Heliyon 10 (2024) e24797

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Design of risk prediction model for esophageal cancer based on machine learning approach

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

ARTICLE INFO

Keywords:
Esophageal cancer
Machine learning
Prediction model
Risk factor
Preventive purposes

ABSTRACT

Background and aim: Esophageal cancer (EC) is a highly prevalent and progressive disease. Early prediction of EC risk in the population is crucial in preventing this disease and enhancing the overall health of individuals. So far, few studies have been conducted on predicting the EC risk based on the prediction models, and most of them focused on statistical methods. The ML approach obtained efficient predictive insights into the clinical domain. Therefore, this study aims to develop a risk prediction model for EC based on risk factors and by leveraging the ML approach to stratify the high-risk EC people and obtain efficient preventive purposes at the community level.

Material and methods: The current retrospective study was performed from 2018 to 2022 in Sari City based on 3256 EC and non-EC cases. The six selected algorithms, including Random Forest (RF), eXtreme Gradient Boosting (XG-Boost), Bagging, K-Nearest Neighbor (K-NN), Support Vector Machine (SVM), and Artificial Neural Networks (ANNs), were used to develop the risk prediction model for EC and achieve the preventive purposes.

Results: Comparing the performance efficiency of algorithms revealed that the XG-Boost model gained the best predictability for EC risk with AU-ROC = 0.92 and AU-ROC-test = 0.889 for internal and validation states, respectively. Based on the XG-Boost, the factors, including sex, drinking hot liquids, fruit consumption, achalasia, and vegetable consumption, were considered the five top predictors of EC risk.

Conclusion: This study showed that the XG-Boost could provide insight into the early prediction of the EC risk for people and clinical providers to stratify the high-risk group of EC and achieve preventive measures based on modifying the risk factors associated with EC and other clinical solutions.

1. Introduction

Esophageal cancer (EC) refers to aggressive malignant tumors in the upper parts of the digestive tract, having two histological types, including squamous cell carcinoma (SCC) and adenocarcinoma (AC), generated from the esophageal epithelium [1]. More than 600,000 new cases are diagnosed as EC annually worldwide [2]. On average, the incidence of EC in men is three to four times higher than in women globally [3]. The SCC type is prevalent in Asian nations such as China and Japan. Also, AC has a high prevalence in the United States and European nations, significantly affected by risk factors [4]. The EC, as the eighth prevalent cancer, accounts for the

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

https://doi.org/10.1016/j.heliyon.2024.e24797
Received 29 August 2023; Received in revised form 11 January 2024; Accepted 15 January 2024
Available online 20 January 2024

sixth deadliest malignancy globally, having a poor prognosis with less than 25 % five-year survival rate [5]. Several factors have a significant role in predicting EC, including smoking, alcohol consumption, red meat consumption, drinking hot tea, socioeconomic status, racial factors, and fruit and vegetable consumption, which are considered essential factors in the occurrence of EC among people [6,7].

The EC has an upward trend worldwide, greatly depending on the lifestyles of people and related risk factors such as smoking [8]. Also, the EC is projected to have an ascending trend worldwide by 2030 [9]. It is stated that the United States has 17,290 new cases of EC and 15,850 deaths induced by this disease annually [10]. One belt with the highest prevalence of EC is from the northern and central regions of China to central Asia and the northern regions of Iran [11]. In the last decade, there has been an upward trend in the prevalence of EC in Iran, especially in men, and the age-standardized rate (ASR) of EC in Iran is approximately seven in 100,000 per person [12]. The northern regions of Iran, especially the Golestan province, have the highest prevalence of EC and mortality rate induced by this disease [13].

EC is commonly asymptomatic and is characterized by various signs and symptoms such as dysphagia, weight loss, and odynophagia at advanced stages [14,15]. The early diagnosis of EC is crucial to improve the efficiency of the curative treatment [15]. Also, the survival rate of EC would be increased in EC patients who are diagnosed at early stages [16]. The poor prognosis and ascending prevalence of EC specify demands for enhanced diagnostic and predictive strategies by suitable screening methods. It is crucial to achieving preventive purposes associated with the EC [17,18].

EC is aggressive and silent, which causes a poor prognosis, so early diagnosis of this disease plays an important role in increasing the five-year survival of these patients and reducing the death rate [19]. To this aim, some studies introduced innovative technologies assisted by Artificial Intelligence (AI) to earlier detection of EC for improving the prognosis and then increasing the survival rate of this disease [20]. Tsai et al. used hyperspectral imaging in combination with a deep learning (DL) approach to detect EC at an early stage by finding the tumor lesions in endoscopic images more efficiently. Their study resulted in 88–91 % accuracy in image segmentation concerning EC patients [21]. Semantic segmentation that used the encoder-decoder architecture of artificial neural networks (ANNs) was another technical solution used for the early detection of EC. In this method, the combination of U-net and ResNet was trained by image data, including white light and narrow band types. This study showed an approximate accuracy of 82 % and 85 % for this algorithm fed by white-light and narrow-band image types, respectively [22]. In another work on this topic, the researchers leveraged the combinations of hyperspectral imaging and band-selective technology with color reproduction on 1780 EC images. They concluded that their early detection method can provide a satisfactory diagnostic performance with an average precision of 80–85 % [23].

Another advantageous strategy to decrease the mortality caused by EC is identifying the high-risk group of EC to prevent the onset of cancer, for example, by screening and stratifying high-risk groups and early resecting the premalignant tissues such as Barrett's esophagus [24]. The premalignant tumors of EC, such as squamous dysplasia, can be detected by endoscopy and biopsy techniques. Indeed, the screening methods based on these techniques face challenges in some of the high-risk groups of people, including older adults or people with various gastrointestinal diseases, which makes this solution impractical [25]. Moreover, the screening methods leveraged in communities, such as endoscopy, are costly and invasive and have questionable validity in stratifying high-risk group of EC, limiting their application at the population level [26]. Considering the limitations of these techniques for screening the EC, we require a more practical solution to stratify the high-risk group and reduce the mortality induced by this disease worldwide [27].

Currently, prediction models are leveraged to assess the risk of various diseases for screening purposes to achieve preventive strategies and enhance the quality of life among high-risk groups of people [28,29]. Some studies have been conducted on stratifying the high-risk group of EC as a preventive strategy, which used risk prediction models based on a statistical regression analysis. Chen et al. developed the risk prediction model based on logistic regression to identify the high-risk group of EC. The risk factors used in the prediction model included age, gender, smoking situations, alarming symptoms such as back pain, nutritional factors, and the family history of upper gastrointestinal tumors. The performance measuring of the prediction model based on logistic regression yielded an AU-ROC of 0.81 for predicting the risk of EC [30]. Wang et al. leveraged competing risk regression models to develop the risk prediction model for EC based on some predictors, including smoking, alcohol consumption, body mass index (BMI), physical activity, and demographic factors. The performance of the risk prediction model was obtained with an AU-ROC of 0.76 and 0.7 for internal and external validations, respectively [31]. Etemadi et al. leveraged multivariate logistic regression to build a risk prediction model for EC based on regional risk factors of water source, tea temperature, oral health, opium use, and demographic factors. The prediction model in their study gained an AUROC of 0.77 to stratify the high-risk group of EC [32].

Despite the statistical methods that are beneficial in determining the relationship between variables, they lose their efficiency for prediction purposes when the volume of data increases. Machine learning (ML) approaches are beneficial in building prediction models with high accuracy, especially in conditions where we have large datasets for prediction purposes. Also, considering the high-volume data and various data types that exist in the medical field, such as image data, the ML approach obtained more predictive competency than statistical predictive approaches [33,34].

ML approach, as an AI subfield, attained celebrity by augmenting the utilization of digital data such as Electronic Health Records (EHR) in various fields of medicine [35]. One sub-field of ML is DL, resulting in high performance when used in high-volume and unstructured data formats. On the contrary, the ML approach has this characteristic by using tabular and structured data types [36,37]. ML approach has gained an increasing utilization trend in clinical prediction processes by using retrospective and longitudinal data [38]. They gained significant prediction insight into various clinical conditions, such as drug discovery [39], elderly status assessment [40], heart diseases [41], COVID-19 [42], and cancer [43]. In the domain of leveraging the ML approach for EC, some studies have been conducted worldwide, for example, in predicting prognosis, drug dosage, survival rates, and treatment complications [36,44–46]. However, establishing an efficient risk prediction model is crucial for stratifying the high-risk group of EC [47]. So far, most previous studies have been focused on developing prediction models for EC based on risk factors and statistical methods, and little attention has been paid to ML approach in this respect. Therefore, this study aims to develop a risk stratification model for EC based on ML approach to promote a healthier lifestyle by applying more efficient screening techniques based on risk factors to achieve the preventive purposes.

2. Material and methods

2.1. Study roadmap

Generally, the current study was conducted in seven phases, including data acquisition, database description and familiarization, data preprocessing, feature selection, ML model selection and development, performance evaluation, and external validation. The roadmap of this study is presented in Fig. 1.

2.2. Study design and setting

This study, as a longitudinal and retrospective type, was performed from 2018 to 2022 in three clinical settings, including the Tooba Clinic, Hekmat, and Imam Khomeini Hospitals in Sari City of Mazandaran province. In this study, we used one integrated electronic database, including the data on suspicious people in terms of EC, who referred to the mentioned clinical centers for diagnostic measures, such as CT scans, endoscopy, etc. The data of 3256 samples in the current database were used for this study. During five years of referral, the cases with positive and negative diagnostic results were 1283 and 1,973, respectively.

2.3. Dependent and independent factors

The outcome feature of this study was the diagnostic results of the suspicious people in terms of EC, characterized by positive and negative consequences. The input features for building the current prediction model for EC were age, sex, BMI, history of smoking, alcohol consumption, gastroesophageal reflux disease (GERD), Barrett's esophagus, obesity, fruit consumption, vegetable consumption, drinking hot liquids such as tea, high fat intake, physical inactivity, achalasia, injury to the esophagus, history of certain other cancers, human papillomavirus (HPV) infection, lye such as drain-washer drinking in childhood, red meat consumption, history of radiotherapy, spicy and salty food consumption, difficult defecation, insufficient sleep, nervousness and anxiety, income level, educational level, race, place of residence, and occupation type.

2.4. Data preprocessing

Before model construction, we leveraged the preprocessing technique in the initial database to promote the quality of the data. We

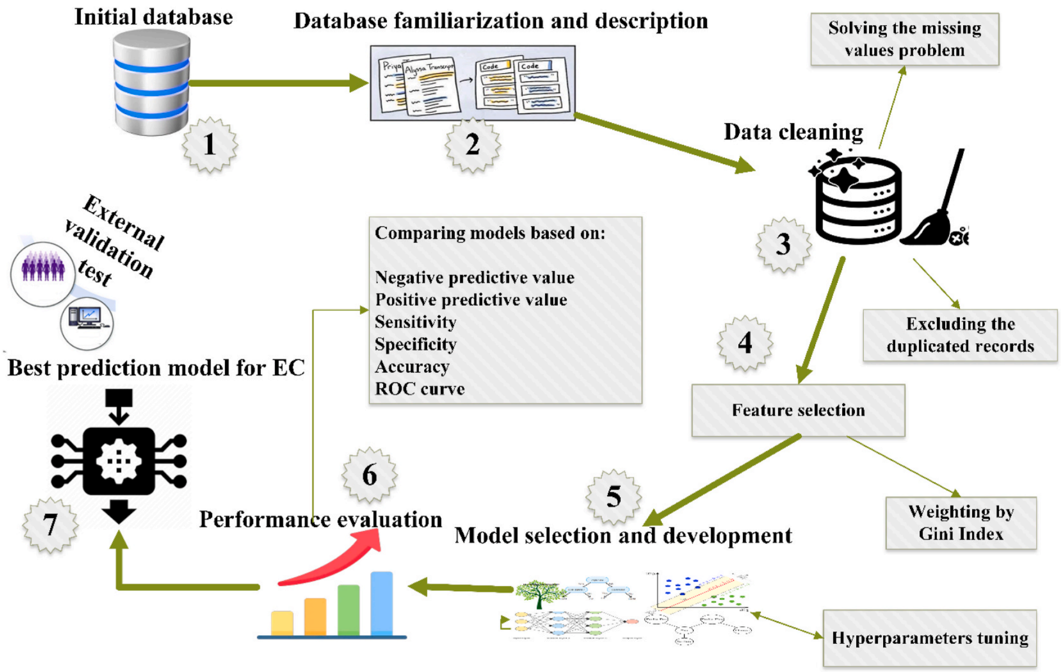
Initial database
1
Database familiarization and description
2
Solving the missing values problem
Data cleaning
3
Excluding the duplicated records
4
Feature selection
Weighting by Gini Index
5
Model selection and development
Hyperparameters tuning
6
Performance evaluation
Comparing models based on:
Negative predictive value
Positive predictive value
Sensitivity
Specificity
Accuracy
ROC curve
7
Best prediction model for EC
External validation test

Fig. 1. The overview of the research methodology.

performed this technique in two steps: first, we investigated the database in terms of existing duplicated cases and excluded any duplicated ones. Second, the input and output features were checked in terms of missing values. In this respect, we adopted two strategies: If the missing values existed in the outcome feature, the cases having this situation were excluded from the study. Otherwise, we faced two conditions: if the cases had more than 10 % missing value in their features, they were excluded from the study; otherwise, we considered the mode of each feature to fill the lost data belonging to the same feature.

2.4.1. Feature selection

It is crucial to select the more relevant features before leveraging the ML approach [48]. This process has some advantages, including executing algorithms faster, lowering the possibility of overfitting, selecting more relevant features for mining purposes, reducing the dataset dimensions, promoting learning functionality, enhancing computational efficiency, decreasing storage space, and increasing the generalizability of the algorithms [49,50]. In this respect, we used the Gini Index (GI) scoring technique (Equation 1) to identify the importance of features and obtain the more relevant ones. The GI determines the randomness of the classified cases by a chosen random attribute. In other words, GI is the probability of randomness classification of cases by a chosen random attribute, ranging from zero to one. GI = 0 means that the classified samples in each group by one attribute are pure or belong to one class type. The GI = 1 indicates the complete random classification by attribute; in other words, the classified cases in each group by one attribute have different class types [51]. A higher value of GI (close to one) by one attribute indicates a lower classification capability, and in this situation, the attribute gains less importance. The lower rate of GI (close to zero) has the opposite situation.

$$\text{Equation 1} : \text{GINI} = 1 - \sum_{i=1}^{n} p_i^2$$

In Equation 1, the P_i represents the probability of one case belonging to one class type.

2.5. Models development and hyperparameter adjustment

We leveraged six selected ML algorithms, including RF, XG-Boost, and Bagging, as ensemble algorithms and K-NN, SVM, and ANNs as simple ones, using the Weka software V 3.9 to develop prediction models. Their selection was due to their popularity and high-performing nature in various fields, such as medicine. The ML learning strategy used in this study was splitting data according to 70:10:20 proportion. In other words, 70 %, 10 %, and 20 % of data were used for training, validating, and testing the selected algorithms, respectively. To gain the most pleasant prediction capability of each ML algorithm, we require the hyperparameters adjustment for each algorithm in the best way. Hyperparameters are the criteria for the learning process in the algorithms. They should be adjusted to the best values to meet the prediction purposes by achieving the highest performance in each algorithm. In the current study, we used the grid search for this purpose.

2.6. Grid search

The grid search is a systematic tuning method of ML hyperparameters, seeking the best combinations of the values in hyperparameters to maximize the performance of the ML algorithms. Contrary to the random search, which selects the values of hyperparameters independently and based on the probability distribution, this method, as a trial-and-error way, performs a comprehensive search on a manually given subset of the hyperparameters for adjusting. This method is beneficial, especially when dealing with a few hyperparameters to optimize. Although the grid search can be used in many optimization problems, it is a celebrated optimization method in the ML approach due to the high accuracy that it gives us when adjusting the hyperparameters. The advantages of leveraging this method are comprehensive search, ensuring all hyperparameter combinations are considered; interpretability of this method, allowing us to comprehend the effect of each hyperparameter combination in the model performance clearly; being fast in parallel computing and implementation, and generating repeatable results after model implementation. In contrast, the disadvantages of the grid search are computational complexity and limited exploration in dealing with high-dimensional hyperparameter spaces [52–55].

2.7. Performance evaluation of algorithms

To measure and compare the performance of ML algorithms, we used several performance criteria, including positive predictive value (PPV), negative predictive value (NPV), sensitivity, specificity, and accuracy obtained by the confusion matrix (Supplementary A). In this study, the area under the receiver operator characteristics (ROC) was measured to compare and evaluate the power of ML algorithms in classifying the EC and non-EC cases.

2.8. Benchmarking to external cases

The efficiency of ML algorithms, in terms of predictive strength in external cases, assures their applicability in other clinical settings. To test the applicability of the ML model for predicting EC and non-EC cases, we leveraged some external data cases from the records belonging to suspicious people in terms of EC referred to Valieasr AJ Hospital in Quaemshahr City for diagnostic tests. In this respect, we used 75 and 101 cases associated with EC and non-EC, respectively. To report the predictive power of the current ML model, we compared the outputs gained by the model when it was used for the test data and the actual outputs that existed in the

Table 1
The characteristics of predictors associated with EC and non-EC cases.

Feature	Value	Total (N = 2955)	EC	Non-EC	P-value
			N = 1175	N = 1780	
Age	<55	1083	452	631	**0.03**
	≥55	1872	723	1149	
Sex	Male female	2071	890	1181	**<0.001**
		884	285	599	
BMI[a]	<18.5	256	176	80	**0.01**
	18.5–25	1623	614	1009	
	25–30	991	336	655	
	>30	85	49	36	
Smoking	Yes	1896	785	1111	**0.01**
	No	1059	390	669	
Alcohol consumption	Yes	325	186	139	0.256
	No	2630	989	1641	
GERD**	Yes	1523	844	679	**<0.001**
	No	1432	331	1101	
Barrett's esophagus	Yes	1611	912	699	**<0.001**
	No	1344	263	1081	
Weight loss	Yes	2068	924	1144	**<0.001**
	No	887	251	636	
Fruit consumption	Low	1613	655	958	**<0.001**
	Medium	845	387	458	
	High	497	133	364	
Vegetable consumption	Low	1598	823	684	**<0.001**
	Medium	929	242	696	
	High	428	110	400	
Drinking hot liquids (tea)	Yes	1927	845	1082	**<0.001**
	No	1028	330	698	
High fat intake	Yes	1452	645	807	**0.03**
	No	1503	530	973	
Physical inactivity	Yes	1827	682	1145	**0.03**
	No	1128	493	635	
Achalasia	Yes	1628	841	787	**<0.001**
	No	1327	334	993	
Injury to the esophagus	Yes	382	255	127	0.25
	No	2573	920	1653	
History of certain other cancers	Yes	1228	510	718	**0.01**
	No	1727	665	1062	
HPV*** infection	Yes	1824	798	1026	**0.01**
	No	1131	377	754	
Lye	Yes	1745	1061	684	**<0.001**
	No	1210	114	1096	
Red meat consumption	Low	675	324	351	**0.01**
	Medium	1687	718	969	
	High	593	133	460	
History of radiotherapy	Yes	1054	426	628	**<0.001**
	No	1901	749	1152	
Spicy and salty food consumption	Yes	1456	686	770	**<0.001**
	No	1499	489	1010	
Difficult defecation	Yes	892	327	565	0.08
	No	2063	848	1215	
Insufficient sleep	Yes	1132	483	649	0.31
	No	1823	692	1131	
Nervousness and anxiety	Yes	1484	827	657	**<0.001**
	No	1471	348	1123	
Income level	Low	1528	712	816	**<0.001**
	Medium	887	411	476	
	High	540	52	488	
Educational level	Illiterate	1527	720	807	**<0.001**
	School level	946	280	666	
	Academic	482	175	307	
Race	Northern	1926	740	1186	0.09
	Turkmen	731	359	372	
	Others	298	76	222	
Place of residence	Rural	1324	735	589	**<0.001**
	Urban	1631	440	1191	
Occupation type	Labors	924	386	538	**0.01**
	Farmer	1045	402	643	
	Housewife	759	250	509	
	Professional/business	227	137	90	

[a] Body mass index, ** Gastroesophageal reflux disease, *** Human papillomavirus.

records. To this aim, the confusion matrix and ROC curve were used to assess the external validation strength of the ML model.

3. Results

3.1. Final dataset description

As stated in the previous section, we prepared the current database before leveraging ML algorithms. After investigating the duplicated records in the database, 126 cases with different IDs that had the same name were excluded from the study. Of 126 cases, 52 and 74 were associated with EC and non-EC, respectively. The 175 cases with more than 10 % missing values were excluded from the study. In this respect, the 56 and 119 cases belonged to EC and non-EC, respectively. The lost data belonging to each attribute in the 152 and 223 EC and non-EC cases, respectively, that had less than 10 % missing values, were filled by the mode of the same attribute in the database. Finally, the 2955 cases remained in the database and were used for the analysis and model construction. The 1175 and 1780 cases were associated with the EC and non-EC, respectively. Of the 1175 EC cases, 890 and 285 belonged to men and women, respectively. Also, the 1181 and 599 cases were associated with men and women, respectively, who reflected the negative diagnostic results. The descriptive statistics of EC risk factors and the difference between the two groups (EC and non-EC) at the statistical level based on the Chi-square test are presented in Table 1.

Based on Table 1, the P-value indicates the difference between the two groups. The risk factors, including age ($P = 0.03$), sex ($P < 0.001$), BMI ($P = 0.01$), smoking ($P = 0.01$), GERD ($P < 0.001$), Barrett's esophagus ($P < 0.001$), weight loss ($P < 0.001$), fruit consumption ($P < 0.001$), vegetable consumption ($P < 0.001$), drinking hot liquids ($P < 0.001$), high fat intake ($P = 0.03$), physical inactivity ($P = 0.03$), achalasia ($P < 0.001$), history of certain other cancers ($P = 0.01$), HPV infection ($P = 0.01$), lye ($P < 0.001$), red meat consumption ($P = 0.01$), history of radiotherapy ($P < 0.001$), spicy and salty food consumption ($P < 0.001$), nervousness and anxiety ($P < 0.001$), income level ($P < 0.001$), educational level ($P < 0.001$), place of residence ($P < 0.001$), and occupation type ($P = 0.01$) gained difference among two different groups. On the contrary, alcohol consumption ($P = 0.256$), injury to the esophagus ($P = 0.25$), difficulty defecation ($P = 0.08$), insufficient sleep ($P = 0.31$), and race ($P = 0.09$) didn't obtain difference in this respect.

3.2. Feature selection

The results of measuring the probability distribution gained by GI, on the importance of the risk factors associated with EC, are illustrated in Fig. 2.

As we stated, the GI indicates the randomness classification capability of each predictor. We considered $GI = 0.5$ as the cut-off point in this regard. In other words, the $GI \leq 0.5$ that belonged to each predictor indicated the low probability of random classification of the cases by the predictor. So, we considered this predictor to build the risk prediction model. On the other hand, $GI > 0.5$ indicated a high probability of random classification by predictors, so this predictor was excluded from the model construction. Based on Fig. 2, the age ($GI = 0.21$), sex ($GI = 0.08$), BMI ($GI = 0.38$), smoking ($GI = 0.32$), GERD ($GI = 0.23$), Barret's esophagus ($GI = 0.27$), weight loss ($GI = 0.18$), fruit consumption ($GI = 0.3$), vegetable consumption ($GI = 0.28$), drinking hot liquids ($GI = 0.19$), high fat intake ($GI = 0.49$), physical inactivity ($GI = 0.43$), achalasia ($GI = 0.17$), history of certain other cancers ($GI = 0.46$), HPV infection ($GI = 0.39$), lye ($GI = 0.34$), red meat consumption ($GI = 0.48$), history of radiotherapy ($GI = 0.36$), spicy and salty food consumption ($GI = 0.41$), nervousness and anxiety ($GI = 0.29$), income level ($GI = 0.35$), educational level ($GI = 0.25$), race ($GI = 0.48$), place of residence ($GI = 0.33$), and occupation type ($GI = 0.46$) with $GI \leq 0.5$ were considered for model construction. The factors, including alcohol consumption ($GI = 0.63$), injury to the esophagus ($GI = 0.59$), difficulty defecation ($GI = 0.65$), and insufficient sleep ($GI = 0.58$) were excluded from the study.

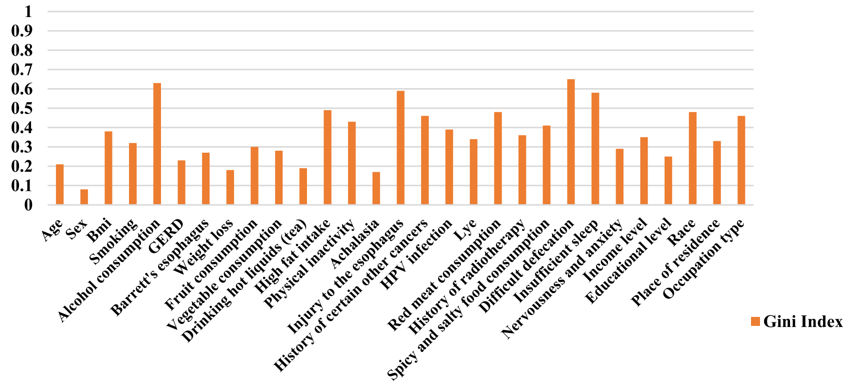

Fig. 2. The results of the GI score belonging to the risk factors associated with EC.

3.3. ML model development and assessment

After gaining the best factors influencing the EC risk prediction, we used them in developing the risk prediction models for EC based on the ML approach. The results of measuring the performance of selected ML algorithms based on the best hyperparameters gained by the grid search method in the total state (average of training, testing, and validation) are presented in Table 2. Also, the performance results in each training, testing, and validation state are brought in Supplementary B.

As mentioned, we compared the performance of the ML algorithms based on the ROC curve. Generally, based on Fig. 3, the XG-Boost with AU-ROC = 0.92 was identified as the best model with higher predictive insight into the EC risk than other ML models. In the second and third ranks, the RF and bagging models were placed with the AU-ROC of 0.85 and 0.83, respectively. The K-NN and SVM models with the AU-ROC of 0.74 and 0.72 gained the fourth and fifth ranks, respectively, in terms of performance efficiency. The lowest performance belonged to ANN, with an AU-ROC of 0.63, so this model obtained lower predictability for the EC risk than other ones. In this study, the XG-Boost gave us a better performance efficiency for predicting the EC risk based on measuring the various performance criteria. Based on the XG-Boost as the better-performing model, we ranked the EC risk factors based on the relative importance (RI) extracted. The results of the ranking of the EC risk factors based on RI are shown in Fig. 4.

As shown in Fig. 4, based on the XG-Boost, the predictors, including sex (RI = 0.28), drinking hot liquids (RI = 0.27), fruit consumption (RI = 0.245), achalasia (RI = 0.23), vegetable consumption (RI = 0.225) gained best predictability for the EC risk. On the contrary, Barret's esophagus (RI = 0.08), lye (RI = 0.13), spicy and salty food consumption (RI = 0.125), educational level (RI = 0.125), and occupation type (RI = 0.125) obtained the less predictive power in this regard.

3.3.1. External validation test

To gain better insight into the generalizability of the current ML model in other clinical environments, we used the 75 and 101 external data cases associated with EC and non-EC, respectively, from the Valieasr AJ Hospital in Quaemshahr City. In this study, the XG-Boost obtained the best performance for predicting the EC risk; hence, we compared the performance of the XG-Boost based on the internal validation obtained in this study and the external validation gained by these external data cases. The classification results of the external cases by the XG-Boost showed that this model correctly classified 62 of 75 EC cases and 88 of 101 non-EC cases, respectively. So, the XG-Boost with TP = 62, FN = 13, FP = 23, and TN = 88 (accuracy = 85 %) obtained favorable performance results in classifying the external data cases. Also, the ROC curve of the XG-Boost (Fig. 5) showed that the predictability of the model wasn't

Table 2
The performance results of the ML algorithms.

Algorithm	hyperparameter	PPV (%)	NPV (%)	Sensitivity (%)	Specificity (%)	Accuracy (%)

ANN	Hidden layers: “15";
	Learning rate: “0.6″
	Normalize attribute: “true";
	Training time: “500";
	Validation threshold: “100";
		59.61	75.4	64.94	70.96	68.56

Bagging	Number of iterations; “20";
	Classifier: “Rep-tree";
	Calculate out of bag: “true";
		78.5	88.32	82.98	85	84.2

K-NN	nearest search algorithm: “KD-tree";
	K-NN: “5";
	Distance weighting: “none"
		63.75	79.79	72	72.98	72.59

RF	Max depth: “10";
	Number of iterations: “100";
	Number of randomly selected features: “5";
	Calculate out of bag: “true";
		83.6	90	85.02	88.99	87.41

SVM	Kernel: “poly kernel";
	Calibrator: “logistic”;
	Tolerance parameter: “0.001";
	C: “10";
		61.2	76.75	66.98	71.97	69.98

XG-Boost	Booster: “gb-tree";
	Eta: “0.1";
	Gamma: “1";
	Max-depth: “8";
	Min-child-weight: “1";
		92.39	94.1	90.98	95.06	93.43

Based on the information provided in Table 2 in the best hyperparameters adjusted by the grid search method, the ANN gained PPV = 59.61 %, NPV = 75.40 %, sensitivity = 64.94 %, specificity = 70.96 %, and accuracy = 68.56 %. Bagging obtained PPV = 78.50 %, NPV = 88.32 %, sensitivity = 82.98 %, specificity = 85 %, and accuracy = 84.20 %. The performance of the K-NN was measured as PPV = 63.75 %, NPV = 79.79 %, sensitivity = 72 %, specificity = 72.98 %, and accuracy = 72.59 %. RF gained PPV = 83.6 %, NPV = 90 %, sensitivity = 85.02 %, specificity = 88.99 %, and accuracy = 87.41 %. SVM obtained PPV = 61.20 %, NPV = 76.75 %, sensitivity = 66.98 %, specificity = 71.97 %, and accuracy = 69.98 %. The performance of XG-Boost was PPV = 92.39 %, NPV = 94.10 %, sensitivity = 90.98 %, specificity = 95.06 %, and accuracy = 93.43 %. We used the ROC curve to compare the performance efficiency of the ML algorithms. The results of measuring the classification capability of the ML algorithms based on the ROC curve in total (average of training, testing, and validation) and testing modes are depicted in Fig. 3 and S-1 (Supplementary C), respectively.

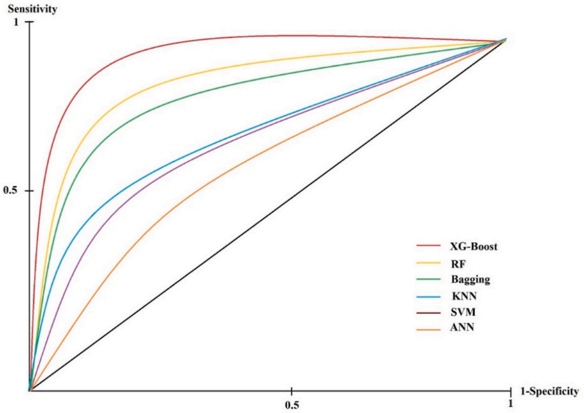

Fig. 3. The ROC of the ML algorithms.

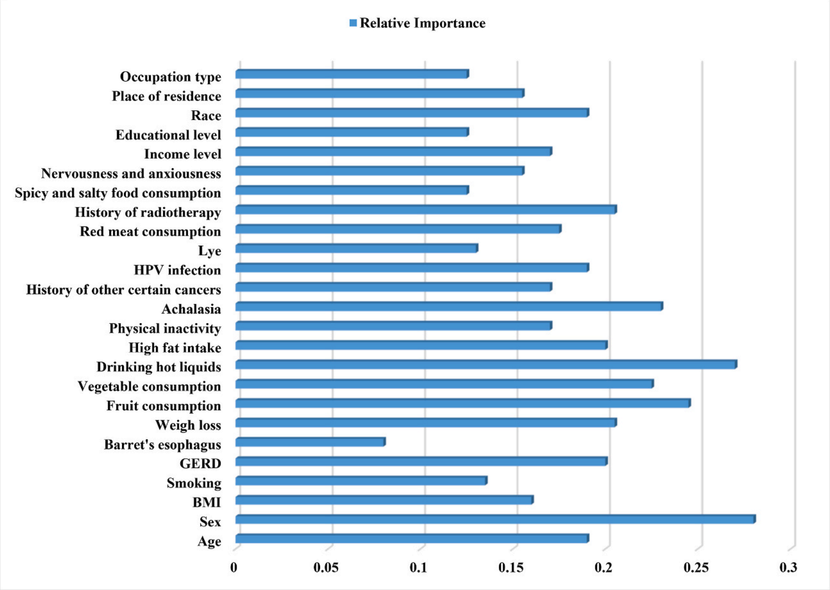

Fig. 4. The RI of the EC risk factors.

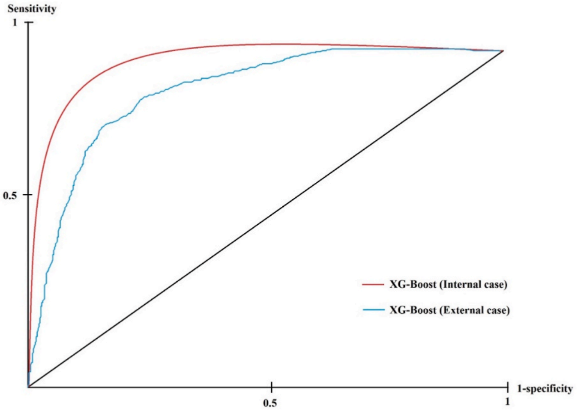

Fig. 5. The ROC of the XG-Boost in the internal and external validation states.

highly reduced. For the internal and external validation, the AU-ROC was 0.92 and 0.83, respectively (difference< 0.1). This difference indicated the desirable generalizability of the XG-Boost for predicting the EC risk in the external state; on the other hand, the XG-Boost model obtained favorable applicability in other clinical environments.

4. Discussion

EC has a high prevalence worldwide, and considering the silent and progressive nature of this disease, the early prediction of EC has a significant role in preventing morbidity and mortality caused by this disease; hence, in this study, we aim to present a new solution for early predicting the EC risk based on the ML approach and risk factors. To achieve this aim, we conducted a retrospective and longitudinal study using one integrated database associated with suspicious people in terms of EC. First, we investigated the database in terms of redundancy and missing values. Second, we used the GI as a feature selection strategy to obtain the best predictors influencing EC risk. Then, based on the best features extracted by the GI, we developed the risk prediction models for EC based on the six selected ML algorithms, including ANNs, bagging, K-NN, RF, SVM, and XG-Boost. In the next step, we measured and compared the performance of each algorithm based on various performance criteria, including PPV, NPV, sensitivity, specificity, accuracy, and ROC curve, to achieve the best model for predicting the EC risk. Finally, we tested our best-performing ML model in terms of predicting the EC risk by the external data cases to evaluate the generalizability of the current model in other clinical environments. Based on comparing the performance of the ML algorithms, we concluded that the XG-Boost with an AU-ROC of 0.92 was the best-performing model in predicting the EC risk. Also, based on the features scored by the XG-Boost model, we observed that the predictors, including sex (RI = 0.28), drinking hot liquids (RI = 0.27), fruit consumption (RI = 0.245), achalasia (RI = 0.23), and vegetable consumption (RI = 0.225) were the five top risk factors for predicting the EC. Based on comparing the performance of the XG-Boost model in the internal and external states, we concluded that this model with an AU-ROC of 0.83 in the external state gained favorable generalizability by a less than 0.1 reduction in the value of AU-ROC.

So far, few studies have been conducted on developing a risk prediction model for EC, and most of them have leveraged statistical methods. In this study, contrary to the previous ones, we used the ML approach to achieve the prediction purpose in this respect. The statistical methods for predicting the EC risk in the previous studies yielded an AUC of 0.7–0.8. In the current study, we obtained the XG-Boost as the best-performing model with an AU-ROC of 0.92 and 0.83 in the internal and external validation states, respectively, which gave us more predictive insight into the EC risk than statistical methods.

The previous studies on the ML approach in the EC have been conducted primarily on the topics of survival rate, early occurrence of tumors after surgery, and complications after chemotherapy, and we mentioned some of them in the current study. Rahman et al. leveraged the ML technique to predict the recurrence after EC surgery using clinical and histopathological characteristics. They developed a prediction model based on elastic net regression (ELR), RF, XG-Boost, and the ensemble of algorithms. Their results showed the AU-ROC of 0.791, 0.801, 0.804, and 0.805 for the ELR, RF, XG-Boost, and ensemble, respectively. Also, the ensemble with an AU-ROC of 0.804 for external validation showed favorable generalizability [56]. In the current study, the XG-Boost gained more performance than other ML models and optimal generalizability, with the AU-ROC of 0.92 and 0.83 for internal and external validation states, respectively. Yoon et al. attempted to predict excessive loss of muscle in Neoadjuvant-chemoradiotherapy (NACRT) based on the ML approach. They used 232 cases belonging to men who underwent NACRT. Also, they used the 70:30 strategy for training and testing the algorithm for model construction. The features, including the percentage of relative change in BMI, platelet to lymphocyte proportion, predictive nourishment indicator, neutrophil to lymphocyte proportion, and albumin during 50 days, were used for this purpose. The results of their study showed the difference between the two different groups (excessive and non-excessive weight loss with $P < 0.001$). The ensemble of LR and SVM with an AU-ROC of 0.808 gained better performance than other ML models for predicting weight loss among NACRT patients [57]. Gong et al. constructed a model to predict the five-year survival of EC patients using the ML approach. They utilized 10,588 cases of EC patients from the SEER database, consisting of 9048 and 1540 non-survived and survived cases, respectively. Several ML algorithms, including gradient boosting models (GBM), XGBoost, CatBoost, LightGBM, gradient boosting decision trees (GBDT), RF, naive Bayes (NB), SVM, and ANN with five-fold cross-validation were leveraged to develop the prediction models for the five-year survival of EC. Their study showed that the XG-Boost with an AU-ROC of 0.852 gained the best performance compared to other ML models in predicting the five-year survival of EC.

This study introduced the ML approach as a less interventional technique assisted by AI to predict the EC risk based on risk factors for screening people and stratifying the high-risk group to achieve the preventive strategy in Iran. However, there were some limitations in the current study that should be considered. In this study, we used the data from three clinical centers in Sari City to develop the prediction model for EC, which might affect the generalizability of the current prediction model to other healthcare environments to some extent. Some missing data were replaced by the mode of features that might impact the performance and generalizability of the model. Some significant predictors might not exist in the current database, so they weren't considered for model construction. The lack of these predictors, due to the retrospective nature of the current study, might impact the efficiency of the current model in terms of performance to some extent. For future studies, we recommend using data from more clinical settings to increase the performance and generalizability of the model and replacing the missing values with the actual data to enhance the generalizability of the model. Also, we recommend the cohort study type instead of a retrospective. This study type assists in collecting more precise data and also considers many factors associated with the purpose of the study that might not be considered in the retrospective type.

5. Conclusion

Due to the high prevalence and rapid progression of EC, early prediction of EC is crucial. So, we got assistance from the ML approach for the efficient prediction of this disease. The results of the study showed that the XG-Boost as an ensemble ML approach could have potential benefits in screening people for EC in terms of risk stratification and identifying the high-risk group. Although the XG-Boost has complexity, this algorithm has shown high predictability in many previous studies, even in the external validation process. In the current study, this algorithm showed a favorable performance in both internal and external validation modes, showing a desirable efficiency regarding the purpose of predicting EC in other clinical environments moreover the internal data from the northern regions of Iran. The use of this strategy to test the predictability of the current model showed its applicability for screening suspicious people in terms of EC in other clinical environments in this country, especially at the regional level. This model with this predictive capability can be useful in preventive medicine that has a higher priority than other clinical aspects such as interventional treatment methods. The current prediction model could predict the EC risk and stratify the high-risk group of individuals by considering the risk factors. Using this model as a predictive solution in this way could promote the clinical solution that could be introduced by various healthcare providers to prevent this disease and improve the overall health of people. These preventive measures are superior to the therapy ones in some aspects. First, in this way, the overall health of people can be more assured and this can eliminate the need for more interventional therapy that the patients tolerate, so the clinical outcomes of people would be improved. Second, leveraging the less interventional measures in this way can diminish the costly interventional measures by care providers, so the clinical efficiency would be increased, leading to a decrease in cost along with improved clinical outcomes in clinical centers at the community level. Third, the knowledge gained by the current ML model can be leveraged in making educational content for people and care providers to increase their knowledge of the risks of EC and effective preventive methods to deal with it. It also can be leveraged in research activities in more widely person groups and adopting the clinical solutions at the community level. Also, the model obtained in the current study could be beneficial in activities such as policy-making and better assigning resources at the community level by assigning the financial and human resources to more suitable clinical and preventive measures in a more efficient manner to improve the overall health of persons. The prediction model obtained by XG-Boost in this study could be leveraged as clinical knowledge for intelligent prediction systems such as clinical decision support systems (CDSS) in clinical environments. In this way, the healthcare providers in their healthcare environments can input the characteristics of suspicious persons on modifiable and non-modifiable factors in terms of EC into the CDSS and get the results on the risk stratification status of these persons. It leads to more suitable preventive measures by various healthcare providers by modifying the modifiable factors, achieving more efficient clinical solutions to the non-modifiable factors as possible, and attempting to promote the quality of life of these high-risk stratified individuals. Although these systems can be used by healthcare providers for stratifying the high-risk group of EC, they cannot replace the decisions of healthcare providers. In other words, the providers should not rely only on the results of these systems. Indeed, these systems can suggest an improvement solution to the overall health of individuals and try to enhance the decision-making power of healthcare providers in achieving suitable preventive and clinical solutions that affect the overall health of individuals.

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.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2024.e24797.

References

[1] J. Li, et al., Esophageal cancer: Epidemiology, risk factors and screening, Chin. J. Cancer Res. 33 (5) (2021) 535–547.
[2] R.J. Kelly, Emerging multimodality approaches to treat localized esophageal cancer, J. Natl. Compr. Cancer Netw. 17 (8) (2019) 1009–1014.
[3] G.K. Malhotra, et al., Global trends in esophageal cancer, J. Surg. Oncol. 115 (5) (2017) 564–579.
[4] M.J. Domper Arnal, Á, Ferrández Arenas, Á, Lanas Arbeloa, Esophageal cancer: risk factors, screening and endoscopic treatment in Western and Eastern countries, World J. Gastroenterol. 21 (26) (2015) 7933–7943.
[5] D.J. Uhlenhopp, et al., Epidemiology of esophageal cancer: update in global trends, etiology and risk factors, Clinical Journal of Gastroenterology 13 (6) (2020) 1010–1021.
[6] C.S. Pramesh, G. Karimundackal, S. Jiwnani, Squamous cell carcinoma of the Oesophagus: the Indian experience, in: N. Ando (Ed.), Esophageal Squamous Cell Carcinoma: Diagnosis and Treatment, Springer Japan, Tokyo, 2015, pp. 279–303.
[7] H. Salehiniya, et al., The incidence of esophageal cancer in Iran: a systematic review and meta-analysis, Biomedical Research Therapy 5 (7) (2018) 2493–2503.
[8] M. DiSiena, et al., Esophageal cancer: an updated review, South. Med. J. 114 (3) (2021) 161–168.

[9] M. Arnold, et al., Predicting the future burden of esophageal cancer by histological subtype: international trends in incidence up to 2030, Official journal of the American College of Gastroenterology | ACG 112 (8) (2017).

[10] M. sadat Yousefi, et al., Esophageal cancer in the world: incidence, mortality and risk factors, Biomedical Research Therapy 5 (7) (2018) 2504–2517.

[11] G. Abbas, M. Krasna, Overview of esophageal cancer, Ann. Cardiothorac. Surg. 6 (2) (2017) 131–136.

[12] H. Rahmani, et al., Burden of esophageal cancer in Iran during 1995-2015: review of findings from the global burden of disease studies, Med. J. Islam. Repub. Iran 32 (2018) 55.

[13] A.-S. Hosseintabar Marzoni, A. Moghimbeigi, J. Faradmal, Gastric and esophageal cancers incidence mapping in golestan province, Iran: using bayesian–gibbs sampling, Osong Public Health and Research Perspectives 6 (2) (2015) 100–105.

[14] J.C. Layke, P.P. Lopez, Esophageal cancer: a review and update, Am. Fam. Physician 73 (12) (2006) 2187–2194.

[15] M.W. Short, K.G. Burgers, V.T. Fry, Esophageal cancer, Am. Fam. Physician 95 (1) (2017) 22–28.

[16] T. DaVee, J.A. Ajani, J.H. Lee, Is endoscopic ultrasound examination necessary in the management of esophageal cancer? World J. Gastroenterol. 23 (5) (2017) 751–762.

[17] F.-L. Huang, S.-J. Yu, Esophageal cancer: risk factors, genetic association, and treatment, Asian J. Surg. 41 (3) (2018) 210–215.

[18] M. Watanabe, et al., Recent progress in multidisciplinary treatment for patients with esophageal cancer, Surg. Today 50 (1) (2020) 12–20.

[19] H. Ge, et al., Symptom experiences before medical help-seeking and psychosocial responses of patients with esophageal cancer: a qualitative study, Eur. J. Cancer Care 2023 (2023) 6506917.

[20] W.-C. Liao, et al., Systematic meta-analysis of computer-aided detection to detect early esophageal cancer using hyperspectral imaging, Biomed. Opt Express 14 (8) (2023) 4383–4405.

[21] C.-L. Tsai, et al., Hyperspectral imaging combined with artificial intelligence in the early detection of esophageal cancer, Cancers 13 (18) (2021) 4593.

[22] Y.-J. Fang, et al., Identification of early esophageal cancer by semantic segmentation, J. Personalized Med. 12 (8) (2022) 1204.

[23] T.-J. Tsai, et al., Intelligent identification of early esophageal cancer by band-selective hyperspectral imaging, Cancers 14 (17) (2022) 4292.

[24] C.S. Yang, X. Chen, S. Tu, Etiology and prevention of esophageal cancer, Gastrointest. Tumors 3 (1) (2016) 3–16.

[25] K. Mönkemüller, L.C. Fry, Gastrointestinal endoscopy: considerations, in: C.S. Pitchumoni, T.S. Dharmarajan (Eds.), Geriatric Gastroenterology, Springer International Publishing, Cham, 2020, pp. 1–31.

[26] G. Roshandel, et al., Endoscopic screening for esophageal squamous cell carcinoma, Arch. Iran. Med. 16 (6) (2013), 0-0.

[27] H. Yang, B. Hu, Recent advances in early esophageal cancer: diagnosis and treatment based on endoscopy, PGM (Postgrad. Med.) 133 (6) (2021) 665–673.

[28] A. Sinha, et al., Risk-based approach for the prediction and prevention of heart failure, Circulation: Heart Fail. 14 (2) (2021) e007761.

[29] P. Fusar-Poli, et al., Preventive psychiatry: a blueprint for improving the mental health of young people, World Psychiatr. 20 (2) (2021) 200–221.

[30] W. Chen, et al., Selection of high-risk individuals for esophageal cancer screening: a prediction model of esophageal squamous cell carcinoma based on a multicenter screening cohort in rural China, Int. J. Cancer 148 (2) (2021) 329–339.

[31] Q.-L. Wang, et al., Development and validation of a risk prediction model for esophageal squamous cell carcinoma using cohort studies, Official journal of the American College of Gastroenterology | ACG 116 (4) (2021).

[32] A. Etemadi, et al., Modeling the risk of esophageal squamous cell carcinoma and squamous dysplasia in a high risk area in Iran, Arch. Iran. Med. 15 (1) (2012) 18–21.

[33] H.S. Rajula, et al., Comparison of conventional statistical methods with machine learning in medicine: diagnosis, drug development, and treatment, Medicina 56 (2020), https://doi.org/10.3390/medicina56090455.

[34] R. Iniesta, D. Stahl, P. McGuffin, Machine learning, statistical learning and the future of biological research in psychiatry, Psychol. Med. 46 (12) (2016) 2455–2465.

[35] J. Ker, et al., Deep learning applications in medical image analysis, IEEE Access 6 (2018) 9375–9389.

[36] 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.

[37] S.B. Atitallah, et al., Leveraging Deep Learning and IoT big data analytics to support the smart cities development: review and future directions, Computer Science Review 38 (2020) 100303.

[38] J.L. Speiser, A random forest method with feature selection for developing medical prediction models with clustered and longitudinal data, J. Biomed. Inf. 117 (2021) 103763.

[39] A. Lavecchia, Machine-learning approaches in drug discovery: methods and applications, Drug Discov. Today 20 (3) (2015) 318–331.

[40] T. Cai, et al., Applying machine learning methods to develop a successful aging maintenance prediction model based on physical fitness tests, Geriatr. Gerontol. Int. 20 (6) (2020) 637–642.

[41] T. Ramesh, et al., Predictive analysis of heart diseases with machine learning approaches, Malays. J. Comput. Sci. (2022) 132–148.

[42] Y. Zoabi, S. Deri-Rozov, N. Shomron, Machine learning-based prediction of COVID-19 diagnosis based on symptoms, npj Digital Medicine 4 (1) (2021) 3.

[43] S. Dalal, et al., A hybrid machine learning model for timely prediction of breast cancer, International Journal of Modeling, Simulation, and Scientific Computing 14 (4) (2023) 2341023.

[44] Y. Cui, et al., Machine learning models predict overall survival and progression free survival of non-surgical esophageal cancer patients with chemoradiotherapy based on CT image radiomics signatures, Radiat. Oncol. 17 (1) (2022) 212.

[45] Z. Zhao, et al., Prediction model of anastomotic leakage among esophageal cancer patients after receiving an esophagectomy: machine learning approach, JMIR medical informatics 9 (7) (2021) e27110.

[46] A.M. Barragán-Montero, et al., Deep learning dose prediction for IMRT of esophageal cancer: the effect of data quality and quantity on model performance, Phys. Med. 83 (2021) 52–63.

[47] R. Chen, et al., Risk prediction model for esophageal cancer among general population: a systematic review, Front. Public Health 9 (2021).

[48] J. Cai, et al., Feature selection in machine learning: a new perspective, Neurocomputing 300 (2018) 70–79.

[49] J. Li, et al., Feature selection: a data perspective, ACM Comput. Surv. 50 (6) (2017) 1–45.

[50] M.A. Khan, et al., Brain tumor detection and classification: a framework of marker-based watershed algorithm and multilevel priority features selection, Microsc. Res. Tech. 82 (6) (2019) 909–922.

[51] S. Tangirala, Evaluating the impact of GINI index and information gain on classification using decision tree classifier algorithm, Int. J. Adv. Comput. Sci. Appl. 11 (2) (2020) 612–619.

[52] G. SijiGeorgeC, B. Sumathi, Grid search tuning of hyperparameters in random forest classifier for customer feedback sentiment prediction, Int. J. Adv. Comput. Sci. Appl. (2020) 11.

[53] H. Alibrahim, S.A. Ludwig, Hyperparameter optimization: comparing genetic algorithm against grid search and bayesian optimization, IEEE Congress on Evolutionary Computation (CEC). 2021. IEEE (2021).

[54] M. Neshat, et al., An effective hyper-parameter can increase the prediction accuracy in a single-step genetic evaluation, Front. Genet. 14 (2023) 1104906.

[55] R Hossain, D Timmer, Machine learning model optimization with hyper parameter tuning approach, Glob. J. Comput. Sci. Technol. D Neural Artif. Intell 21 (2) (2021).

[56] S.A. Rahman, et al., Machine learning to predict early recurrence after oesophageal cancer surgery, Br. J. Surg. 107 (8) (2020) 1042–1052.

[57] H.G. Yoon, et al., Machine learning model for predicting excessive muscle loss during neoadjuvant chemoradiotherapy in oesophageal cancer, Journal of Cachexia, Sarcopenia and Muscle 12 (5) (2021) 1144–1152.