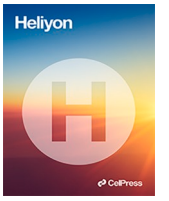
Heliyon 9 (2023) e22654

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Prediction of five-year survival among esophageal cancer patients using machine learning

Raoof Nopour

Department of Health Information Management, Student Research Committee, School of Health Management and Information Sciences Branch, Iran University of Medical Sciences, Tehran, Iran

A R T I C L E I N F O

Keywords:
Esophageal cancer
Machine learning
Survival
Prediction model
Public health challenges

A B S T R A C T

Background and aim: Considering the silent progression of esophageal cancer, the survival prediction of this disease is crucial in enhancing the quality of life of these patients globally. So far, no prediction solution has been introduced for the survival of EC in Iran based on the machine learning approach. So, this study aims to develop a prediction model for the five-year survival of EC based on the ML approach to promote clinical outcomes and various treatment and preventive plans.

Material and methods: In this retrospective study, we investigated the 1656 cases of survived and non-survived EC patients belonging to Imam Khomeini Hospital in Sari City from 2013 to 2020. The multivariable regression analysis was used to select the best predictors of five-year survival. We leveraged random forest, eXtreme Gradient Boosting, support vector machine, artificial neural networks, Bayesian networks, J-48 decision tree, and K-nearest neighborhood to develop the prediction models. To get the best model for predicting the five-year survival of EC, we compared them using the area under the receiver operator characteristics.

Results: The age at diagnosis, body mass index, smoking, obstruction, dysphagia, weight loss, lymphadenopathy, chemotherapy, radiotherapy, family history of EC, tumor stage, type of appearance, histological type, grade of differentiation, tumor location, tumor size, lymphatic invasion, vascular invasion, and platelet albumin ratio were considered as the best predictors associated with the five-year survival of EC based on the regression analysis. In this respect, the random forest with the area under the receiver operator characteristics of 0.95 was identified as a superior model.

Conclusion: The experimental results of the current study showed that the random forest could have a significant role in enhancing the quality of care in EC patients by increasing the effectiveness of follow-up and treatment measures introduced by care providers.

1. Introduction

Cancer is multi-factorial and sophisticated, challenging public health by its growth trends [1]. Esophageal cancer (EC) refers to the emergence and growth of cancerous masses in the upper, middle, and lower regions of the esophagus that histologically have adenocarcinoma and squamous cell carcinoma types [2]. EC has a poor prognosis, and most EC patients are diagnosed at advanced stages, usually with dysphagia and weight loss [3]. The pathology of EC is hardly comprehended compared to other cancers, requiring

E-mail address: raoof.n1370@gmail.com.

https://doi.org/10.1016/j.heliyon.2023.e22654
Received 12 October 2023; Received in revised form 16 November 2023; Accepted 16 November 2023
Available online 29 November 2023

Table 1
The characteristics of survived and non-survived EC patients.

Feature	Value	Total	Non-Survived	Survived	Skewness	Kurtosis
		n	n	n		
Age at diagnosis (years)	<55	676	472	204	0.895	−1.752
	≥55	929	744	185		
Gender	Male	1070	776	294	1.125	0.526
	Female	535	440	95		
Education	Illiterate	1225	1012	213	−2.267	6.824
	Literate	380	204	176		
Place of residence	Rural	955	635	320	−1.012	0.437
	Urban	650	581	69		
Income	Low	731	542	189	−0.724	−1.484
	Medium	554	470	84		
	High	320	204	116		
BMI	<18.5	320	217	103	−0.321	1.591
	18.5–25	570	401	169		
	25–30	421	386	35		
	>30	294	212	82		
Smoking	Yes	839	655	184	0.153	−2.385
	No	766	561	205		
Alcohol	Yes	131	91	40	−2.695	−5.626
	No	1474	1125	349		
Obstruction	Yes	970	638	332	0.947	−1.526
	No	635	578	57		
Dysphasia	Yes	1175	964	211	1.284	0.523
	No	430	252	178		
Weight loss	Yes	916	806	110	0.795	−0.126
	No	689	410	279		
Lymphadenopathy	Yes	604	417	187	−1.064	0.155
	No	1001	799	202		
Chemotherapy	Yes	1379	1101	278	2.497	−5.854
	No	226	115	111		
Open surgery	Yes	1475	1159	316	2.639	−6.924
	No	130	57	73		
Radiotherapy	Yes	988	809	179	1.026	−0.138
	No	617	407	210		
Family history of esophageal cancer	Yes	549	378	171	−1.154	0.227
	No	1056	838	218		
Tumor stage	II	480	317	163	−0.196	1.775
	III	709	629	80		
	IV	416	270	146		
Type of appearance	Fungating,	314	194	120	−0.226	1.694
	Ulcerating,	749	587	162		
	Fungating and ulcerating,	241	181	60		
	Without any Fungating and ulcerating	301	254	47		
Histological type	Squamous cell carcinoma	1455	1126	329	−2.602	0.155
	Adenocarcinoma	150	90	60		
Grade of differentiation	Well	275	187	88	−0.389	1.524
	Moderate	805	659	146		
	Poor	525	370	155		
Tumor location	Overlapping	540	376	164	−0.317	1.575
	Lower thoracic	145	82	63		
	Middle thoracic	426	357	69		
	Upper thoracic	494	401	93		
Tumor size	T1	165	126	39	−0.618	1.321
	T2	712	556	156		
	T3	506	357	149		
	T4	222	177	45		
Lymphatic invasion	Yes	632	494	138	−1.018	0.236
	No	973	722	251		
Vascular invasion	Yes	1182	964	218	1.308	0.484
	No	423	252	171		
Platelet albumin ratio	<4000	238	179	59	−0.194	1.790
	4000–6000	630	489	141		
	6000–8000	503	381	122		
	>8000	234	167	67		

more precise ways to diagnose and manage this disease [4]. Another sophisticated characteristic of the EC is geographic location-based variability and socioeconomic conditions globally [5].

According to presented reports from GLOBOCAN, 604,100 new cases and 544,000 deaths from EC existed worldwide in 2020 [6]. This malignancy has the seventh and sixth rank among cancers concerning morbidity and mortality, respectively [7]. It is shown that there has been an increasing trend regarding EC since 2020 [8]. By this fixed increasing trend, it is estimated that there will be more than 900,000 new cases and 800,000 mortality rate by 2040 associated with EC [6]. The incidence of this cancer in men is two to three times higher than in women [5]. Also, increasing the age would heighten the probability of getting EC [9]. The highest incidence of EC rate is obtained from African nations, India, China, and Iran, which imposed too much economic burden on them and increased the attention to enhance preventive strategies for EC [10].

Iran is among the most hazardous countries concerning EC [11]. The EC rate in Iran is higher than the average incidence in both genders globally. This country ranks fifth and eighth concerning EC with an age-standardized rate of 0.88 and 6.15 per 100,000 person-years in men and women, respectively [12]. As a large country with a variable cancer prevalence, gastrointestinal cancer, especially the esophageal type, is more prevalent in the northern regions of Iran [13,14].

Considering the sophisticated nature of EC, it is not uncommon for the five-year survival rate of EC to be less than 25% worldwide and less than other gastrointestinal cancer types [15]. EC has a variable five-year survival rate in different points of the world; for example, 20% in the USA, 15% in the UK, and 11% in China. This disease imposes public health concerns globally due to the variable nature concerning location, as mentioned previously [16]. The five-year survival rate of EC in Iran is estimated at roughly 11.3%, indicating a relatively lower rate than in developed countries [17]. According to the analysis of the cancer national surveillance on cancer survival in Iran, it is recommended to implement an appropriate and timely detection program and increase the quality of care to increase the survival of cancer patients [18].

The Tumor, Nodes, and Metastasis (TNM) staging system, generated by extensive multi-center studies covering many patients, is used for classifying cancer patients [19]. As a classification system for malignancy, it can be leveraged for cancer prognosis staging and assess the malignancy based on the tumor size, regional lymph node involvement, and metastasis conditions [20]. Although the TNM staging system is used for screening the prognosis of EC patients, it is not considered a comprehensive approach for accurate prediction due to the limited features used in this system [21]. Also, considering the low five-year survival rate of EC and the heterogeneity of these cancer cases concerning pathological characteristics and age, enhancing the survival rate and saving EC patients at advanced stages is a global challenge. Considering this complexity, an efficient strategy for early detection of this disease is crucial [22].

The machine learning (ML) approach as a sub-field of artificial intelligence (AI) has potential medical advantages, including efficient diagnosis, risk stratification, and therapy [23]. It also covers suitable preventive, predictive, and therapy strategies for various medical conditions and diseases [24]. So far, AI strategies such as ML methods have played a crucial role in predicting the cancer survival rate and recurrence [25]. Despite leveraging the deep learning (DL) techniques in recent studies, especially in image data and high volume of data, the ML techniques are the most typical approach when dealing with structured or tabular datasets due to the lower computational cost, faster training, and better tuning of the hyperparameters in algorithms [26,27]. So far, no research has been conducted concerning the ML approach for predicting the survival of EC in Iran. Therefore, this study aims to develop a prediction model for the survival rate of EC in northern Iran based on ML algorithms. In this respect, we first assess all factors associated with the five-year survival of EC using statistical analysis and then build a prediction model for the survival of EC using the ML algorithms based on the best-related factors. The next sections of the manuscript include as follows.

2. Material and methods

This study was a retrospective and applied approach, including five phases as follows.

2.1. Data gathering and familiarization

The research community in this study was the EC patients referred to Imam Khomeini Hospital in Sari City in the Mazandaran province from 2013 to 2020, where their information was retained in that center. We used one single-center database, including the data of EC patients who were referred for treatment measures after EC diagnosis, and their five-year vital status was recorded. The 1656 cases pertained to the five-year survival status of EC patients were recorded in this database. The 1255 and 401 cases belonged to the non-survived and survived cases, respectively. The non-survived cases were patients referred to this center after confirmation of EC diagnosis, and despite the follow-up and treatment measures, they died after five years or less than five years following the EC diagnosis. The surviving cases were patients with similar conditions to non-survived cases but survived after five years. The data in the database included demographic characteristics, signs and symptoms, socioeconomic status, history of personal situations, history of treatment, and laboratory information as input features. The output class was the five-year survival status of EC patients, classified into two types: non-survived (coded as 1 in the database) and survived (coded as 0). The characteristics of EC patients in each survived, and non-survived cases are presented in Table 1.

In Table 1, the skewness and kurtosis are based on the distribution of total EC and non-EC cases, calculated by embedding standard error (SE). The SE for skewness and kurtosis are 0.064 and 0.128, respectively.

2.2. Preparing and analyzing the dataset

In this step, we performed three main tasks for preparing the dataset to build the prediction model for the five-year survival of EC. First, any redundant cases in the database belonging to one person were removed without further processes. Second, we explored the dataset regarding the missing values in the features or output class. In the scenario of lost data existence in independent variables, we faced two conditions: first, if the missing values in the attributes were more than 10%, we omitted the cases with lost features. Otherwise, we filled the lost data with the mode of each feature. For the lost data attributed to the class feature, the cases with the lost data class were excluded from the study. Third, we used the feature selection technique to obtain a subset of features without irrelevant ones, enhance learning performance, develop a more generalized model, decrease memory storage capacity, and promote calculation efficiency [28,29]. We leveraged the binary logistic regression (LR) as a multivariable approach to select the best predictors associated with the five-year survival of EC. The $P < 0.05$ was considered as a significant statistical level. The statistical analysis in this step was performed by IBM SPSS Statistics V 25.0.

2.3. Model development and assessment

In this phase, we developed the prediction models for the five-year survival of EC based on ML algorithms. In this respect, we used the seven famous and widely used algorithms, including the Random-Forest (RF), eXtreme Gradient Boosting (XG-Boost), K-nearest neighborhood (KNN), J-48 decision tree, Bayesian Network (BN), Artificial Neural Network (ANN), and Support Vector Machine (SVM) in Weka 3.9.1 software. We assessed all the hyperparameters of each algorithm when training the ML algorithms. The grid search-based hyperparameters adjustment was leveraged during developing prediction models to obtain high-performing ones. In the grid search, various combinations of hyperparameters in training iterations are used to get the high-performing models. To evaluate

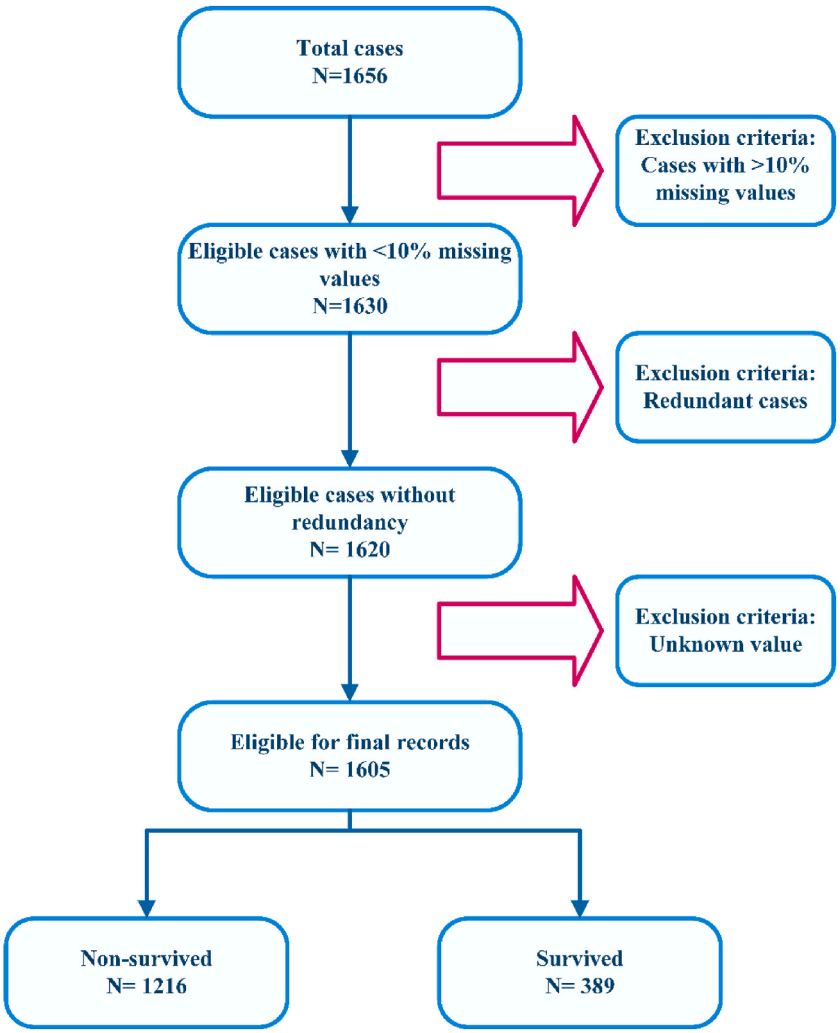


Fig. 1. Flowchart of selecting survived and non-survived cases.

the performance of each ML algorithm, we used negative predictive value (NPV), positive predictive value (PPV), sensitivity, specificity, accuracy, Kappa, and F-Score criteria. The area under receiver operator characteristics (AU-ROC) was used for comparing the performance effectiveness.

2.4. K-fold cross-validation

We require a data-splitting strategy for classification and regression purposes covering training the algorithms using one section of the data and testing using the other section. It should be considered during the learning process because some high-performing algorithms may be ineffective on future test data after training. K-fold cross-validation is one solution in this respect. In this technique, all data samples are divided into K sections. One section is used for testing, and other (K-1) branches are used for training. In this technique, the performance of the classification or regression is the average performance obtained in each fold. This method is beneficial because the various fold is selected for testing in each step; thus, the obtained performance results are more generalizable. Also, there may be some imbalance in the number of data classes. Hence, the stratified K-fold cross-validation is utilized to select instances in each fold based on the class distribution. In this study, the stratified 10-fold cross-validation is leveraged for performance measuring as a commonly used method with the best efficiency in terms of performance [30,31].

2.5. External validation cohort

We performed the external validation test to evaluate the generalizability of the current prediction model for the five-year survival of EC. We used the external records associated with EC patients in Tehran Province. The 54 non-survived and 46 survived samples from Imam Khomeini Hospital in Tehran City were chosen to test the generalizability of the current prediction model. We calculated the True Positive (TP), False Negative (FN), False Positive (FP), and True Negative (TN) as assessment criteria for comparing the output from the best-trained algorithm using external test data and the actual output of the external data. TP and TN refer to non-survived and survived cases correctly classified by the model, respectively. FN and FP are incorrectly classified cases associated with non-survived and survived patients, respectively. The AU-ROC curve of the internal and external validation situations was presented and compared for better insight into the generalizability of the current prediction model.

3. Results

3.1. Preprocessing of the database

We removed 26 cases from the study by excluding the cases having more than 10% missing values. The 50 cases with less than 10% lost data were replaced by the mode of features. By removing the duplicated cases on the five-year survival EC patients, 10 cases were excluded. The process of excluding cases lacking qualification for analysis is shown in the schematic chart in Fig. 1. Also, after consulting with the oncology experts, we removed the cases with the unknown value associated with the grade of differentiation,

Table 2
Analyzing factors influencing the five-year survival of EC based on LR.

Variable	β	Odd ratio (OR)	CI	P
Age at diagnosis (years)	0.11	1.18	[1.12–1.26]	0.01
Gender	0.09	0.882	[0.894–1.025]	0.08
Education	0.08	0.941	[0.842–1.148]	0.1
Place of residence	0.13	0.92	[0.755–1.229]	0.12
Income	0.1	0.967	[0.912–1.06]	0.07
BMI	−0.16	0.783	[0.755–0.821]	0.04
Smoking	0.14	1.324	[1.275–1.523]	0.02
Alcohol	0.14	1.285	[0.975–1.471]	0.08
Obstruction	0.1	1.163	[1.12–1.223]	0.01
Dysphasia	0.17	1.331	[1.26–1.432]	<0.01
Weight loss	0.15	1.159	[1.124–1.18]	0.01
Lymphadenopathy	0.13	1.07	[1.05–1.11]	0.01
Chemotherapy	0.13	1.483	[1.421–1.56]	<0.01
Surgery	0.1	1.435	[1.354–1.55]	<0.01
Radiotherapy	0.16	1.497	[1.252–1.824]	<0.01
Family history of esophageal cancer	0.14	1.351	[1.123–1.527]	<0.01
Tumor stage	0.3	1.775	[1.324–2.11]	<0.01
Type of appearance	0.25	1.528	[1.221–1.745]	<0.01
Histological type	0.36	1.769	[1.326–2.077]	<0.01
Grade of differentiation	−0.33	0.435	[0.34–0.572]	<0.01
Tumor location	−0.4	0.565	[0.225–0.701]	0.01
Tumor size	0.43	1.746	[1.125–2.527]	0.01
Lymphatic invasion	0.28	1.265	[1.214–1.374]	0.01
Vascular invasion	0.18	1.211	[1.142–1.271]	<0.01
Platelet albumin ratio	0.15	1.499	[1.472–1.521]	0.03

tumor location, and tumor stage from analysis due to their clinical importance in predicting the five-year survival of EC upon their opinions and few numbers of cases containing the unknown type. In this respect, we didn't want to replace them with other case values, so the 15 cases associated with tumor location, tumor stage, and differentiation grade with unknown values were excluded from the study. Finally, after exerting the exclusion criteria on the cases, 1605 EC cases remained. 1216 and 389 cases were associated with the non-survived and survived EC cases, respectively. Among the non-survived cases, 765 and 451 belonged to men and women, respectively. Also, 290 and 99 survived patients were associated with men and women, respectively. The results of feature selection based on the LR as a multivariable correlation technique are shown in Table 2.

Table 2 shows the importance of each predictor influencing the five-year survival of EC. β implies the correlation of each factor affecting the survival in the presence of other factors. OR indicates the occurrence ratio of each predictor state, and P is the statistical level, which was considered at $P < 0.05$. The factors, including gender, education, place of residence, income, and alcohol were excluded from the further process.

3.2. Model development and assessment

In this step, we obtained the best-performing ML models for predicting the five-year survival of EC by adjusting the hyper-parameters of the algorithms. The results of the classification capability of the ML-trained algorithms based on the performance criteria, including negative predictive value (NPV), positive predictive value (PPV), sensitivity, specificity, accuracy, Kappa, and F-Score with the best-adjusted hyperparameters and 10-fold cross-validation are presented in Tables 3 and 4, respectively.

Based on the information given in Table 3, the RF model with NPV = 96.1%, PPV = 97.1%, sensitivity = 98.8%, specificity = 91%, accuracy 96.9%, kappa = 91.5%, and F-Score = 98% obtained a more favorable classification capability than other models based on the 10-fold cross-validation. The results obtained by the confusion matrices of the RF-trained algorithm based on different folds of data splitting are presented in Table 5. The K-fold cross-validation data-splitting strategies are in stratified conditions.

As shown in Table 5, the RF-trained algorithm with TP = 1202, FN = 14, FP = 35, and TN = 354 with the stratified 10-fold strategy obtained higher classification capability than other conditions. On the contrary, the model with TP = 1024, FN = 192, FP = 101, and TN = 288 had a lower performance than other data-splitting conditions based on the K = 20. The results of the RF-trained algorithm performance based on the confusion matrices in various numbers of folds are presented in Table 6.

We used the ROC curve to compare and evaluate the performance of ML-trained algorithms in predicting the five-year survival of EC. In this regard, we plotted the ROC of ML-trained algorithms; Fig. 2 shows this scenario. The 10-fold cross-validation was considered due to having optimal performance than other folds of data splitting.

Based on Fig. 2, the RF (AU-ROC = 0.95) obtained the higher competency in predicting five-year survival of EC among patients, and the ROC curve was closer to the sensitivity axis. The XG-Boost (AU-ROC = 0.89) received the second rank in predictability for the five-year survival of EC. The ANN and J-48 decision tree with AUC of 0.82 and 0.81 obtained the third and fourth ranks regarding predictive power, respectively. Also, the KNN, BN, and SVM-RBF models with the AU-ROC of 0.77, 0.75, and 0.71 had a pleasant performance in predicting survival, respectively (AUC>0.7). For the SVM-linear model (AU-ROC = 0.65), the curve was closer to the dividing line of two axes; therefore, it had a lower predictive strength associated with the five-year survival of EC than others. The current study showed that the RF and XG-Boost as the ensemble ML approaches gave us better insight into predicting the five-year survival of EC than other ML algorithms. The importance of each predictor of the five-year survival based on the weighting by the Gini Index (GI) score gained by the RF model is shown in Fig. 3.

As shown in Fig. 3, the tumor characteristics, including the grade of differentiation, tumor size, tumor stage, tumor location, histological type, and vascular invasion, were considered the most important predictors for the five-year survival of EC. The dysphagia as a sign and symptom and family history of EC gained pleasant forecasting power in this respect. On the contrary, BMI and smoking obtained the lowest predictive capability. Generally, based on the results of the current study, the pathological characteristics of tumors played an essential role in prediction purposes. Also, the importance of features to predict the five-year survival of EC among patients based on the RF-trained algorithm is presented based on the SHAP (SHapley Additive exPlanations) values and permutation feature importance, shown in Figs. 4 and 5, respectively. Based on the SHAP values, the factors, including the grade of differentiation, tumor size, vascular invasion, tumor stage, histological type of tumor, and tumor location, were considered the six essential features in predicting the five-year survival of EC. Also, based on the permutation feature importance gained by the RF-trained algorithm, the grade of differentiation, tumor size, vascular invasion, tumor stage, histological type, and tumor location gained the most effectiveness associated with predicting the five-year survival of EC.

Table 3
The performance of ML models based on performance criteria.

Model	NPV (%)	PPV (%)	Sensitivity (%)	Specificity (%)	Accuracy (%)	Kappa (%)	F-Score (%)
RF	96.1	97.1	98.8	91	96.9	91.5	98
XG-Boost	80.5	93.9	93.7	80.9	90.6	74.5	93.8
SVM-RBF	57	90	82.6	71.9	80	50.1	86.2
SVM-linear	49.8	88.3	78.1	67.8	75.6	40.9	82.9
KNN	60.9	91.2	84.7	74.5	82.2	55	87.8
ANN	71.9	92.1	90.5	75.8	86.9	65.1	91.3
BN	55.3	89.5	81.9	70.1	79	47.7	85.5
J-48	68.4	92.4	88.5	77.3	85.8	63.1	90.4

Table 4
The hyperparameters of ML-trained algorithms.

ML algorithm	Best hyperparameters tuned
RF	Maximum depth = 8, maximum iteration number = 100, number of estimators = 15, maximum number of features = 8, maximum leaf nodes = 3, maximum samples = 200, minimum sample split = 2.
XG-Boost	Booster = gradient boosted tree, silent = 0, number of threads = default, eta = 0.2, minimum child weight = 1, maximum depth = 8, subsample = 1, scale positive weight = 1.
SVM-Radial basis function (RBF)	Control parameter (C) = 15, kernel type = RBF, RBF_gamma = 0.1, gamma = 1, epsilon = 0.1.
SVM-linear	C = 10, kernel type = Linear, gamma = 1, epsilon = 0.1.
KNN	3 < K < 9, distance computation = Euclidean metric, cross validate = true, distance weighting = 1/distance.
ANN	Hidden layers = 15, learning rate = 0.5, normalize attribute = true, validation threshold = 50, maximum epoch = 100.
BN	Estimator = BMAE, search algorithm = K2, significance level = 0.05, independence test = pearson Chi square.
J-48	Confidence factor = 0.2, minimum number of object = 1, binary splitting = false, reduced error pruning = true, sub-tree raising = true.

Table 5
The classification capability of RF-trained algorithm by different fold numbers.

RF algorithm K	TP	FN	FP	TN
K = 5	1035	181	93	296
K = 10	1202	14	35	354
K = 15	1186	30	57	332
K = 20	1024	192	101	288

Table 6
The results of RF model performance in various numbers of folds.

RF algorithm K	NPV	PPV	Sensitivity	Specificity	F-Score	Accuracy	Kappa
K = 5	62.1%	92.8%	85.1%	76.1%	88.3%	83.9%	57.8%
K = 10	96.2%	97.2%	99.8%	91%	98%	97.9%	92.5%
K = 15	92.7%	95.4%	98.5%	85.3%	96.5%	95.6%	85.9%
K = 20	60%	91%	84.2%	74%	87.5%	82.7%	54%

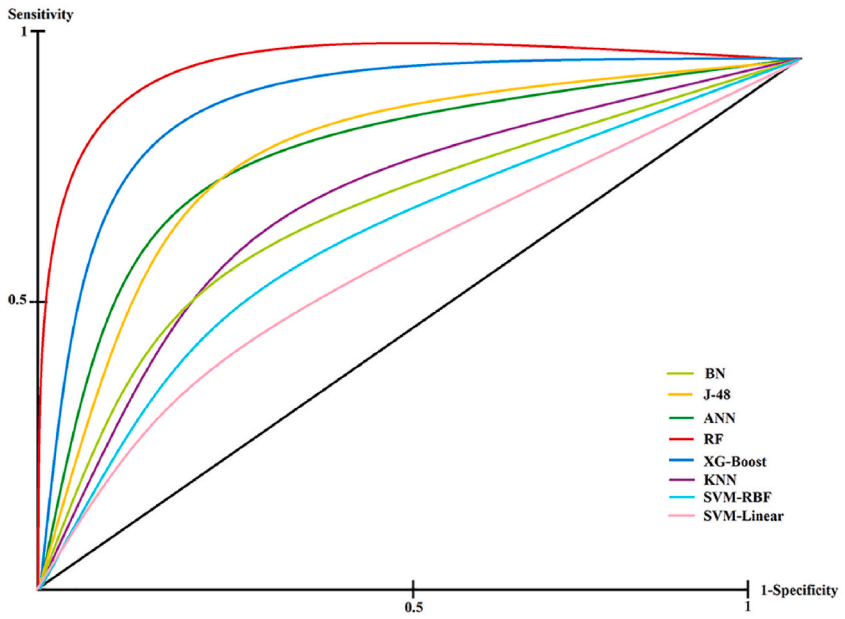


Fig. 2. The ROC of the selected ML models.

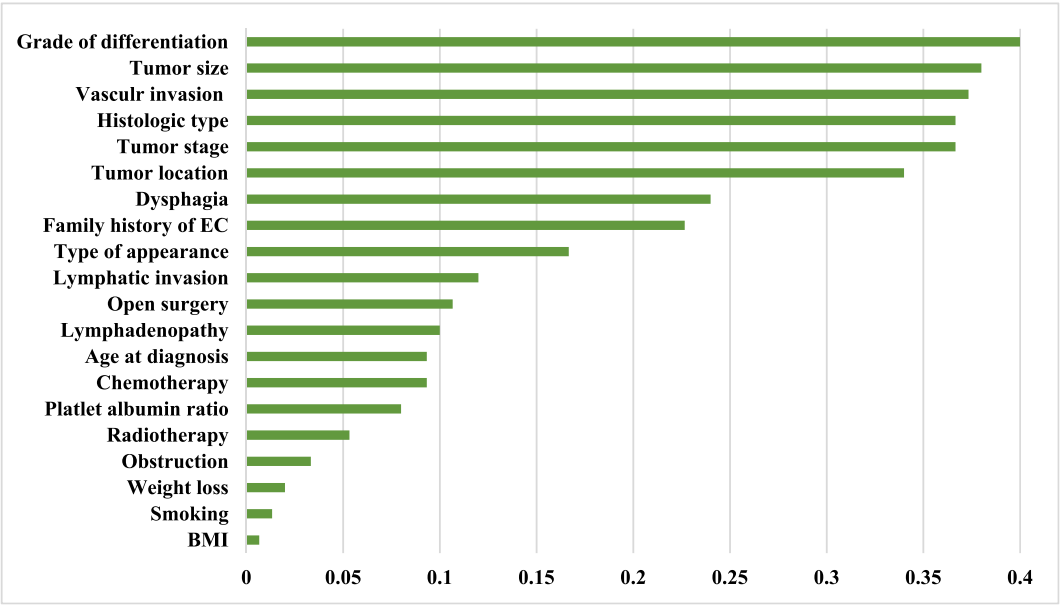


Fig. 3. The RF-based feature importance of five-year survival.

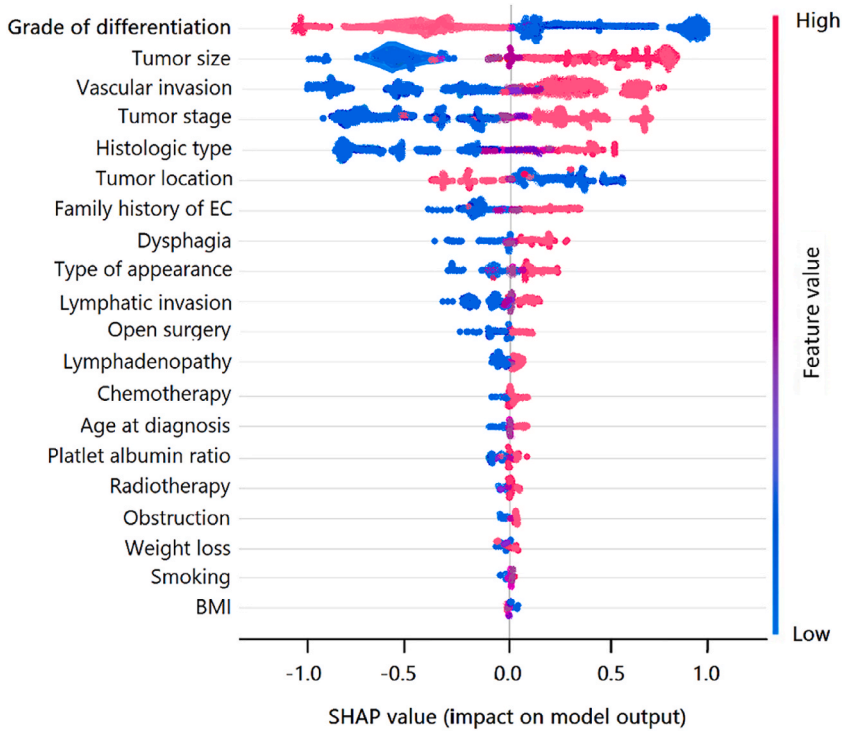


Fig. 4. The SHAP values of predictors influencing the five-year survival of EC.

3.3. External validation cohort

We utilized the 54 non-survived and 46 survived samples from Imam Khomeini Hospital in Tehran City to assess the comprehensiveness of the best ML-trained algorithm in other clinical settings. The results of evaluating the prediction model based on confusion matrices in different numbers of K for the five-year survival of EC patients in external mode are given in Table 7.

Table 7 shows that the RF model with TP = 44, FN = 10, FP = 13, and TN = 33 in K = 10 fold had the higher classification power for

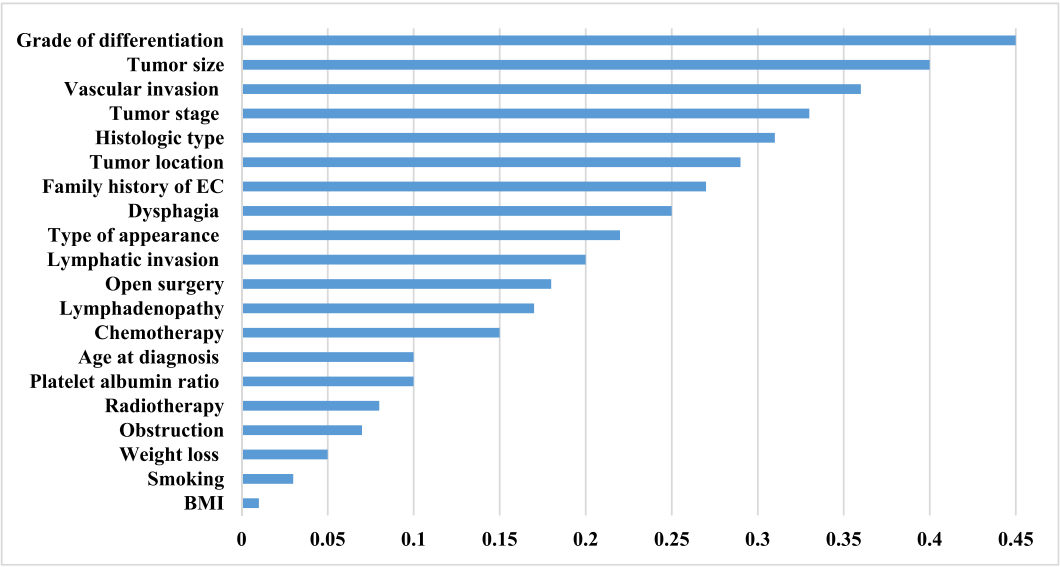


Fig. 5. The permutation feature importance of predictors associated with the five-year survival of EC.

survival based on the external data cases. On the contrary, the RF-trained algorithm with TP = 37, FN = 17, TP = 18, and TN = 28 in K = 5 fold gained lower classification strength. So, The RF-trained algorithm with sensitivity = 81.4% and specificity = 71.7% in K = 10 folds gained nearly suitable predictive power in the external data setting by an average relative decrease of 15%–20% in terms of predictive power than the internal state (sensitivity = 98.8% and specificity = 91%). Also, the ROC curve of the RF model for the internal and external data samples is shown in Fig. 6. According to this, we observed that the ROC curve of RF in internal (AU-ROC of 0.95) and external (AU-ROC of 0.76) situations was almost close (19% reduction in AUC), confirming almost desirable generalizability. The importance of factors influencing the five-year survival of EC based on weighting by GI in the internal and external modes is presented in Fig. 7.

Based on information presented in Fig. 7, the factors, including grade of differentiation, tumor size, vascular invasion, histological type, tumor stage, and tumor location, were considered the most critical factors influencing the five-year survival of EC with a significant increase than other factors based on external data cases, so they were considered as the essential predictors for the five-year survival of EC. Also, they were identified as the best predictors in training the RF algorithm in internal mode. Hence, this similarity in the gained predictors in the two datasets indicates their generalizable strength in predicting the five-year survival of EC in various clinical environments in Iran.

4. Discussion

EC is one of the most common types of cancer worldwide [5]. This cancer has a very high incidence in Iran [32], especially in its northern regions, and the survival rate of EC is low due to its progressive and silent nature and lack of effective preventive and treatment measures [33]. Therefore, this study aimed to introduce ML as an AI solution for the early prognosis and increase the five-year survival of EC among patients. To this aim, we used the multivariable regression analysis to obtain the best predictors to develop prediction models associated with the five-year survival of EC. Based on the results of the current study, the RF-trained algorithm with NPV = 96.1%, PPV = 97.1%, sensitivity = 98.8%, specificity = 91%, accuracy 96.9%, kappa = 91.5%, F-Score = 98%, and AU-ROC = 0.95 obtained the more effective predictability than other ML-trained algorithms. Based on the RF, we extracted the best predictors associated with the five-year survival of EC. The pathological findings, including the differentiation grade, tumor size, tumor stage, tumor location, histological type, and vascular invasion, were recognized as the best-influencing factors for survival, confirmed by three feature importance measurement methods. Although this study was the first research that leveraged ML to predict the survival of EC in Iran, some works have been performed globally, as shown in Table 8.

Table 7
The classification capability of the RF model in external cases.

RF algorithm	TP	FN	FP	TN
K				
K = 5	37	17	18	28
K = 10	44	10	13	33
K = 15	41	13	15	31
K = 20	38	16	17	29

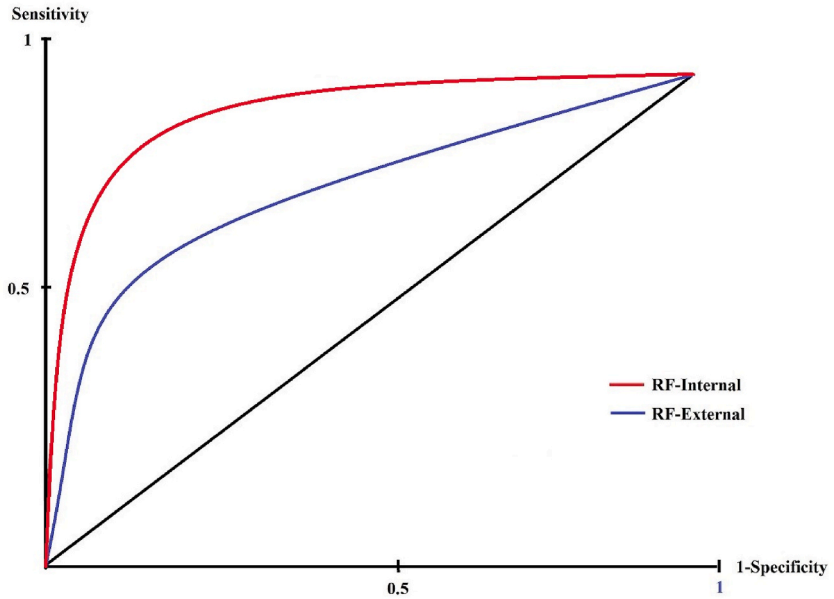


Fig. 6. The ROC of the RF model in internal and external states.

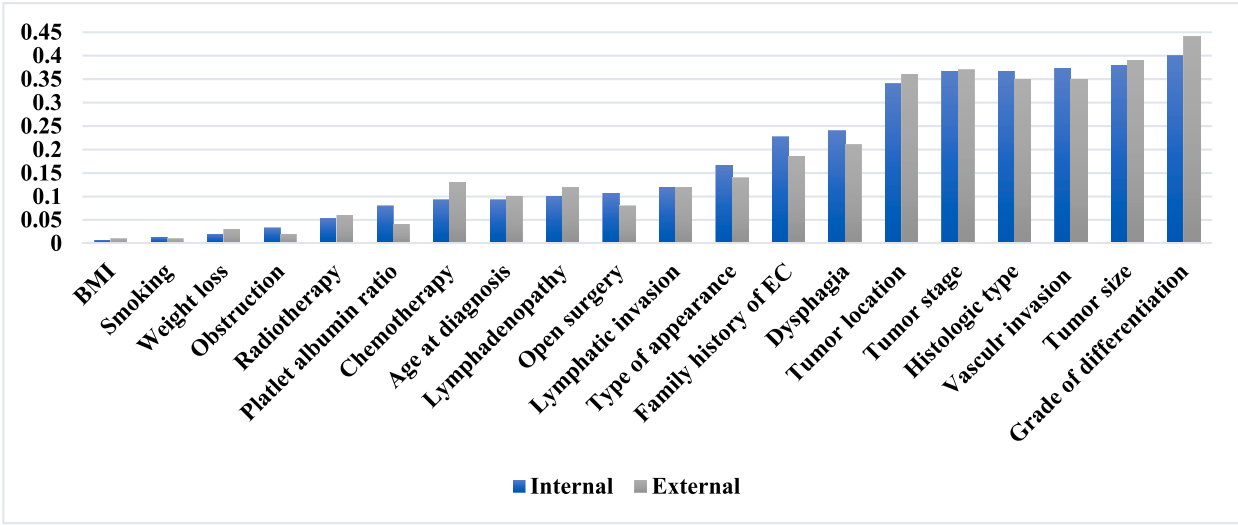


Fig. 7. Feature importance in internal and external modes based on weighting by GI.

As presented in Table 8, Gong et al. obtained the XG-Boost with an AUC of 0.852 as the best-performing prediction model for the five-year survival of EC. In the current study, we got the RF model as the best prediction model with an AUC of 0.95 for the five-year survival of EC. So, the recent research obtained better predictability than the study of Gong. In both studies, the ensemble algorithms gained better predictive power than simple algorithms for survival prediction. Also, based on the analysis of Gong, the clinicopathological features such as tumor size were recognized as crucial predictors for the survival of EC, similar to the current study. They utilized the chi-square technique to select the best feature to construct the prediction model. Although the chi-square is used in studies, it is not considered a powerful approach because it doesn't consider the combinatory effects of factors. Hence, in the current study, we leveraged the multivariable regression analysis to assess the hybrid effects of variables. As Gong et al. mentioned, radiotherapy and chemotherapy negatively impact the survival of EC due to complications that may be generated from these therapies. Although these factors were considered necessary after multivariable regression analysis in the current study, they were not regarded as essential based on the feature importance extracted from the RF. In the study of Wang et al., the GA algorithm was leveraged as a feature selection technique to train the BES-LSSVM algorithm. As a heuristic approach, the GA algorithm generates various scenarios by choosing various sub-features of the original dataset to develop prediction models. Although this approach is beneficial specifically in databases with many variables, in the current study, we used the multivariable logistic regression, and considering the hybrid

Table 8
The previous studies associated with the survival of EC based on the ML approach.

Author (reference)	Measured survival time	Number of features used	Feature selection	Data used	List of variables	Study sample	Models used	External validation	Performance evaluation
Gong et al. [26]	Five-year	24 features	Yes (Chi-Square)	SEER database	Race, sex, primary site labeled, diagnostic confirmation, ICD-O-3 histology behavior, Derived AJCC stage group, Derived AJCC T, Derived AJCC N, Derived AJCC M, RX Summ—Surg Prim Site, RX Summ—Scope Reg LN, RX Summ—Surg Oth Reg/Dis, SEER combined mets at DX-bone, SEER combined mets at DX-brain, SEER combined mets at DX-liver, SEER combined mets at DX-lung, CS tumor size, CS lymph nodes, CS mets at DX, Sequence number, Reason no cancer-directed surgery, Age recode with single ages and more than 85, Regional nodes examined, Regional nodes positive	10,588 EC patients	XG-Boost, CAT-Boost, LightGBM, GBDT, RF, ANN, NB, SVM	No	XG-Boost had the best performance with AUC = 0.852
Wang et al. [34]	Five-year	21 features	Yes (GA)	Clinical data patients affiliated with the Hospital of Zhengzhou University	The 17 blood factors, including white blood cell count, lymphocyte count, globulin, prothrombin time, albumin, red blood cell count, thrombin time, basophil count, eosinophil count, international normalized ratio, neutrophil count, total protein, monocyte count, fibrinogen, hemoglobin concentration, platelet count, and activated partial thromboplastin time, age, and TNM information.	360 patients with ESCC	Bald eagle search and least-squares support vector machine	No	BES-LSSVM had a higher accuracy rate, with 86.538% for the high-age group and 86.495% for the low-age group.
Xu et al. [21]	Five-year	16 features	Yes (Univariate and multivariate regression analysis)	clinicopathological characteristics and follow-up data of ESCC patients at the Department of Thoracic Surgery in Northern Jiangsu People's Hospital	Gender, age, type of surgery, hypertension, diabetes, smoking, drinking, tumor size, tumor center location, histological grade, PT stage, pN stage, vascular invasion, nerve invasion, pathological types, surgical margins,	810 patients with ESCC	Decision tree, RF, SVM, GBM, XG-Boost	No	The XG-Boost model with (AUC = 0.855; 95% CI, 0.808–0.902) was considered optimal.
Zhang et al. [35]	Three-year and five-year survival	27 features	Yes (LASSO regularization and univariable Cox regression analysis)	One single-center database of Sichuan Cancer Hospital	Age, sex, Karnofsky performance scale score, tumor length, tumor grade, tumor location, vascular invasion, surgical margin, dissected lymph nodes number, nerve invasion, T stage, N stage, AJCC8th stage, surgical intervention alone, hematocrit,	2441 ESCC patients	R-part, Elastic Net, GBM, RF, GLMboost, and ML-extended CoxPH method	No	ML-extended CoxPH has a 75.4%, 45.8%, and 26.9% prediction capability for stratifying the low, medium, and high-risk groups for three-year survival. Also, it gained 65.3%, 29.7%, and 11% for 5-year survival.

(*continued on next page*)

Table 8 (*continued*)

Author (reference)	Measured survival time	Number of features used	Feature selection	Data used	List of variables	Study sample	Models used	External validation	Performance evaluation
					mean platelet volume, neutrophil to lymphocyte ratio, monocytes, eosinophil, direct bilirubin, albumin, aspartate aminotransferase, alkaline phosphatase, sodium, magnesium, fibrinogen, lymphocyte -to-monocytes ratio				
Wang [36]	Five-year survival	17 blood indicators	No	One dataset from the State Key Laboratory of EC Prevention and Control of the First Affiliated Hospital of Zhengzhou University and the Key Laboratory of EC Research in Henan Province.	White blood cell count, lymphocyte count, monocyte count, neutrophil count, eosinophil count, basophil count, red blood cell count, hemoglobin concentration, platelet count, total protein, albumin globulin, prothrombin time, international normalized ratio, activated partial thromboplastin time, thrombin time, and fibrinogen	340 EC patients	AMSSA-KELM, ABC-SVM, TLRF, GP-SVM, Cox-LMM	No	AMSAA-KELM gained better capability with the accuracy of 95% and 87.5% for low-risk and high-risk groups of five-year survival prediction.
Current study	Five-year survival	25 features	Yes	One single-center database	Age at diagnosis, gender, education, place of residence, income, BMI, smoking, alcohol, obstruction, dysphasia, weight loss, lymphadenopathy, chemotherapy, surgery (open surgery), radiotherapy, family history of esophageal cancer, tumor stage, type of appearance, histological type, grade of differentiation, tumor location, tumor size, lymphatic invasion, vascular invasion, platelet albumin ratio	1656 EC patients	RF, XG-Boost, SVM, J-48, ANN, KNN, and NB	Yes	The random forest with AU-AUC = 0.95 was identified as a superior model for predicting the survival of EC.

Abbreviations: AJCC, American Joint Committee on Cancer; ESCC, Esophageal squamous cell carcinoma; GA, genetic algorithm.

correlation generated by this technique, the optimal predictive strength is obtained.

Wang et al. focused more on laboratory variables than the pathological data, contrary to the current research, but both studies obtained favorable predictive performance for the survival of EC. In the study by Xu et al., the XG-Boost model with (AUC = 0.855; 95% CI, 0.808–0.902) was considered the best-trained algorithm for predicting the survival of EC. Similar to the current study, the XG-Boost as an ensemble approach was considered the best ML algorithm for prediction purposes. They utilized some clinicopathological and follow-up factors to predict survival, similar to the current study. Although they leveraged the feature selection process to gain the best factors influencing the prediction purposes, they first performed this based on univariate regression. Then, they continued this process by multivariate regression analysis. In this way, we may lose much information on the topic by executing this feature selection process when not dealing with many features. It is due to the variable stratified and abandoned in univariate regression analysis, which might be essential when combined with other factors. Hence, we leveraged the multivariable regression analysis to cover all features associated with the five-year survival of EC. Wang et al. leveraged the 17 laboratory indicators to develop the prediction model for the five-year survival of the EC. The prediction strategy was considered for low-risk and high-risk EC patient groups, estimated with an accuracy of 95% and 87.5%, respectively. In the current study, we used the platelet albumin ratio as a laboratory indicator and clinicopathological variables for the prediction purpose by the ML approach. In addition to the present study, previous studies on this topic leveraged pathological factors for prediction purposes; hence, these factors can significantly enhance the performance and interoperability of the ML models. Zhang et al. predicted the three-year and five-year survival of EC among patients, contrary to other studies focused on the five-year survival of EC. To this end, they estimated the survival among low-risk, medium-risk, and high-risk groups. Similar to the current study, Zhang et al. leveraged one single-center database, including clinicopathological data and laboratory indicators.

Based on their study, the ML-extended CoxPH was obtained as the best predictive solution for the three-year and five-year survival of EC among three patient-stratified groups, as shown in Table 8. In their study, the pathological factors, including tumor grade and tumor stage, were obtained as the best predictors, similar to the current research. Also, the pathological factors were recognized as essential in most previous studies on this topic. So, considering these factors in EC patients and attempting to modify them is crucial for prediction. The pathological factors in this respect can be leveraged for patient counseling based on screening these factors in specific periods by care providers in clinical practice and establishing surveillance protocols as the alternative for treatment. The screening of pathological factors for surveillance leads to improved cancer prognosis in these patients by timely detection of cancer recurrence at earlier stages. This subject leads to a less interventional approach by identifying high-risk EC groups regarding tumor recurrence and leveraging more efficient preventive solutions to increase their survival likelihood. This study showed that the pathological factors have predictive insights into the EC prognosis, so focusing on advancing the pathological diagnostic tools and techniques used in clinical environments has a crucial role in improving the prognosis purposes and survival rate among EC patients by replacing the advanced radiotherapy procedures and chemotherapy having complications with less interventional therapy for patients, leading to more quality of life and survival. In addition to the better clinical insights that can be obtained by focusing on these factors, it also decreases the clinical cost of patients and clinical providers by leveraging fewer interventional therapy measures for EC patients at the community level.

One lack observed in the previous studies on this topic was not performing the external cross-validation. By leveraging this external test, we can interpret the interoperability and applicability of the prediction model in other clinical settings, and it is crucial when leveraging data belonging to a limited number of centers. The current study showed that the RF model with AU-ROC of 0.76 for external validation gained the desirable generalizability in predicting the five-year survival of EC despite leveraging the one single-center database. Although Gong et al. leveraged the SEER database to build the comprehensive prediction model, it is not specified the interoperability of the model in other clinical environments due to the lack of external validity. Also, in other studies that leveraged one single-center database and obtained high performance, the applicability of the model in other clinical centers is not determined.

However, there are some limitations in the present study, including the single-centrality of the database associated with the survival of EC for mining purposes. Although using one single-center database may act optimal in the center that produced the data, it will significantly impact the generalizability of the ML models due to unfamiliar data patterns in other clinical settings. Some missing values were replaced by values from other cases, influencing the performance and generalizability of ML algorithms. Some factors, including follow-up information, more detailed data on treatment measures, and laboratory indicators, may impact the predictability of the five-year survival of EC that was lacking in the current database and so were not considered. Another research concern was the small number of survived (389) samples compared to dead ones (1216). In the current research, the ML algorithms achieved high performance by utilizing this number of survived patients. Although in other studies, the oversampling techniques, such as the synthetic over-sampling technique (SMOTE), were used to increase the model performance, this may affect the external validation results [37]. Also, many ML algorithms, such as XG-Boost, can resolve the imbalance problem intrinsically. Finally, we used the stratified cross-validation method that fully considers the data class imbalance problems. For future studies, it is suggested to use samples belonging to several (at least six) centers to train the ML algorithms as possible. From this viewpoint, using these more diverse samples efficiently enhances the generalizability of the ML models. It is also suggested to use the actual values to replace the lost data to improve the comprehensiveness of ML models as much as possible. Also, we recommend collecting more real samples to maintain the comprehensiveness of the model and improve its performance. The stratified cross-validation can potentially solve the imbalance challenge in the number of data classes, so this data-splitting strategy should be considered.

5. Conclusion

This study aimed to develop a prediction model for the five-year survival of EC using ML algorithms, considering the sophisticated nature of this disease. In this respect, we implemented models using best-selected factors from the regression analysis. Based on the results, the RF with NPV = 96.1%, PPV = 97.1%, sensitivity = 98.8%, specificity = 91%, accuracy 96.9%, kappa = 91.5%, F-Score = 98%, and AU-ROC = 0.95 was recognized as the best model for predicting the five-year survival of EC. Also, assessing the external validity of the RF with the ROC curve showed desirable generalizability results (AU-ROC of 0.76) when used in other clinical settings. The pathological findings were obtained as the best predictors in this respect. This study showed that the RF as an ensemble technique plays a significant role in predicting the five-year survival of EC, considering the poor prognosis of this disease and the less comprehensiveness of other approaches, such as the TNM system. The early prediction of survival based on the RF can introduce more effective follow-up and treatment measures by care providers considering patients' conditions, such as pathological findings and other crucial predictors. Therefore, this solution can significantly impact the increment of survival and quality of life among patients by early predicting tumors through efficient evaluation and modification. It induces a decrement in the cost at the community level through the increased efficiency of treatment measures of care providers and improved patient health outcomes. The disadvantage of this study was the type of dataset used for analyzing the survival of EC among the patients. The single-center dataset may limit the generalizability of the current prediction model on this topic. Also, the few survived cases affected the performance indicators and generalizability. The balanced data based on the actual ones could increase the interoperability of the RF-trained algorithm in other clinical environments, and the current study didn't have this competency.

Data availability statement

Data will be made available by the corresponding author upon reasonable request.

CRediT authorship contribution statement

Raoof Nopour: Writing – review & editing.

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.

References

[1] G. Roshandel, et al., Esophageal cancer crisis in golestan province, Iran; focus on risk factors: back to future, GOVARESH 25 (2) (2020). Summer, 2020.
[2] D.G. Shin, Patterns of lymph node metastasis in esophageal carcinoma and their importance in esophageal cancer treatment, Foregut. Surg. 3 (2) (2023) 49–58, https://doi.org/10.51666/fs.2023.3.e5.
[3] A.W. Asombang, et al., Systematic review and meta-analysis of esophageal cancer in Africa: epidemiology, risk factors, management and outcomes, World J. Gastroenterol. 25 (31) (2019) 4512–4533, https://doi.org/10.3748/wjg.v25.i31.4512.
[4] J. Yang, et al., Understanding esophageal cancer: the challenges and opportunities for the next decade, Front. Oncol. 10 (2020) 1–13, https://doi.org/10.3389/fonc.2020.00102.
[5] D.J. Uhlenhopp, et al., Epidemiology of esophageal cancer: update in global trends, etiology and risk factors, Clin. J. Gastroenterol. 13 (6) (2020) 1010–1021, https://doi.org/10.1007/s12328-020-01237-x.
[6] E. Morgan, et al., The global landscape of esophageal squamous cell carcinoma and esophageal adenocarcinoma incidence and mortality in 2020 and projections to 2040: new estimates from GLOBOCAN 2020, Gastroenterology 163 (3) (2022) 649–658.e2, https://doi.org/10.1053/j.gastro.2022.05.054.
[7] J. Fan, et al., Global trends in the incidence and mortality of esophageal cancer from 1990 to 2017, Cancer Med. 9 (18) (2020), e03338, https://doi.org/10.1002/cam4.3338.
[8] C.-Q. Liu, et al., Epidemiology of esophageal cancer in 2020 and projections to 2030 and 2040, Thoracic Cancer 14 (1) (2023) 3–11, https://doi.org/10.1111/1759-7714.14745.
[9] B. Li, et al., Trends of esophageal cancer incidence and mortality and its influencing factors in China, Risk Manag. Healthc. Pol. 14 (2021) 4809–4821, https://doi.org/10.2147/RMHP.S312790.
[10] R. Daroudi, et al., The economic burden of esophageal cancer in Iran, Indian J. Cancer 59 (4) (2022), https://doi.org/10.4103/ijc.IJC_1009_19.
[11] M. Sheikh, et al., Current status and future prospects for esophageal cancer, Cancers 15 (2023), https://doi.org/10.3390/cancers15030765.
[12] F.S. Asgarian, M. Mahdian, N. Amori, Epidemiology and trends of gastrointestinal cancer in Iran (2004–2008), J. Cancer Res. Therapeut. 17 (4) (2021), https://doi.org/10.4103/jcrt.JCRT_509_19.
[13] E. Najafi, et al., The association of gastrointestinal cancers (esophagus, stomach, and colon) with solar ultraviolet radiation in Iran—an ecological study, Environ. Monit. Assess. 191 (3) (2019) 152, https://doi.org/10.1007/s10661-019-7263-0.
[14] E. Zarean, et al., Determining risk factors for gastric and esophageal cancers between 2009-2015 in east-azarbayjan, Iran using parametric survival models, Asian Pac. J. Cancer Prev. APJCP 20 (2) (2019) 443–449, https://doi.org/10.31557/APJCP.2019.20.2.443.
[15] E.O. Then, et al., Esophageal cancer: an updated surveillance epidemiology and end results database analysis, World J. Oncol. 11 (2) (2020) 55–64, https://doi.org/10.14740/wjon1254.
[16] S. Nemati, et al., Improvement in the survival of esophageal cancer patients at cancer institute of Iran after implementation of the neo-adjuvant chemo-radiation: retrospective cohort study, Middle East J. Cancer 12 (4) (2021) 535–542, 21.84185.1205.
[17] A. Talebi, et al., Survival analysis in gastric cancer: a multi-center study among Iranian patients, BMC Surg. 20 (1) (2020) 152, https://doi.org/10.1186/s12893-020-00816-6.
[18] S. Nemati, et al., National surveillance of cancer survival in Iran (IRANCANSURV): analysis of data of 15 cancer sites from nine population-based cancer registries, Int. J. Cancer 151 (12) (2022) 2128–2135, https://doi.org/10.1002/ijc.34224.
[19] V. Ficarra, et al., TNM staging system for renal-cell carcinoma: current status and future perspectives, Lancet Oncol. 8 (6) (2007) 554–558, https://doi.org/10.1016/S1470-2045(07)70173-0.

[20] R.D. Rosen, SA *TNM classification*. 2023 [2023 Feb 13 2023 Sep 21], Available from: https://www.ncbi.nlm.nih.gov/books/NBK553187/.
[21] J. Xu, et al., Development and validation of a machine learning model for survival risk stratification after esophageal cancer surgery, Front. Oncol. 12 (2022), https://doi.org/10.3389/fonc.2022.1068198.
[22] J. Sun, et al., Five-year prognosis model of esophageal cancer based on genetic algorithm improved deep neural network, IRBM 44 (3) (2023), 100748, https://doi.org/10.1016/j.irbm.2022.100748.
[23] T. Averbuch, et al., Applications of artificial intelligence and machine learning in heart failure, Eur. Heart J.- Digital Health 3 (2) (2022) 311–322, https://doi.org/10.1093/ehjdh/ztac025.
[24] L.J. Muhammad, et al., Supervised machine learning models for prediction of COVID-19 infection using epidemiology dataset, SN Computer Sci. 2 (1) (2020) 11, https://doi.org/10.1007/s42979-020-00394-7.
[25] S. Huang, et al., Artificial intelligence in cancer diagnosis and prognosis: opportunities and challenges, Cancer Lett. 471 (2020) 61–71, https://doi.org/10.1016/j.canlet.2019.12.007.
[26] X. Gong, et al., Application of machine learning approaches to predict the 5-year survival status of patients with esophageal cancer, J. Thorac. Dis. 13 (11) (2021) 6240–6251, https://doi.org/10.21037/jtd-21-1107.
[27] S.B. Atitallah, et al., Leveraging Deep Learning and IoT big data analytics to support the smart cities development: review and future directions, Computer Sci. Rev. 38 (2020), 100303, https://doi.org/10.1016/j.cosrev.2020.100303.
[28] G. Chandrashekar, F. Sahin, A survey on feature selection methods, Comput. Electr. Eng. 40 (1) (2014) 16–28, https://doi.org/10.1016/j.compeleceng.2013.11.024.
[29] J. Li, et al., Feature selection: a data perspective, ACM Comput. Surv. 50 (6) (2017), https://doi.org/10.1145/3136625. Article 94.
[30] S. Yadav, S. Shukla, Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification, in: 2016 IEEE 6th International Conference on Advanced Computing (IACC), 2016, https://doi.org/10.1109/IACC.2016.25.
[31] I.K. Nti, O. Nyarko-Boateng, J. Aning, Performance of machine learning algorithms with different K values in K-fold cross-validation, J. Inf. Technol. Comput. Sci 6 (2021) 61–71, https://doi.org/10.5815/ijitcs.2021.06.05.
[32] H. Salehiniya, et al., The incidence of esophageal cancer in Iran: a systematic review and meta-analysis, Biomed. Res. Therapy 5 (7) (2018) 2493–2503, https://doi.org/10.15419/bmrat.v5i7.459.
[33] M. Gholipour, et al., Esophageal cancer in Golestan province, Iran: a review of genetic susceptibility and environmental risk factors, Middle East J. Dig. Dis. 8 (4) (2016) 249–266, https://doi.org/10.15171/mejdd.2016.34.
[34] Y. Wang, et al., Survival risk prediction of esophageal squamous cell carcinoma based on BES-LSSVM, Computat. Intellig. Neurosci. 2022 (2022), https://doi.org/10.1155/2022/3895590.
[35] K. Zhang, et al., Machine learning-based prediction of survival prognosis in esophageal squamous cell carcinoma, Sci. Rep. 13 (1) (2023), 13532, https://doi.org/10.1038/s41598-023-40780-8.
[36] Y. Wang, et al., Survival risk prediction of esophageal cancer based on the kohonen network clustering algorithm and kernel extreme learning machine, Mathematics 10 (9) (2022) 1367, https://doi.org/10.3390/math10091367.
[37] P.-N. Chen, et al., General deep learning model for detecting diabetic retinopathy, BMC Bioinf. 22 (5) (2021) 84, https://doi.org/10.1186/s12859-021-04005-x.