Int J Cancer Manag. January-December 2023 ; 16(1):e138653. https://doi.org/10.5812/ijcm-138653.

Published online 2023 September 30. **Research Article**

Comparing the Performance of Feature Selection Methods for Predicting Gastric Cancer

Hamed Mazreati [1], Reza Radfar [2, *], Mohammad-Reza Sohrabi [3, 4, **], Babak Sabet Divshali [5] and Mohammad Ali Afshar Kazemi [6]

[1]Department of Information Technology Management, Science and Research Branch, Islamic Azad University, Tehran, Iran
[2]Department of Industrial Management, Science and Research Branch, Islamic Azad University, Tehran, Iran
[3]Department of Community Medicine, Shahid Beheshti University of Medical Sciences, Tehran, Iran
[4]Cancer Research Center, Shahid Beheshti University of Medical Sciences, Tehran, Iran
[5]Department of Surgery, Shahid Beheshti University of Medical Sciences, Tehran, Iran
[6]Department of Industrial Management, Central Tehran Branch, Islamic Azad University, Tehran, Iran

[*]*Corresponding author*: Department of Industrial Management, Science and Research Branch, Islamic Azad University, Tehran, Iran. Email: r.radfar@srbiau.ac.ir
[**]*Corresponding author*: Department of Community Medicine, Shahid Beheshti University of Medical Sciences, Tehran, Iran. Email: m.sohrabi@sbmu.ac.ir

Received 2023 June 30; **Accepted** 2023 July 19.

Abstract

Background: Gastric cancer (GC) is a leading cause of cancer-related deaths, emphasizing the importance of timely diagnosis for effective treatment. Machine learning models have shown promise in assisting with GC diagnosis.
Objectives: This study aimed at comparing the performance of various feature selection methods in identifying influential factors related to GC based on lifestyle using machine learning models. The ultimate goal was to enhance early detection and treatment of the disease.
Methods: The data of patients from Shahid Ayatollah Modarres Hospital and Shohadaye Tajrish Hospital between 2013 and 2021 were utilized. Three feature selection methods (filter, wrapper, and filter-wrapper) were employed. The k-fold method validated each model. Four classifiers k Nearest Neighbor (kNN), Decision Tree (DT), Random Forest (RF), and Gradient-Boosted Decision Trees (GBDT) compared their outputs based on feature selection methods.
Results: The filter-wrapper method outperformed others, achieving an area under the ROC curve and F1 score of 95.8% and 94.7%, respectively. GBDT also performed well. The wrapper and RF classifiers achieved an area under the ROC curve and F1 scores of 95.7% and 93.6%, respectively, after the filter-wrapper method. Without feature selection methods, the RF classifier had an area under the ROC curve and F1 scores of 95.6% and 91.7%, respectively, surpassing other classifiers.
Conclusions: This study suggests that appropriate feature selection methods for identifying influential factors related to GC based on lifestyle can facilitate early diagnosis and treatment. The filter-wrapper method demonstrated the best performance in this regard.

Keywords: Artificial intelligence, Machine Learning, Gastric cancer

1. Background

Cancer represents a major global health challenge with a high incidence rate worldwide (1). Gastric cancer (GC) is the fifth most common type of cancer globally (2). However, once the cancer has metastasized to the serosa, the 5-year survival rate is below 5% (3). Medical professionals use intelligent computer applications to improve clinical decision-making, thereby reducing the potential for errors and saving time (4). Given the low survival rate, it is critical to identify appropriate methods for predicting cancer (1). These methods rely on the effective factors identified in GC. Thus, selecting effective factors to enhance the performance of prediction models is crucial (5).

Identifying the risk factors associated with GC is crucial to enable early diagnosis. Given the large number of risk factors involved, it is necessary to use feature selection methods to reduce the number of factors (6). Feature selection involves identifying relevant features for a given problem while discarding redundant or irrelevant ones, to improve classification accuracy (7). The aim of feature selection methods is to decrease the number of necessary features while improving classification accuracy (8). This

method comprises 4 approaches, namely filter, wrapper, embedded, and ensemble (9).

Li et al. (10) used the minimal redundancy maximal relevance (mRMR) algorithm, sequential forward selection (SFS), and K-nearest neighbor classifier to classify lymph node metastasis in GC. Thara and Gunasundari used the infinite feature selection mechanism (IFS) and similarity preserving feature selection (SPFS) to predict GC (11, 12). Qi et al. (13) developed a new feature selection method, using sequential feature selection. To predict Parkinson's disease, Saeed et al. employed various filter and wrapper methods to select features (14). Got et al. achieved superior performance to that obtained using each feature selection method separately by using the whale optimization algorithm and the filter-wrapper feature selection method (15). Singh and Singh used a hybrid filter-wrapper feature selection method that involved 4 stages of validation, filter, wrapper, and classification with an appropriate model for disease diagnosis, resulting in acceptable performance (8). Mandal et al. developed a 3-stage wrapper-filter feature selection framework involving an ensemble formed by 4 filter methods to classify diseases (16). Afrash et al. (17) used the relief feature selection algorithm with 6 classifiers to predict the early risk of GC.

Many previous studies have used expensive methods such as imaging and endoscopy. Some of these methods also have harmful effects on human health. Therefore, in this study, data related to the lifestyle of individuals with GC and healthy individuals were used because these data were collected without the need for expensive and harmful methods. However, since many variables in the field of lifestyle affect the incidence of GC, selecting effective features is of great importance. Therefore, to identify effective features, feature selection methods including filter, wrapper, and filter-wrapper have been compared on this type of data. For this comparison, 4 classifiers include k Nearest Neighbor (kNN), Decision Tree (DT), Random Forest (RF), and Gradient-Boosted Decision Trees (GBDT).

Based on this study, there are a few potential technical gaps that could be addressed: (A) Lack of effective feature selection methods: While there have been previous studies that have used feature selection methods for predicting GC, there may still be gaps in the effectiveness of these methods. This study aims at comparing and evaluating different feature selection methods to identify effective ones for predicting GC based on lifestyle data. (B) Limited use of lifestyle data for predicting GC: Many previous studies have used expensive and invasive methods such as imaging and endoscopy for predicting GC. However, our study aims at using lifestyle data, which is less invasive and more readily available. (C) Need for improved prediction accuracy: GC has a low survival rate, particularly once it has spread to the serosa. Therefore, accurately predicting the risk of GC is critical for early diagnosis and treatment. This study aims at identifying effective feature selection methods to improve the accuracy of predicting GC.

In addition to the introduction section, which discusses the objectives, rationale, and related research, this study consists of 4 other sections, including Methods, Results, Discussion, Conclusions, and Introducing the Tool. In the Methods section, we introduced the methodology used in this study, including dataset preparation and various feature selection methods. In the Results section, we presented the results of accuracy, precision, recall, F1-score, and AUC-ROC based on the calculations performed. The Discussion section compared and examined the results of this study with other similar research. The Conclusions section discussed the practical implications of the results obtained from this study. Finally, in Introducing the Tool section, we introduced the software tool used in this study.

2. Objectives

This study aimed at comparing the performance of various feature selection methods in identifying influential factors related to GC based on lifestyle using machine learning models. The ultimate goal was to enhance early detection and treatment of the disease.

3. Methods

In the initial phase of the study, a dataset of the hospitals and clinics affiliated with Shahid Beheshti University of Medical Sciences and Health Services (SBMU) was utilized. Subsequently, feature selection techniques were employed to identify personal lifestyle-related factors that have a significant impact on GC. The model was, then, validated, using the k-fold method. Following this, each of the classifier models, including DT, RF, GBDT, and kNN was developed and assessed, using the collected data on each of the influential factors. In this study, we used TRIPOD reporting guidelines. The implementation of the designed model was carried out through the use of Python and Jupyter Tool V. 6.4.5.

3.1. Dataset

This study is extracted from the Ph.D. thesis entitled "Designing an intelligent model in Predicting the Pattern of GC in Iran". This research has been approved by Council No. 36 of the Vice-Chancellor in Research Affairs of Islamic Azad University, Science and Research Branch. Then, the

98th Committee of Vice-Chancellor in Research Affairs of SBMU approved it.

In the present study, we used existing medical records, all patients' information was considered confidential, and their identity information was eliminated. Moreover, written consent was received from all patients in this dataset before the authors received it. Therefore, a number was allocated to each patient, and this number was entered into the software anonymously. So, information that leads to the disclosure of patients' identities was not published by the main study team. The data were coded and labeled in a way that masked the identities of the predictors and outcome variables. Each variable was assigned a unique identifier that was unrelated to its content, ensuring that the assessors remained blind to the specific predictors being assessed.

The thematic scope of the study consists of people with GC covered by the hospitals and clinics of SBMU. The spatial scope of the study encompasses the selected hospitals and clinics of this hospital. Besides, the temporal scope of the study consists of the period 2013 - 2021.

The dataset was two classes, including people with and without GC. Based on this dataset, 51 factors were identified as effective factors (Table 1). Since some of these factors should have been categorized into several sub-factors to enter the software, we used one-hot encoding and the number of these factors has reached 86 effective factors. One-hot encoding is a commonly employed technique in machine learning for handling categorical data.

Moreover, the characteristics of the study population are listed in Table 2. In our dataset, missing data were observed to be zero for all variables.

3.2. Feature Selection

There are two dimensionality reduction methods: feature selection and extraction. The feature selection method chooses solely a set of the first options that contain relevant info. In distinction, feature extraction transforms the input area into a lower-dimensional mathematical space to preserve the foremost relevant information. New opportunities will not be created throughout feature selection; however, through feature extraction (18).

Using dimensionality reduction techniques must be selected to prevent over-fitting caused by a large number of factors and a small sample size. The feature selection method is a strategy for preprocessing high-dimensional data that can lead to more straightforward, more understandable, and better-performing models in data mining methods (19).

There are 3 types of feature selection methods. The first type is the "Filter method", which operates independently of learning algorithms. The second type is the "Wrapper method", which depends on learning algorithms, and the third type is the combined "Embedded method", which selects features based on a specific learning algorithm (20).

Filter and wrapper techniques are the two most typical feature selection techniques. The benefits of those models are unit standard procedure cost and smart generalizability. Researchers have agreed that no "best" (absolute) technique exists for feature selection. Thus, this study uses the "Filter" and "Wrapper" methods." A comparison of the execs and cons of those two ways has been summarized in Table 3.

3.2.1. Filter Method

Filtering methods are usually used as a pre-processing step. The filter method is independent of any machine learning algorithm. Instead, features are selected based on their scores in various statistical tests for their correlation with the outcome variable (Table 4). Here the correlation is a subjective term.

Since the features and response of the feature selection method are categorical, the chi-square statistical test should be used, whose formula and the notation in the equation are as follows:

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

χ_c^2: Chi-squared statistic, a measure of the difference between observed and expected values.

$\sum$: The summation symbol, which indicates that the values inside the parentheses should be summed up.

O_i: Observed value for the i-th category.

E_i: Expected value for the i-th category.

$(O_i - E_i)^2$: The difference between observed and expected values squared.

C: Degrees of freedom, which is the number of categories minus one.

3.2.2. Wrapper Method

In wrapper methods, our aim is to utilize a selected set of features to train a model. By analyzing the insights gained from the initial model, we make decisions regarding the addition or removal of features within the subset. This method includes the following three groups:

- Forward Selection
- Backward Elimination
- Recursive Feature Elimination

3.2.3. Embedded Methods

Embedded methods combine the qualities of filter and wrapper methods. It is implemented by algorithms that have built-in feature selection methods.

Table 1. Factors Affecting Gastric Cancer Based on the Personal Lifestyle of Individuals [a]

No.	Factors	No.	Factors	No.	Factors
1	Sex	22	History of helicobacter pylori	43	Consumption of spicy foods
2	Age	23	History of acid reflux	44	Consumption of refined beans
3	Height	24	Radiation history	45	Consumption of fried foods in oil
4	Weight	25	Stomach ache	46	Consumption of carbonated beverages
5	Residence	26	High blood pressure	47	Consumption of vegetables
6	Education	27	High blood fats	48	Fruit consumption
7	physical activity (daily)	28	Feeling discomfort in the abdomen after a meal	49	Smoking
8	Alcohol	29	Flatulence	50	Job
9	Breakfast	30	Early satiety	51	Monthly family income
10	High-salt diet	31	Belching		
11	Eating fast	32	History of aspirin use		
12	Dust exposure	33	History of taking stomach pills		
13	Facing with cement	34	History of metformin use		
14	Exposure to metals	35	History of use of glipizide, gliclazide and glibenclamide		
15	Exposure to volcanic material	36	Consume red meat		
16	Exposure to air pollution	37	Consumption of fish		
17	Family history of gastric cancer	38	Tea consumption		
18	Family history of other cancers	39	Consumption of hot drinks		
19	History of esophageal cancer	40	Consumption of pickles		
20	History of gastric ulcer	41	Consumption of frozen foods		
21	History of gastric surgery	42	Consumption of salty foods		

[a] In this table, 51 factors affecting gastric cancer are shown, which were collected based on the data of gastric cancer patients in the hospitals of SBMU.

Table 2. Characteristics of the Study Population [a]

Factors	People with Gastric Cancer (n = 173)	People without Gastric Cancer (n = 157)
Female	115 (66.5)	103 (65.6)
Age, y	51 ± 18.4	51 ± 18.3
Height, cm	170 ± 10.7	167 ± 10.7
Weight, kg	70 ± 17.4	71 ± 19.93

[a] Values are expressed as No. (%) or mean ± SD.

Table 3. A Comparison of the Pros and Cons of the Filter Method" and "Wrapper Method" (21) [a]

Method	Filter Method	Wrapper Method
Pros	Independent of learning model Fast execution; appropriate for high dimensional data; Generalizability	Better performance attainability; considering the interaction between features; Recognizing feature interactions of higher order.
Cons	Ignorance of Interactions between features; unable to handle the redundancy problem; Lack of interaction with the learning algorithm	High cost in terms of execution times; susceptible to overfitting; Creating a learning algorithm from scratch for each subset.

[a] A comparison of the advantages and disadvantages of filter and wrapper methods is shown.

Table 4. Classification of Statistical Tests Based on the Type of Features and the Response of Feature Selection Methods [a]

Feature/Response	Continuous	Categorical
Continuous	Pearson's correlation	LDA [b]
Categorical	ANOVA	Chi-square

[a] Statistical tests are determined based on the type of features and the response of the methods. For example, if the selected features and the response of the method are categorical, the chi-square test is used.
[b] Linear discriminant analysis

3.3. Classification Algorithms

3.3.1. Decision Tree

A decision tree (DT) is among the most frequently used algorithms in data mining, where DT serves as a predictive model applicable to both regression and classification models. According to the DT structure, predictions generated by the tree are explained as a set of rules. Each path from the root to a DT leaf represents a classification rule. Finally, the desired leaf is labeled with the class with the highest number of records.

3.3.2. Random Forest

DT serves as one of the most popular models in hybrid methods. Robust models consist of several trees, known as forests. The trees making up a forest can be shallow or deep. Shallow trees have low variance and high bias, rendering them suitable for hybrid methods. In contrast, deep trees have low bias and high variance, making them ideal for bagging methods focused on reducing conflict.

3.3.3. Gradient-Boosted Decision Trees

The gradient-boosting algorithmic program is among the foremost powerful machine learning algorithms introduced over the past 20 years. Though this algorithmic program was designed to wear down classification issues, it can even be applied for regression. Gradient boosting aimed at developing a technique for combining the output of many "weak" classifiers to get a strong "committee".

The purpose of the gradient boosting algorithmic program is to consecutive apply the weak classification algorithmic program to repeatedly changed versions of the information, thereby manufacturing a sequence of weak classifiers.

3.3.4. K-Nearest-Neighbor

The nearest neighbor method (aka kNN) is an instance-based learning method and is among the simplest ML algorithms. The classification of a sample in this algorithm is based on a majority (plurality) vote from its neighboring samples. The sample is assigned to the most prevalent class among its k nearest neighbors, where k is a small, positive integer. When k equals 1, the sample is directly assigned to the class of its closest neighbor. It is important to choose an odd value for k to prevent any ties in the classification process.

The performance of each of the above classifiers is compared, using the area under the ROC curve and F1 curves, calculated using the following formulas and notations:

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

$$Recall = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$F1 = 2 \times \frac{Recision \times Recall}{Precision + Recall}$$

TP: True Positive (refers to the number of correct positive predictions made by the model)

TN: True Negative (refers to the number of correct negative predictions made by the model)

FP: False Positive (refers to the number of incorrect positive predictions made by the model)

FN: False Negative (refers to the number of incorrect negative predictions made by the model)

4. Results

4.1. Feature Selection

Filter-wrapper hybrid methods are used because there is no best technique for feature selection. So, the filter technique is applied first to the practical issue knowledge collected; hence, the wrapper technique is applied to the output. Four standard classifiers utilized in similar alternative studies (DT, RF, GBDT, and kNN) were evaluated. Finally, 3 methods (filter, wrapper, and filter-wrapper methods) were compared.

4.1.1. Filter Method

In this step, data collected on 86 efficiency factors are entered into the filter part of the model. Since the features and response of the filter method are categorical, we have used the chi-square statistical test. Therefore, features that have a P-value above 0.05 have been removed due to their lower correlation with the response of the filter method.

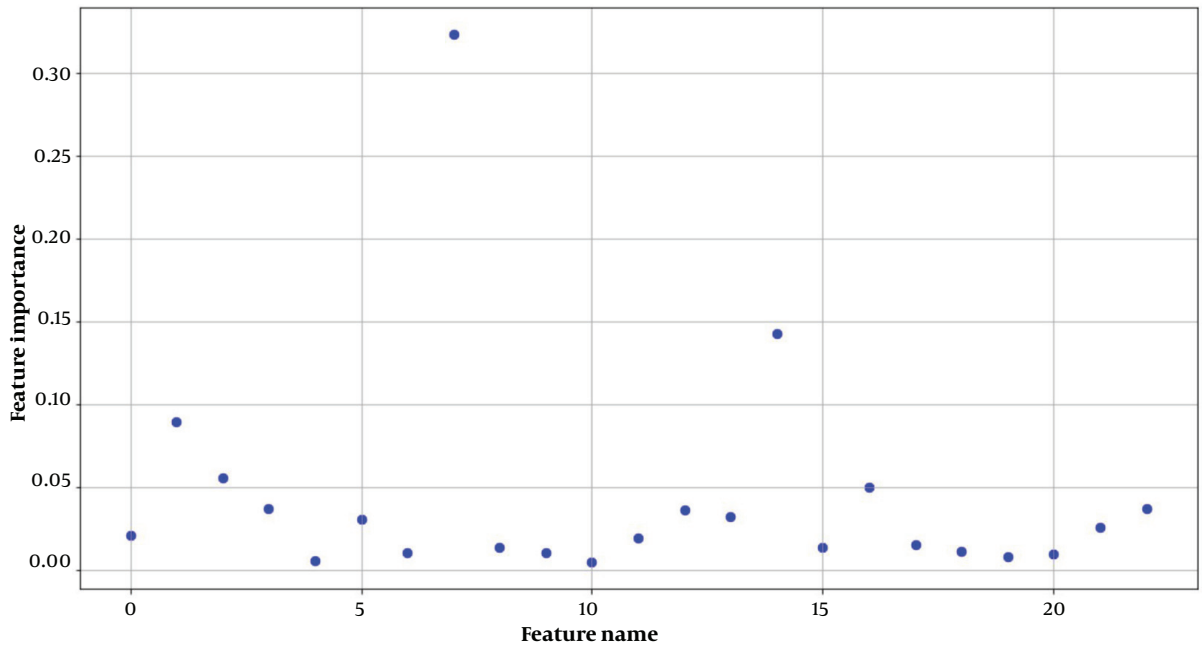

Figure 1. Dispersion diagram of factors selected by the wrapper method.

4.1.2. Wrapper Method

In this step, 23 efficiency factors determined by the filter method are imported into the wrapper section of the model, producing the following output. The procedure involves initially considering all features and subsequently eliminating the least significant feature in each iteration, aiming at enhancing the model's performance. This process continues until no further improvement is discernible through feature removal. Finally, Linear Support Vector Classification has been used as a classifier model.

Figure 1 illustrates the dispersion of the 23 efficiency factors generated by the filter model according to their relative importance.

Accordingly, 5 factors were identified influencing GC based on personal lifestyle including education, physical activity (days per week), history of gastric surgery, consumption of salty foods, and consumption of spicy foods.

4.2. Cross-Validation (CV)

Several parameters in many classification models can control complexity. The good values for the complex parameters were found to achieve the best prediction performance in the new data, resulting in the best model. To achieve this objective, the input data undergoes a partitioning process known as k-fold cross-validation, where it is split into k = 10 sets comprising both training and testing datasets.

4.3. Implementing Classifier

The performance of each of the above classifiers was compared, using the area under the ROC curve and F1 score. As shown in Table 5, when the filter-wrapper method is used, the area under the ROC curve and F1 score are higher (95.8%, 94.7%) than the other methods. Furthermore, it can be seen that in the Filter-Wrapper method, the GBDT classifier performs better. After the Filter-Wrapper method, the RF classifier and wrapper method have more areas under the ROC curve and F1 score (95.7%, 93.6%). Finally, the filter method and the RF classifier have more areas under the ROC curve and F1 score (95.6%, 91.7%) than other classification methods of this method. This model is shown in Figure 2. Moreover, the area under the ROC and PR curves are presented in Appendices 1 to 4 in the Supplementary File.

5. Discussion

In this study, we proposed a filter-wrapper hybrid method for feature selection in predicting GC based on personal lifestyle. Our results showed that the GBDT classifier using the filter-wrapper method outperformed other classifiers with a higher area under the ROC curve

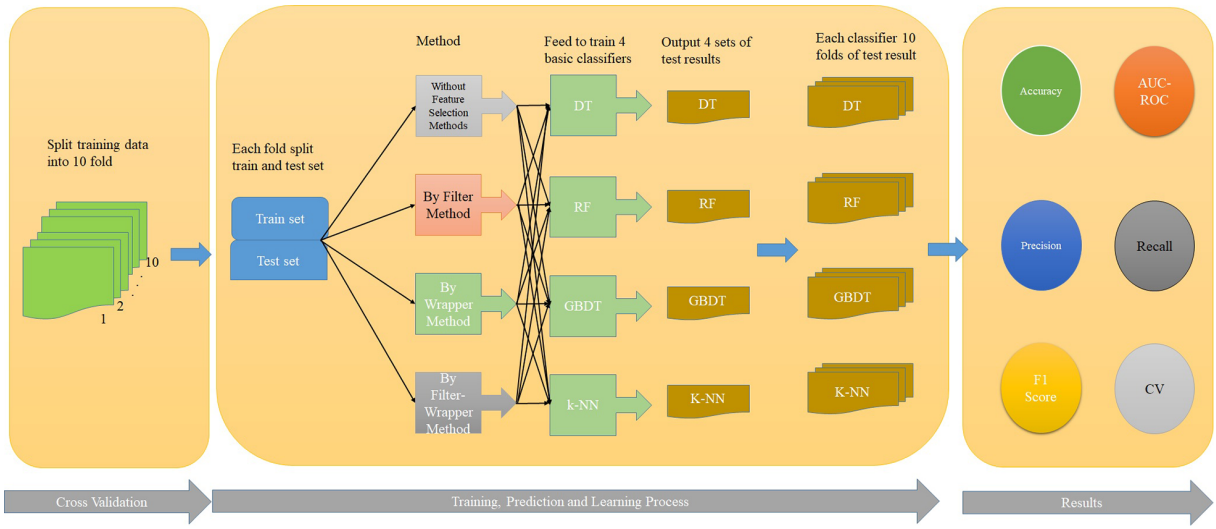

Figure 2. Components of the model

Table 5. A Comparison of the Classifiers for Performance

Method and Classifier	CV	Precision	Recall	F1 Score	Accuracy	AUC-ROC
Without feature selection methods						
GBDT	0.980	0.960	0.952	0.947	0.962	0.952
RF	0.977	0.960	0.953	0.917	0.994	0.956
kNN	0.862	0.734	0.734	0.680	0.810	0.897
DT	0.936	0.939	0.915	0.899	0.939	0.915
By filter method						
GBDT	0.982	0.960	0.949	0.942	0.958	0.948
RF	0.984	0.960	0.952	0.929	0.978	0.952
kNN	0.953	0.859	0.906	0.857	0.965	0.952
DT	0.947	0.949	0.940	0.937	0.947	0.939
By wrapper method						
GBDT	0.971	0.939	0.954	0.940	0.971	0.955
RF	0.977	0.939	0.958	0.936	0.984	0.957
kNN	0.961	0.889	0.906	0.857	0.963	0.948
DT	0.909	0.909	0.932	0.928	0.940	0.930
By filter-wrapper method						
GBDT	0.976	0.939	0.958	0.947	0.973	0.958
RF	0.976	0.929	0.954	0.941	0.969	0.955
kNN	0.972	0.919	0.950	0.934	0.969	0.955
DT	0.928	0.929	0.940	0.935	0.950	0.936

Abbreviations: GBDT, gradient-boosted decision trees; RF, random forest; Knn, k nearest neighbor; DT, decision tree.

(95.8%) and F1 score (94.7%). This method can reduce costs and prevent physical complications compared to endoscopic image-based diagnosis.

Comparing our results to previous studies, we found that our method performed better than other classifiers and feature selection methods. For example, Biglarian et al. (22) created a neural network model with an accuracy of 85.6%. 436 GC patients, who underwent surgery between 2002 and 2007 at Taleghani Hospital in Tehran, Iran were included in the study. Feng et al. (23) achieved an accuracy of 76.4% by using 490 patients, who were diagnosed with GC between January 2002 and December 2016 in diagnosing GC and CT images. Zhu et al. (24) achieved an accuracy of 89.16 % by using the CNN model to diagnose GC based on endoscopy images. A total of 993 endoscopic images of GC tumors were acquired from the Endoscopy Center of Zhongshan Hospital for their study. Wu et al. (25) reached an accuracy of 78.5% by using deep learning models and endoscopy images. The study utilized a dataset comprising 100 consecutive patients who underwent magnifying narrow-band (M-NBI) endoscopy at Peking University Cancer Hospital between June 9, 2020, and November 17, 2020.

Taninaga et al. (26) conducted a study at a single facility in Japan, involving 25 942 participants who underwent multiple endoscopies between 2006 and 2017. They employed the XGBoost algorithm to predict GC and achieved an accuracy rate of 77.7%. Amirgaliyev et al. (27) compared 4 following algorithms: Logit, k-NN, XGBoost, and light GBM, and concluded that Boost has better accuracy (95%) than other algorithms. Mortezagholi et al. (28) compared 4 following algorithms: SVM, DT, and naive Bayesian, and concluded that SVM has better accuracy (90.08%) than other algorithms.

The filter-wrapper hybrid method was used in our study because it combines the advantages of filter and wrapper methods. We first removed irrelevant features based on the P-value defined in the filtering step, then imported the remaining features into the wrapper model to make the model run faster. Finally, 5 features were selected as factors influencing GC based on personal lifestyle.

In conclusion, our study provides a promising approach for predicting GC based on personal lifestyle, using the GBDT classifier and the filter-wrapper hybrid method. The proposed method can help reduce the cost and physical complications of endoscopic image-based diagnosis. We hope that our findings will contribute to future research in this field.

5.1. Conclusions

Cancer is a major health issue that is also one of the leading causes of death around the world. GC is one of the most common types of cancer. Because many factors influence GC, identifying the most important factors is necessary. On the other hand, reducing the number of factors by feature selection methods can increase the performance of predicting models.

The factors affecting GC were identified based on personal lifestyle, according to the methods used including filter, wrapper, and filter-wrapper methods. Then, 4 classifiers were created, using the feature selection methods. The results revealed that the developed filter-wrapper method and GBDT classifier outperformed the higher performance than other classifiers and feature selection methods. As a result, physicians can use this model as a decision support system (DSS) to make preliminary identifying GC risk factors. Further, by developing predictive models, they can predict GC probability based on factors related to people's lifestyles.

5.2. Introducing the Tool

We used Python as the programming language to develop the proposed GBDT classifier and filter-wrapper method. Python is a popular choice for machine learning due to its simplicity, versatility, and rich libraries that offer a wide range of functionalities for machine learning applications. We utilized Jupyter Notebook, an open-source web application, to generate and distribute documents that incorporate live code, equations, visualizations, and narrative text. This allowed us to organize and document our work effectively and share it with other researchers in a reproducible manner.

Additionally, we utilized SciKit-learn, a widely-used Python library for machine learning, which offers effective tools for data mining and analysis. In the implementation, we modified some of the default parameters in the classifiers to obtain the best possible results for our dataset. We also utilized various Python libraries such as pandas, Numpy, and Matplotlib for data processing, manipulation, and visualization.

We acknowledge that the codes used in this study are available from the corresponding author upon reasonable request.

Supplementary Material

Supplementary material(s) is available here [To read supplementary materials, please refer to the journal website and open PDF/HTML].

Acknowledgments

The authors appreciate the assistance of the Cancer Research Center of SBMU, especially Professor Mohammad Esmaeil Akbari (Head of the Center) in collecting information on patients with GC.

Footnotes

Authors' Contribution: H.M.: Software, formal analysis, investigation, methodology, writing original draft. R.R.: Conceptualization, methodology, writing review and editing, supervision. M-R.S.: Conceptualization, methodology, writing review and editing, supervision. B.S.D: Validation, data curation. M.A.A.K.: Validation.

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

Data Reproducibility: The dataset presented in the study is available on request from the corresponding author during submission or after publication.

Ethical Approval: This study is extracted from the Ph.D thesis entitled "Designing an intelligent model in predicting the pattern of gastric cancer in Iran". This research has been approved by Council No. 36 of the Vice-Chancellor in Research Affairs of Islamic Azad University, Science and Research Branch. Then, the 98th committee of Vice-Chancellor in Research Affairs Shahid Beheshti University of Medical Sciences approved it.

Funding/Support: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Informed Consent: Written consent was received from all patients.

References

1. Xiao Y, Wu J, Lin Z, Zhao X. A deep learning-based multi-model ensemble method for cancer prediction. *Comput Methods Programs Biomed*. 2018;**153**:1–9. [PubMed ID: 29157442]. https://doi.org/10.1016/j.cmpb.2017.09.005.
2. Rawla P, Barsouk A. Epidemiology of gastric cancer: global trends, risk factors and prevention. *Prz Gastroenterol*. 2019;**14**(1):26–38. [PubMed ID: 30944675]. [PubMed Central ID: PMC6444111]. https://doi.org/10.5114/pg.2018.80001.
3. De Vuysere S, Vandecaveye V, De Bruecker Y, Carton S, Vermeiren K, Tollens T, et al. Accuracy of whole-body diffusion-weighted MRI (WB-DWI/MRI) in diagnosis, staging and follow-up of gastric cancer, in comparison to CT: A pilot study. *BMC Med Imaging*. 2021;**21**(1):18. [PubMed ID: 33546626]. [PubMed Central ID: PMC7866710]. https://doi.org/10.1186/s12880-021-00550-2.
4. Ghaderzadeh M, Sadoughi F, Ketabat A. A computer-aided detection system for automatic classification of prostate cancer from benign hyperplasia of prostate. *Front Health Inform*. 2013;**2**(2):1–5.
5. Cueto-Lopez N, Garcia-Ordas MT, Davila-Batista V, Moreno V, Aragones N, Alaiz-Rodriguez R. A comparative study on feature selection for a risk prediction model for colorectal cancer. *Comput Methods Programs Biomed*. 2019;**177**:219–29. [PubMed ID: 31319951]. https://doi.org/10.1016/j.cmpb.2019.06.001.
6. Jiménez F, Sánchez G, Palma J, Sciavicco G. Three-objective constrained evolutionary instance selection for classification: Wrapper and filter approaches. *Eng Appl Artif Intell*. 2022;**107**:104531. https://doi.org/10.1016/j.engappai.2021.104531.
7. Bolón-Canedo V, Alonso-Betanzos A. Ensembles for feature selection: A review and future trends. *Inf Fusion*. 2019;**52**:1–12. https://doi.org/10.1016/j.inffus.2018.11.008.
8. Singh N, Singh P. A hybrid ensemble-filter wrapper feature selection approach for medical data classification. *Chemom Intell Lab Syst*. 2021;**217**:104396. https://doi.org/10.1016/j.chemolab.2021.104396.
9. Chandrashekar G, Sahin F. A survey on feature selection methods. *Comput Electr Eng*. 2014;**40**(1):16–28. https://doi.org/10.1016/j.compeleceng.2013.11.024.
10. Li C, Zhang S, Zhang H, Pang L, Lam K, Hui C, et al. Using the K-nearest neighbor algorithm for the classification of lymph node metastasis in gastric cancer. *Comput Math Methods Med*. 2012;**2012**:876545. [PubMed ID: 23150740]. [PubMed Central ID: PMC3488413]. https://doi.org/10.1155/2012/876545.
11. Thara L, Gunasundari R. Adaptive feature selection method based on particle swarm optimization for gastric cancer prediction. *2017 2nd International Conference on Communication and Electronics Systems (ICCES)*. 2017Oct19-20. IEEE; 2017. p. 808–13.
12. Thara L, Gunasundari R. Swarm Intelligence Based Feature Selection Algorithms and Classifiers for Gastric Cancer Prediction. *International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI) 2018*. Springer International Publishing; 2019. p. 1194–201.
13. Qi Y, Su B, Lin X, Zhou H. A new feature selection method based on feature distinguishing ability and network influence. *J Biomed Inform*. 2022;**128**:104048. [PubMed ID: 35248795]. https://doi.org/10.1016/j.jbi.2022.104048.
14. Saeed F, Al-Sarem M, Al-Mohaimeed M, Emara A, Boulila W, Alasli M, et al. Enhancing Parkinson's Disease Prediction Using Machine Learning and Feature Selection Methods. *Comput Mater Contin*. 2022;**71**(3):5639–58. https://doi.org/10.32604/cmc.2022.023124.
15. Got A, Moussaoui A, Zouache D. Hybrid filter-wrapper feature selection using whale optimization algorithm: A multi-objective approach. *Expert Syst Appl*. 2021;**183**:115312. https://doi.org/10.1016/j.eswa.2021.115312.
16. Mandal M, Singh PK, Ijaz MF, Shafi J, Sarkar R. A Tri-Stage Wrapper-Filter Feature Selection Framework for Disease Classification. *Sensors (Basel)*. 2021;**21**(16). [PubMed ID: 34451013]. [PubMed Central ID: PMC8402295]. https://doi.org/10.3390/s21165571.
17. Afrash MR, Shafiee M, Kazemi-Arpanahi H. Establishing machine learning models to predict the early risk of gastric cancer based on lifestyle factors. *BMC Gastroenterol*. 2023;**23**(1):6. [PubMed ID: 36627564]. [PubMed Central ID: PMC9832798]. https://doi.org/10.1186/s12876-022-02626-x.
18. Jia W, Sun M, Lian J, Hou S. Feature dimensionality reduction: A review. *Complex Intell Syst*. 2022;**8**(3):2663–93. https://doi.org/10.1007/s40747-021-00637-x.
19. Li J, Cheng K, Wang S, Morstatter F, Trevino RP, Tang J, et al. Feature Selection. *ACM Computing Surveys*. 2017;**50**(6):1–45. https://doi.org/10.1145/3136625.
20. Kumar V. Feature Selection: A literature Review. *Smart Comput Rev*. 2014;**4**(3). https://doi.org/10.6029/smartcr.2014.03.007.
21. Slimani K, Bouchlaghem Y, Akhiat Y, Amjad S, Gerasymov O, Ait Kbir M, et al. Feature Selection: A Review and Comparative Study. *E3S Web Conf*. 2022;**351**:1046. https://doi.org/10.1051/e3sconf/202235101046.

22. Biglarian A, Hajizadeh E, Kazemnejad A, Zayeri F. Determining of prognostic factors in gastric cancer patients using artificial neural networks. *Asian Pac J Cancer Prev*. 2010;**11**(2):533–6. [PubMed ID: 20843146].
23. Feng QX, Liu C, Qi L, Sun SW, Song Y, Yang G, et al. An Intelligent Clinical Decision Support System for Preoperative Prediction of Lymph Node Metastasis in Gastric Cancer. *J Am Coll Radiol*. 2019;**16**(7):952–60. [PubMed ID: 30733162]. https://doi.org/10.1016/j.jacr.2018.12.017.
24. Zhu Y, Wang QC, Xu MD, Zhang Z, Cheng J, Zhong YS, et al. Application of convolutional neural network in the diagnosis of the invasion depth of gastric cancer based on conventional endoscopy. *Gastrointest Endosc*. 2019;**89**(4):806–815 e1. [PubMed ID: 30452913]. https://doi.org/10.1016/j.gie.2018.11.011.
25. Wu L, Wang J, He X, Zhu Y, Jiang X, Chen Y, et al. Deep learning system compared with expert endoscopists in predicting early gastric cancer and its invasion depth and differentiation status (with videos). *Gastrointest Endosc*. 2022;**95**(1):92–104 e3. [PubMed ID: 34245752]. https://doi.org/10.1016/j.gie.2021.06.033.
26. Taninaga J, Nishiyama Y, Fujibayashi K, Gunji T, Sasabe N, Iijima K, et al. Prediction of future gastric cancer risk using a machine learning algorithm and comprehensive medical check-up data: A case-control study. *Sci Rep*. 2019;**9**(1):12384. [PubMed ID: 31455831]. [PubMed Central ID: PMC6712020]. https://doi.org/10.1038/s41598-019-48769-y.
27. Amirgaliyev Y, Shamiluulu S, Merembayev T, Yedilkhan D. Using Machine Learning Algorithm for Diagnosis of Stomach Disorders. *18th International Conference on Mathematical Optimization Theory and Operations Research (MOTOR 2019)*. Ekaterinburg, Russia. Springer; 2019. p. 343–55.
28. Mortezagholi A, Khosravizadeh O, Menhaj MB, Shafigh Y, Kalhor R. Make Intelligent of Gastric Cancer Diagnosis Error in Qazvin's Medical Centers: Using Data Mining Method. *Asian Pac J Cancer Prev*. 2019;**20**(9):2607–10. [PubMed ID: 31554353]. [PubMed Central ID: PMC6976843]. https://doi.org/10.31557/APJCP.2019.20.9.2607.