To develop the prediction model for predicting survival risk in gastric cancer patients, 5 ML algorithms, including the Bagging

classifier, AdaBoost classifier, HGB classifier, SVM (with RBF) and J48, were trained. For the development and validation of ML models, a 10-fold cross-validation method was used to train and test these models over the full and selected features. The final data set was randomly split into training (877 records, 90%) and testing (97 records, 10%) sets using methods in Scikit-learn (as shown in Figure 2). The training set is a piece of data used for model development and hyperparameter tuning (to teach ML models) and the testing set to evaluate the performance of the trained models. Data splitting prevents random data bias and ensures balanced distribution of data in training and testing sets. It is important to note that

testing set data, which was used to evaluate the performance of ML algorithms, was never used when training algorithms during the training process.

We experimentally tuned the hyperparameters over the training set based on the cross-validation method.

Once the classification algorithms were implemented over the trained data set, the next phase was to test these trained algorithms over the testing set to assess the performance of classifiers on unseen data. The performance of 5 classification models for predicting survival among gastric cancer patients was evaluated using 5 commonly used performance testing metrics including accuracy, specificity, sensitivity, AUC, and F1-score (Equations 1 to 4). Afterward, the performance of each trained classifier was compared with all other ML algorithms according to the 5 selected performance metrics. Then, the best-performing model was further applied to predict the survival of patients with gastric cancer. The performance evaluation metrics of the classifiers are listed below:

1. classification accuracy =  $\frac{TP + TN}{TP + TN + FP + FN} \times 100$
2. classification sensitivity =  $\frac{Tp}{TP + FN} \times 100$
3. classification specificity =  $\frac{TN}{TN + FP} \times 100$
4. F1-score =  $2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$

### *Ethical consideration*

Ethical Committee approved the study conducted by Abadan University of Medical Sciences (Ethics code: IR.ABADANUMS.REC.1401.003). To protect the privacy and confidentiality of the patients, we concealed the unique identification information of all the patients in the process of data collection and presentation. It adhered to the principles expressed in Declaration of Helsinki.

## **Results**

### *Characteristics of patients*

Overall, 974 patients with gastric cancer met the prespecified inclusion criteria. Of 974 eligible patients in our study, 648 (66.53%) cases were male and 326 (33.47%) cases were women and the median age of the participants was 57.25 (age of cases ranged from 23 to 79 years old). Of these, 399 (40.96%) cases survived and 575 (59.04%) dead. The detailed descriptions of all the variables are listed in Table 2.

### *Variables included in the ML models*

The variables that would be important for the prediction of the 5-year survival status of gastric cancer patients were selected from a large number of features for modeling. The Boruta algorithm was used to select important features. The

Boruta algorithm selects the most important features based on the random forest (RF) algorithm, which determines all the variables that are either potentially or faintly related to the decision features. The 8 features that were selected as the most important predictors by the Boruta algorithm and their scores and ranks are shown in Table 3.

The 8 most important features were tumor stage, tumor site, tumor size, age, metastatic status, type of treatment, lymphatic invasion, and body weight. As shown in Figure 3, tumor stage, tumor site, and tumor size obtained the highest score for the survival prediction among the patients with gastric cancer.

Moreover, between these 8 selected features, body weight and lymphatic invasion had the lowest rank for prediction of gastric cancer survival; additionally, 10 features were not selected for the survival prediction model and were deleted from the data set.

### *Results of hyperparameters tuning*

The performance of prediction models depends on the setting of the hyperparameter. In this study, to select the best model architecture, the Randomized Search CV method was used for parameter tuning and optimization models. Table 4 represents the best hyperparameters selected in this study for feeding into ML algorithms.

### *Performance of ML models*

In this experiment, we first trained 5 ML algorithms (Bagging, SVM, AdaBoost, HGB and J48 DT) over all and selected features. Afterward, we tested these trained algorithms over the testing set. The performances of 5 ML models were tested with a 10-fold cross-validation method using evaluation metrics including the mean of accuracy, sensitivity, specificity, F1-score, and area under the receiver operating characteristic (ROC). Table 5 describes the 10-fold cross-validation performance of the applied ML algorithms when using the full features data set and selected feature.

As indicated in Table 5, the Bagging classifier achieved 85.37 accuracy, 86.395% specificity, 86.54% sensitivity, 83.77% AUC, and 85.64% the F1-score value. The AdaBoost classifier had 87.322% accuracy, 83.62% specificity, 87.15% sensitivity, 86.93% AUC, and 85.15% F1-score value. The J48 DT classifier was given with prediction accuracy of 85.63%, specificity of 84.19%, sensitivity of 83.1%, AUC of 84.72%, and F1-score of 83.45%.

The HGB classifier performance for the prediction of survival among gastric cancer was 88.37% accuracy, 86.24% specificity, 89.72% sensitivity, 88.11% AUC, and 89.91% F1-score value. Finally, the SVM model with RBF kernel had 86.25% accuracy, 87% specificity, 86.43% sensitivity, 86.103% AUC, and 85.971% F1-score (Figure 4).

As indicated in Figure 4, the best ML model for predicting survival in the patients was with gastric cancer HGB classifier,

**Table 2.** The descriptive statistics of variables of the study after preprocessing.

NO.	FEATURE NAME	CLASSIFICATIONS	TOTAL	SURVIVED	DID NOT SURVIVE
				N	N
1	Age at diagnosis	<45	249	197	36
		>>45	725	483	258
2	Sex	Female	326	218	108
		Male	648	462	186
3	Body weight	<60	263	174	89
		>>60	711	506	205
4	Weight loss	Yes	369	231	138
		No	605	449	156
5	Addiction	Yes	70	27	43
		No	604	353	251
6	History of other cancers	Yes	155	74	81
		No	819	606	213
7	Family history of gastric cancer	Yes	23	7	16
		No	951	673	278
8	Family history of other cancers	Yes	62	27	35
		No	912	653	259
9	Tumor size	<<<3 CM	326	269	57
		3-6 CM	459	324	135
		>6	189	87	102
10	Tumor stage	IA	43	31	12
		IB	134	107	27
		IIA	159	127	32
		IIB	198	153	45
		IIIA	183	132	51
		IIIB	139	47	82
		IIIC	152	83	69
11	Tumor location	Lower third	315	288	27
		Middle third	340	256	84
		Upper third	284	132	152
		Whole stomach	35	4	31
12	Metastatic status	Yes	227	93	134
		No	549	437	112
		Unknown	198	150	48
14	Lymphatic invasion	Positive	542	433	209
		Negative	332	247	85

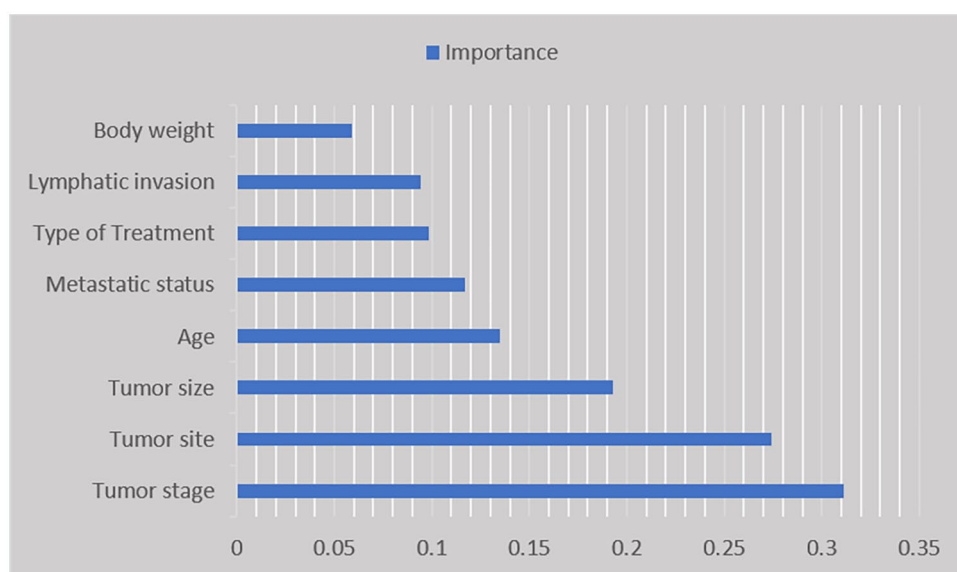
(Continued)

**Table 2.** (Continued)

NO.	FEATURE NAME	CLASSIFICATIONS	TOTAL	SURVIVED		DID NOT SURVIVE	
				N	N	N	N
15	Vascular invasion	Positive	583	364	219		
		Negative	391	316	75		
16	Histopathology type	Adenocarcinoma	670	507	163		
		Lymphoma	146	98	48		
		Sarcoma	158	75	83		
17	Type of treatment	Surgery	192	75	117		
		Chemotherapy	366	292	74		
		Surgery + chemotherapy + radiotherapy	416	313	103		
18	Class	Survived	974	399	575		
		Did not survive					

**Table 3.** The most important selected variables of survival prediction.

NO.	FEATURE NAME	IMPORTANCE
1	Tumor stage	0.311
2	Tumor site	0.274
3	Tumor size	0.193
4	Age	0.135
5	Metastatic status	0.117
6	Type of treatment	0.098
7	Lymphatic invasion	0.0941
8	Body weight	0.059



**Figure 3.** The most important predictors of survival among patients with gastric cancer.

**Table 4.** Best hyperparameters selected for machine learning algorithms.

NO.	ML MODELS	HYPERPARAMETERS	F-SCORE
1	Bagging classifier	("verbose": 2, "random_state": 933, "n_estimators": 8, "max_samples": 0.4, "bootstrap": "true")	64.21
2	SVM (kernel=RBF)	C=10, G=0.001	59.70
3	AdaBoost classifier	("random_state": 933, "n_estimators": 87, "learning_rate": 0.1, "algorithm": "samme.R")	55.18
4	Hist gradient boosting	("verbose": 2, "random_state": 93, "n_estimators": 8, "max_leaf_nodes": 58, "max_iter": 130, "max_depth": 8, "learning_rate": 0.1)	53.95
5	Decision tree (j48)	("random_state": 93, "min_sample_splits": 8, "min_sample_leaf": 1, "max_features": "lpg2, "criteria": "Gini")	51.037

Abbreviations: ML, machine learning; SVM, support vector machine; RBF, radial basic function.

with mean accuracy value, mean specificity value, mean sensitivity value, mean AUC value, and mean F1-score value of 88.37%, 86.24%, 89.72%, 88.11%, and 89.91%, respectively. Figure 5 depicted the classification report matrix and AUC curve of the HGB model which was selected as the best prediction model in terms of the highest performance metrics. The AdaBoost classifier was the second-best classifier that had the accuracy of 87.322%. The worst ML model's performance was observed for the Bagging classifier out of 5 prediction models in terms of the average accuracy, sensitivity, specificity, AUC, and F1-measure.

### System development

Using the best-performing ML model developed from among the 5 models, a windows-based clinical decision support system (CDSS) was designed and implemented between August 2021 and December 2021. The user interface of the gastric cancer survival prediction system was developed by C# programming language. To help medical oncologists' decision-making and to predict the survival among the patients with gastric cancer, the CDSS was installed at Ayatollah Taleghani Hospital of Abadan city, Iran. Screenshots of the developed CDSS are shown in Figure 6.

### Discussion

Accurate evaluation of the gastric cancer prognosis is of great value in understanding the disease and providing effective treatment for each patient. In the last few decades, the TNM grading system has been the most accepted and used global gastric cancer classification system in the anatomic extent of disease. However, the TNM gastric cancer grading system has led to a substantial difference in the survival of patients with the same tumor stage and similar survival results between distinctive steps.<sup>44,45</sup> Presently, TNM staging cannot still meet the individual and precise treatment of patients' requirements in the health center. The TNM staging system is inherently limited, with large survival variations for same-stage tumors and low accuracy in determining a patient-specific prognosis. Relevant literature has revealed that the recital of making a

prognostic model by Cox proportional hazards model<sup>45</sup> and SVM<sup>46</sup> is significantly better than the TNM staging system. However, determining more illustrative variables for precise prediction of prognosis is a crucial problem that needs to be addressed. ML algorithms can be a good alternative for solving this problem. In the present work, the selected ML models were evaluated to predict future gastric cancer survival. Then, a CDSS was developed based on the best model.

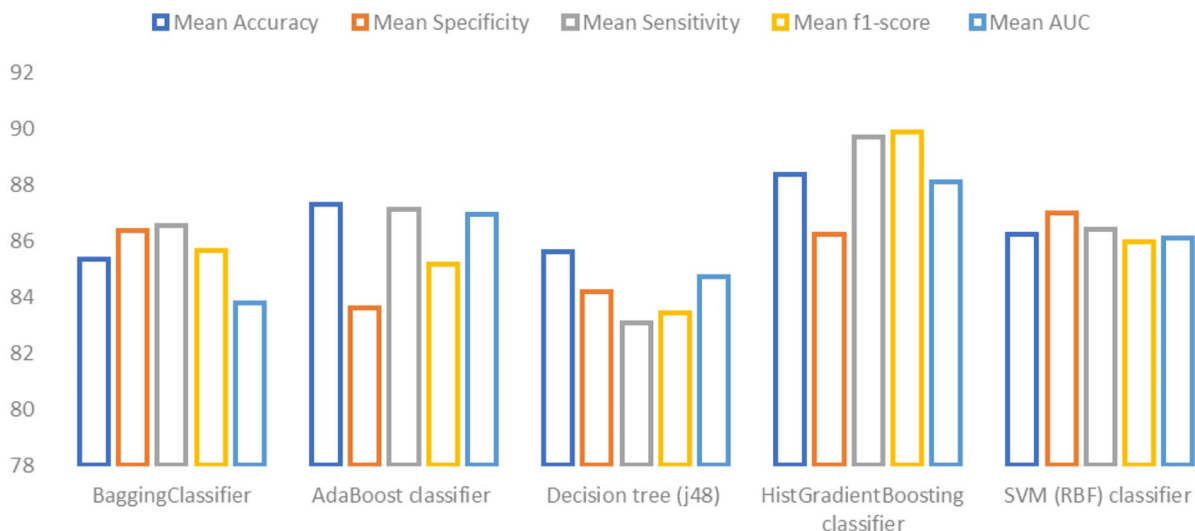
So far, several studies have been conducted to compare ML techniques and design optimal and efficient CDSSs for the survival prognosis of the patients with gastric cancer. Liu et al<sup>8</sup> used ML methods in the survival prediction of gastric cancer. Out of 6 models, the light gradients boosting machine (GBM) had the best accuracy and the highest precision rate for survivability analysis. By implementing 6 ML models, Akcay et al<sup>6</sup> concluded that XGBoost with 86% accuracy (95% confidence interval, 0.74-0.97, AUC: 0.86) along with RF is the most successful algorithms for gastric cancer survival and recurrence prediction. Similarly in Bang's study,<sup>35</sup> among the 18 ML models, the XBoost classifier showed the best performance in early gastric cancer prediction and survivability with the accuracy of 93.4%, precision of 92.6%, recall of 99.0%, and F1 score of 95.7%. Fan et al<sup>36</sup> retrospectively compared 3 ML techniques for the prediction of metastatic, relapse, and patient survival chances in the early stage of gastric cancer. In their study, the AdaBoost model achieved better performance with the AUC of 0.849. Accordingly, Lee et al<sup>47</sup> applied 7 ML methods for a 2-year survival analysis of patients with gastric cancer. They found that the gradient Boosting algorithm (GBA) with the AUC of 0.80 gained the highest performance. In addition, Gao et al<sup>48</sup> implemented the selected ML models for gastric cancer recurrence and survival prediction. Their results showed that the GBA would present optimum performance. Chen et al<sup>49</sup> proposed a gradient-boosting decision tree (GBDT)-based prediction method for projecting the GC clinical deterioration and survival chance. Ultimately, the proposed model attained appropriate performance with 0.89% of AUC. Mirniaharikandehi et al<sup>50</sup> compared 5 gradients boosting machine (GBM) model performance for predicting gastric cancer metastatic risk and patient survivability. The results



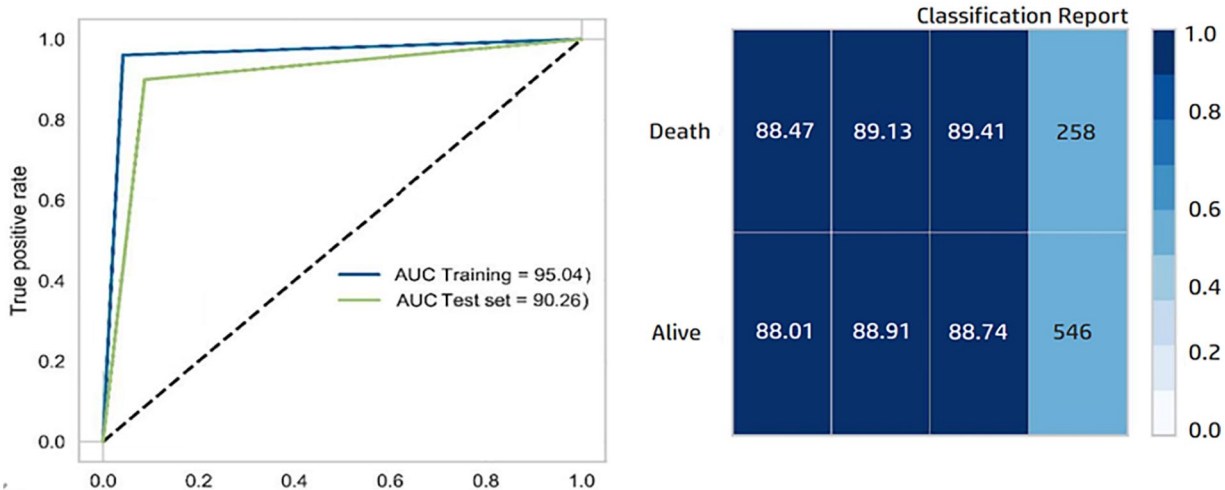
**Table 5.** Overall predictive performance for each ML model to predict survival of gastric cancer.

	BAGGING CLASSIFIER		ADABOOST CLASSIFIER		DECISION TREE (J48)		HIST GRADIENT BOOSTING CLASSIFIER		SVM (RBF) CLASSIFIER	
	FULL FEATURE	SELECTED FEATURE	FULL FEATURE	SELECTED FEATURE	FULL FEATURE	SELECTED FEATURE	FULL FEATURE	SELECTED FEATURE	FULL FEATURE	SELECTED FEATURE
Mean accuracy	81.05	85.37	76.019	87.322	74	85.63	84.10	89.37	76.91	86.25
95% CI	(0.801, 0.827)	(0.839, 0.862)	(0.75, 0.771)	(0.86, 0.882)	(0.73, 0.751)	(0.847, 0.871)	(0.83, 0.851)	(0.881, 0.91)	(0.75, 0.78)	(0.857, 0.88)
STD	0.0114	0.018	0.01874	0.015	0.019	0.0108	0.014	0.0487	0.00974	0.0141
Mean specificity	78.320	86.395	77.20	83.62	73.58	84.19	78.31	87.24	71.9	87
95% CI	(0.779, 0.792)	(0.85, 0.88)	(0.759, 0.786)	(0.827, 0.841)	(0.72, 0.74)	(0.83, 0.8502)	(0.771, 0.792)	(0.86, 0.891)	(0.70, 0.74)	(0.86, 0.889)
STD	0.012	0.016	0.017	0.0126	0.0191	0.014	0.0132	0.0280	0.0223	0.01496
Mean sensitivity	79.065	86.54	74.31	87.15	76.02	83.1	80.46	89.841	72.68	86.43
95% CI	(0.784, 0.81)	(0.85, 0.88)	(0.731, 0.765)	(0.86, 0.882)	(0.751, 0.774)	(0.82, 0.841)	(0.79, 0.82)	(0.89, 0.913)	(0.718, 0.745)	(0.85, 0.872)
STD	0.0174	0.0210	0.019	0.019	0.0158	0.0109	0.0160	0.0141	0.0235	0.0188
Mean AUC	80.14	83.77	75.16	86.93	73.04	84.72	81.93	88.11	72.3011	86.103
95% CI	(0.79, 0.816)	(0.82, 0.85)	(0.740, 0.763)	(0.851, 0.883)	(0.72, 0.751)	(0.83, 0.861)	(0.80, 0.831)	(0.87, 0.891)	(0.71, 0.74)	(0.841, 0.881)
STD	0.0154	0.0107	0.0291	0.0179	0.0186	0.0291	0.0217	0.0194	0.019	0.0209
MeanF1-score	79.30	85.64	74.801	85.15	71.9	83.45	83.441	89.91	74.08	85.971
95% CI	(0.781, 0.810)	(0.84, 0.87)	(0.73, 0.76)	(0.838, 0.87)	(0.71, 0.72)	(0.82, 0.849)	(0.82, 0.91)	(0.88, 0.91)	(0.73, 0.75)	(0.84, 0.871)
STD	0.0128	0.0293	0.01075	0.01936	0.0105	0.0138	0.0125	0.0164	0.0194	0.0172

Abbreviations: AUC, area under the curve; CI, confidence interval; ML, machine learning; SVM, support vector machine; RBF, radial basic function; STD, standard deviation.



**Figure 4.** Comparing machine learning models' performance on selected features. AUC indicates area under the curve; SVM, support vector machine; RBF, radial basis function.



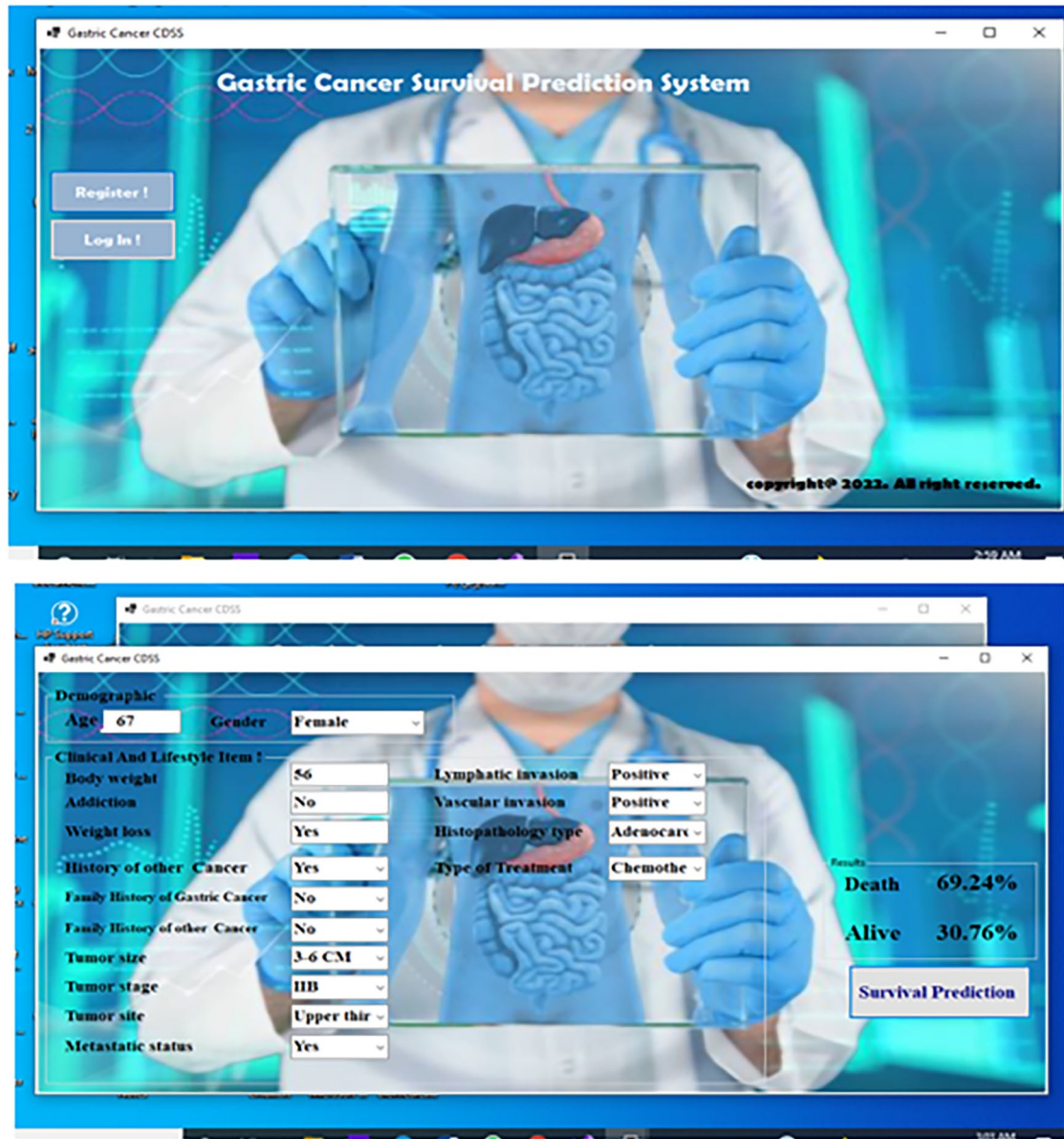
**Figure 5.** AUC curve and classification report for hist gradient boosting classifier. AUC indicates area under the curve.

showed GBM technique combined with a random projection algorithm yielded significantly higher prediction performance (accuracy = 71.2%).

Many clinical predictors influence gastric cancer. In the reviewed studies, after doing feature ranking, the variables such as age,<sup>6,35,48,49</sup> gender,<sup>36,47,50</sup> body mass index,<sup>6,47,50</sup> Karnofsky performance scale,<sup>8,48,51</sup> TNM stage,<sup>36,47-50</sup> tumor grade,<sup>7,8,35,47-50</sup> tumor size,<sup>6,7,47,49-51</sup> tumor location,<sup>6,7,35,36,48,49</sup> lymphovascular invasion,<sup>7,8,47,49,50</sup> active and timely treatment,<sup>7,8,36</sup> type of treatment,<sup>35,49</sup> disease stage and severity,<sup>6,8,35,36,48,49</sup> and weight loss<sup>36,47,49</sup> were determined as the most important risk factors affecting gastric cancer survival outcome. Similarly, in our study, feature selection analysis was performed to rank the important set of variables. Among 17 primary variables, 8 variables including tumor stage, tumor site, tumor size, age, metastatic status, type of treatment, lymphatic invasion, and body weight were ultimately selected as the most important variables. These variables were used as input to construct ML models. After

implementing the selected classifiers, the HGB with 88.37% accuracy, 86.24% specificity, 89.72% sensitivity, 88.11% AUC, and 89.91% F1 score achieved the highest performance in the survival prognosis of gastric cancer patients.

It is proven that ML technologies will improve health care quality and, consequently, reduce the serious complications and deaths associated with gastric cancer. The developed models in our study can help to better adhere to the best treatment standards. Such models may assist in early and effective diagnosis and accurate survival prediction of gastric cancer cases. Early detection of gastric cancer and active patient triaging help to evade the advanced stages of the disease and increase survival chances. This requirement is more important since numerous risk factors are involved in gastric cancer emergence and development. Therefore, in the present study, initially, the most important effective variables in the survival and the prognosis of patients with gastric cancer were identified using Boruta feature selection.



**Figure 6.** Screenshots of CDSS for prediction survival among patients with gastric cancer. CDSS indicates clinical decision support system.

However, the present study faced several potential limitations and challenges that need to be addressed. These challenges may negatively affect the quality of modeling. The most important limitations in the present study were (1) single-center and small size of the selected data set, (2) retrospective data collection nature and the existence of missing fields and noise, (3) the selected data set lacks some important variables such as history and lifestyle, and (4) we did not use external validation to evaluate the proposed model. Therefore, to improve the quality of modeling and reduce prejudice in future research, more ML algorithms with further variables on multicenter and larger databases should be trained. In addition, it is suggested that the present study be conducted as a prospective to follow-up on the 5-year status of patients and use more external validations to further validate our findings.

## Conclusions

Using ML techniques, accurate models can be made based on appropriate algorithms that can guide patient care and treatment, and increase workflow efficiency based on the available big data. Using ML techniques to predict survival in gastric cancer patients is an important opportunity to further improve decision support systems and provide the objective assessment of the comparative benefits of different types of treatment options for each case by determining factors using ML algorithms. The possibility of personalizing the treatment of patients is provided. Further ML studies with a larger number of patients are needed to determine the optimum algorithm and support the decision-making process for personalized treatment.

## Acknowledgements

We thank the research deputy of the Abadan University of medical sciences for financially supporting this project. Also, we thank all patients who freely participated in this study.

## Author Contributions

**MRA, HKA:** Conceptualization; Data curation; Formal analysis; Investigation; Software; Roles/Writing – original draft.

**HKA, MSH:** Conceptualization; Formal analysis; Investigation; Roles/Writing – original draft; Funding acquisition; Methodology; Project administration; Resources; Supervision; Writing – review & editing.

**MRA, MSH, HKA:** Conceptualization; Investigation; Methodology; Validation; Writing – review & editing.

## Availability of Data and Materials

The data sets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

## Ethics Approval and Consent to Participate

This study is the result of a research project approved by the Research Committee at Abadan University of Medical Sciences (Iran) (ethic code number: IR.ABADANUMS.REC.1401.003). All methods of the present study were performed following the relevant guidelines and regulations of the ethical committee of the Abadan University of Medical Sciences. Participation was voluntary, the consent was verbal, but all participants responded via email or text message to approve their participation. Participants had the right to withdraw from the study at any time without prejudice.

## Consent for Publication

Not applicable.

## ORCID iD

Mohammad Reza Afrash  <https://orcid.org/0000-0001-9571-2112>

## REFERENCES

- Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin.* 2021;71:209-249.
- Moradian F, Fararouei M, Karami M, et al. Trend of geographical distribution of stomach cancer in Iran from 2004 to 2014. *BMC Gastroenterol.* 2022;22:4.
- Abbaszadegan MR, Mojarrad M, Rahimi HR, Moghbeli M. Genetic and molecular biology of gastric cancer among Iranian patients: an update. *Egypt J Med Hum Genet.* 2022;23:17.
- Thapa S, Fischbach LA, Delongchamp R, Faramawi MF, Orloff MS. Using machine learning to predict progression in the gastric precancerous process in a population from a developing country who underwent a gastroscopy for dyspeptic symptoms. *Gastroenterol Res Pract.* 2019;2019:8321942.
- Sapra R, Mehrotra S, Nundy S. Artificial neural networks: prediction of mortality/survival in gastroenterology. *Curr Med Res Pract.* 2015;5:119-129.
- Akçay M, Etiz D, Celik O. Prediction of survival and recurrence patterns by machine learning in gastric cancer cases undergoing radiation therapy and chemotherapy. *Adv Radiat Oncol.* 2020;5:1179-1187.
- Gao W, Tang X, Yi H, et al. Mesoporous molecular sieve-based materials for catalytic oxidation of VOC: a review. *J Environ Sci (China).* 2023;125:112-134.
- Liu P, Li L, Yu C, Fei S (eds). Two staged prediction of gastric cancer patient's survival via machine learning techniques. *Paper presented at: Proceedings of the 7th International Conference on Artificial Intelligence Applications*; 2020. <https://aircconline.com/csit/papers/vol10/csit100308.pdf>.
- Wohnrath DR, Araujo RLC. Positive node-ratio in curative-intent treatment for gastric cancer is a strong independent prognostic factor for 5-year overall survival. *J Surg Oncol.* 2020;121:777-783.
- Wang Y, Wang D, Ye X, Wang Y, Yin Y, Jin Y. A tree ensemble-based two-stage model for advanced-stage colorectal cancer survival prediction. *Inform Sci.* 2019;474:106-124.
- Jiang Y, Jin C, Yu H, et al. Development and validation of a deep learning CT signature to predict survival and chemotherapy benefit in gastric cancer: a multicenter, retrospective study. *Ann Surg.* 2021;274:e1153-e1161.
- Glare P, Sinclair C, Downing M, Stone P, Maltoni M, Vigano A. Predicting survival in patients with advanced disease. *Eur J Cancer.* 2008;44:1146-1156.
- Li Z, Li S, Ying X, et al. The clinical value and usage of inflammatory and nutritional markers in survival prediction for gastric cancer patients with neoadjuvant chemotherapy and D2 lymphadenectomy. *Gastric Cancer.* 2020;23:540-549.
- Oh SE, Seo SW, Choi MG, Sohn TS, Bae JM, Kim S. Prediction of overall survival and novel classification of patients with gastric cancer using the survival recurrent network. *Ann Surg Oncol.* 2018;25:1153-1159.
- Lei K-F, Liu B-Y, Zhang X-Q, et al. Development of a survival prediction model for gastric cancer using serine proteases and their inhibitors. *Exp Ther Med.* 2012;3:109-116.
- Dalhammar K, Malmström M, Schelin M, Falkenback D, Kristensson J. The impact of initial treatment strategy and survival time on quality of end-of-life care among patients with oesophageal and gastric cancer: a population-based cohort study. *PLoS ONE.* 2020;15:e0235045.
- Claassen YHM, Bastiaannet E, Hartgrink HH, et al. International comparison of treatment strategy and survival in metastatic gastric cancer. *BJS Open.* 2019;3:56-61.
- Ma W, Bai H, Li F. Study on the evaluation of survival time of patients with gastric cancer and the construction of systemic inflammatory markers score. *Chin J Postgrad Med.* 2020;36:973-979.
- Kangi AK, Bahrapour A. Predicting the survival of gastric cancer patients using artificial and Bayesian neural networks. *Asian Pac J Cancer Prev.* 2018;19:487.
- Umut K, Yaprak G, Yilmaz A, Alaattin Ö. Is there any advantage of machine learning to multivariate regression analysis for predicting disease-related deaths in patients with gastric cancer? Reevaluation of retrospective data. *Turk J Oncol.* 2021;36:184-190.
- Safari M, Abbasi M, Gohari Ensaf F, Berangi Z, Roshanaei G. Identification of factors affecting metastatic gastric cancer patients' survival using the random survival forest and comparison with Cox regression model. *Iran J Epidemiol.* 2020;15:343-351.
- Zare A, Hosseini M, Mahmoodi M, Mohammad K, Zeraati H, Holakouie Naineni K. A comparison between accelerated failure-time and Cox proportional hazard models in analyzing the survival of gastric cancer patients. *Iran J Public Health.* 2015;44:1095-1102.
- Biglarian A, Hajizadeh E, Kazemnejad A, Zali M. Application of artificial neural network in predicting the survival rate of gastric cancer patients. *Iran J Public Health.* 2011;40:80-86.
- Zhang Z, He T, Huang L, Li J, Wang P. Immune gene prognostic signature for disease free survival of gastric cancer: translational research of an artificial intelligence survival predictive system. *Comput Struct Biotechnol J.* 2021;19:2329-2346.
- Yousefi Z, Goshayeshi L, Miri HH, et al. Survival rate of patients with gastric cancer and effective factors in two referral hospitals of Mashhad, Iran: a multiple regression analysis. *Electron Physician.* 2021;13:7831-7837.
- Huang S, Yang J, Fong S, Zhao Q. Artificial intelligence in cancer diagnosis and prognosis: opportunities and challenges. *Cancer Lett.* 2020;471:61-71.
- Deo RC. Machine learning in medicine. *Circulation.* 2015;132:1920-1930.
- Jha S, Topol EJ. Adapting to artificial intelligence: radiologists and pathologists as information specialists. *JAMA.* 2016;316:2353-2354.
- Wong D, Yip S. Machine learning classifies cancer. *Nature* 2018;555:446-447.
- Simmons CP, McMillan DC, McWilliams K, et al. Prognostic tools in patients with advanced cancer: a systematic review. *J Pain Symptom Manage.* 2017;53:962-970.e10.
- Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. *Comput Struct Biotechnol J.* 2015;13:8-17.
- Zolbanin HM, Delen D, Zadeh AH. Predicting overall survivability in comorbidity of cancers: a data mining approach. *Decis Support Syst.* 2015;74:150-161.
- Chen D, Xing K, Henson D, Sheng L, Schwartz AM, Cheng X. Developing prognostic systems of cancer patients by ensemble clustering. *J Biomed Biotechnol.* 2009;2009:632786.

34. Zhou C, Hu J, Wang Y, et al. A machine learning-based predictor for the identification of the recurrence of patients with gastric cancer after operation. *Sci Rep.* 2021;11:1571.
35. Bang CS, Ahn JY, Kim J-H, Kim Y-I, Choi IJ, Shin WG. Establishing machine learning models to predict curative resection in early gastric cancer with undifferentiated histology: development and usability study. *J Med Internet Res.* 2021;23:e25053.
36. Fan L, Li J, Zhang H, et al. Machine learning analysis for the noninvasive prediction of lymphovascular invasion in gastric cancer using PET/CT and enhanced CT-based radiomics and clinical variables. *Abdom Radiol (NY).* 2022;47:1209-1222.
37. Das AK, Mishra S, Mishra DK, Gopalan SS. Machine learning to predict 5-year survival among pediatric acute myeloid leukemia patients and development of OSPAM-C online survival prediction tool. *medRxiv* 2020. <https://www.medrxiv.org/content/10.1101/2020.04.16.20068221v2>.
38. Hauser RG, Esserman D, Beste LA, et al. A machine learning model to successfully predict future diagnosis of chronic myelogenous leukemia with retrospective electronic health records data. *Am J Clin Pathol.* 2021;156:1142-1148.
39. Okagbue HI, Adamu PI, Oguntunde PE, Obasi E, Odetunmbi OA. Machine learning prediction of breast cancer survival using age, sex, length of stay, mode of diagnosis and location of cancer. *Health Technol.* 2021;11:887-893.
40. Kaur P, Singh A, Chana I. BSense: a parallel Bayesian hyperparameter optimized Stacked ensemble model for breast cancer survival prediction. *J Comput Sci.* 2022;60:101570.
41. Liu P, Fu B, Yang SX, Deng L, Zhong X, Zheng H. Optimizing survival analysis of XGBoost for ties to predict disease progression of breast cancer. *IEEE Trans Biomed Eng.* 2020;68:148-160.
42. Feng Y, Wang X, Zhang J. A heterogeneous ensemble learning method for neuroblastoma survival prediction. *IEEE J Biomed Health Inform.* 2021;26:1472-1483.
43. Zhu S-L, Dong J, Zhang C, Huang YB, Pan W. Application of machine learning in the diagnosis of gastric cancer based on noninvasive characteristics. *PLoS ONE.* 2020;15:e0244869.
44. Dikken JL, van de Velde CJ, Gönen M, Verheij M, Brennan MF, Coit DG. The New American Joint Committee on Cancer/International Union Against Cancer staging system for adenocarcinoma of the stomach: increased complexity without clear improvement in predictive accuracy. *Ann Surg Oncol.* 2012;19:2443-2451.
45. Woo Y, Son T, Song K, et al. A novel prediction model of prognosis after gastrectomy for gastric carcinoma. *Ann Surg.* 2016;264:114-120.
46. Jiang Y, Xie J, Han Z, et al. Immunomarker support vector machine classifier for prediction of gastric cancer survival and adjuvant chemotherapeutic benefit. *Clin Cancer Res.* 2018;24:5574-5584.
47. Lee E, Jung SY, Hwang HJ, Jung J. Patient-level cancer prediction models from a nationwide patient cohort: model development and validation. *JMIR Med Inform.* 2021;9:e29807.
48. Gao Y, Wang H, Guo M, Li Y (eds). An adaptive machine learning pipeline for predicting the recurrence of gastric cancer. Paper presented at: 2020 5th international conference on information science, computer technology and transportation (ISCTT); November 13-15, 2020; Shenyang, China. New York, NY: IEEE.
49. Chen Q, Zhang J, Bao B, Zhang F, Zhou J. Large-scale gastric cancer susceptibility gene identification based on gradient boosting decision tree. *Front Mol Biosci.* 2021;8:815243.
50. Mirniaharikandehi S, Heidari M, Danala G, Lakshmirarahan S, Zheng B. Applying a random projection algorithm to optimize machine learning model for predicting peritoneal metastasis in gastric cancer patients using CT images. *Comput Methods Programs Biomed.* 2021;200:105937.
51. Connolly NP, Simpkin A, Mylotte D, et al. Impact on percutaneous coronary intervention for acute coronary syndromes during the COVID-19 outbreak in a non-overwhelmed European healthcare system: COVID-19 ACS-PCI experience in Ireland. *BMJ Open.* 2021;11:e045590.