Cukurova Medical Journal
ÇUKUROVA ÜNİVERSİTESİ TIP FAKÜLTESİ

Cukurova Med J 2022;47(3):1179-1186
DOI: 10.17826/cumj.112853

ARAŞTIRMA / RESEARCH

Classification of colorectal cancer based on gene sequencing data with XGBoost model: An application of public health informatics

XGBoost modeli ile gen dizileme verilerine dayalı kolorektal kanserin sınıflandırılması: Bir halk sağlığı bilişimi uygulaması

Sami Akbulut[1,2,3], Zeynep Küçükakçalı[2], Cemil Çolak[3]

[1]Department of Surgery, Inonu University Faculty of Medicine, Malatya, Turkey
[2]Department of Biostatistics and Medical Informatics, Inonu University Faculty of Medicine, Malatya, Turkey
[3]Department of Public Health, Inonu University Faculty of Medicine, Malatya, Turkey

Cukurova Medical Journal 2022;47(3):1179-1186

Abstract

Purpose: This study aims to classify open-access colorectal cancer gene data and identify essential genes with the XGBoost method, a machine learning method.
Materials and Methods: The open-access colorectal cancer gene dataset was used in the study. The dataset included gene sequencing results of 10 mucosae from healthy controls and the colonic mucosa of 12 patients with colorectal cancer. XGboost, one of the machine learning methods, was used to classify the disease. Accuracy, balanced accuracy, sensitivity, selectivity, positive predictive value, and negative predictive value performance metrics were evaluated for model performance.
Results: According to the variable selection method, 17 genes were selected, and modeling was performed with these input variables. Accuracy, balanced accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and F1 score obtained from modeling results were 95.5%, 95.8%, 91.7%, 1%, 1%, and 90.9%, and 95.7%, respectively. According to the variable impotance acquired from the XGboost technique results, the CYR61, NR4A, FOSB, and NR4A2 genes can be employed as biomarkers for colorectal cancer.
Conclusion: As a consequence of this research, genes that may be linked to colorectal cancer and genetic biomarkers for the illness were identified. In the future, the detected genes' reliability can be verified, therapeutic procedures can be established based on these genes, and their usefulness in clinical practice may be documented.
Keywords:. Colorectal cancer, genomics, machine learning, XGboost model

Öz

Amaç: Bu çalışma, bir makine öğrenmesi yöntemi olan XGBoost yöntemi ile açık erişimli kolorektal kanser gen verilerini sınıflandırmayı ve temel genleri tanımlamayı amaçlamaktadır.
Gereç ve Yöntem: Çalışmada açık erişimli kolorektal kanser gen veri seti kullanıldı. Veri seti, sağlıklı kontrollerden 10 mukozanın ve kolorektal kanserli 12 hastanın kolon mukozasının gen dizileme sonuçlarını içeriyordu. Hastalığı sınıflandırmak için makine öğrenmesi yöntemlerinden biri olan XGboost kullanıldı. Model performansı için doğruluk, dengelenmiş doğruluk, duyarlılık, seçicilik, pozitif tahmin değeri ve negatif tahmin değeri performans metrikleri değerlendirildi.
Bulgular: Değişken seçim yöntemine göre 17 gen seçilmiş ve bu girdi değişkenleri ile modelleme yapılmıştır. Modelleme sonuçlarından elde edilen doğruluk, dengeli doğruluk, duyarlılık, özgüllük, pozitif tahmin değeri, negatif tahmin değeri ve F1 puanı sırasıyla %95.5, %95.8, %91.7, %1, %1 ve %90.9 ve %95.7 idi. XGboost tekniği sonucundan elde edilen değişken önemliliklerine göre, CYR61, NR4A, FOSB ve NR4A2 genleri kolorektal kanser için biyolojik belirteçler olarak kullanılabilir.
Sonuç: Bu araştırma sonucunda kolorektal kanserle bağlantılı olabilecek genlerin yanı sıra hastalığa yönelik genetik biyobelirteçler de belirlendi. Gelecekte, tespit edilen genlerin güvenilirliği doğrulanabilir, bu genlere dayalı olarak terapötik prosedürler oluşturulabilir ve klinik pratikteki yararları belgelenebilir.

Anahtar kelimeler: Kolorektal kanser, genomik, makine öğrenimi, XGboost modeli

Yazışma Adresi/Address for Correspondence: Dr. Sami Akbulut, Department of Surgery, Department of Public Health, Department Biostatistics and Medical Informatics Inonu University Faculty of Medicine, Malatya, Turkey. E-mail: akbulutsami@gmail.com
Geliş tarihi/Received: 10.06.2022 Kabul tarihi/Accepted: 25.07.2022

INTRODUCTION

Cancer is a heterogeneous group of diseases, also defined as a malignant neoplasm characterized by a clonal increase of abnormal cells. Another definition of cancer is an increase in the number of abnormal cells that grow excessively, multiply beyond the area it is in, and is likely to spread to the junction points and other organs of the body, defined as "metastasis".

Cancers are most commonly caused by epithelial cells that coat the body's surfaces. Colon cancers (CRC) are carcinomas with an adenocarcinoma form that arise from epithelial cells in the colorectal mucosa[1]. Colorectal cancers are among the most common cancer-related disorders and causes of death. This form of cancer is one of the leading causes of mortality worldwide[2]. Colorectal cancer is the third most common disease in both men and women worldwide, with more than a million new cases recorded every year[3]. The highest incidences have been reported in Asia, America, Europe, Australia, and New Zealand from developing countries where a generally western diet and lifestyle are common. Underdeveloped and crowded countries such as India have the lowest incidence[4]. The incidence of these cancers increases significantly with age and can generally be considered a disease of advanced age[5]. According to United States data, the five-year survival of colorectal cancer is 65%. Between 2000 and 2016, there was a 34% reduction in colorectal cancer-related mortality with cancer drug approvals[6].

Genomic technology, which processes and stores its outputs utilizing information technologies, is a science developed by advances in automation and bioinformatics. With the correct setup of genomic technology, research can be carried out in almost every field of medicine (Oncology, Pharmacology, Immunology, Biochemistry, Microbiology, etc.)[7]. Comparative studies provide opportunities for research such as cancerization and prediction of prognosis, prediction of drug response and personalized drug development, nature of the immune response, and even prediction of transplantation outcome[8].

Recent improvements in the analysis of genetic changes in cancer research and clinical application have been made possible by next-generation sequencing (NGS)[9]. Colorectal cancer develops due to a build-up of genetic (gene mutations, gene amplification, and so on) and epigenetic (abnormal DNA methylation, chromatin changes, and so forth) alterations that transform colonic epithelial cells into colon adenocarcinoma cells. Understanding the causes and roles of genomic and epigenomic instability in colon tumor formation can lead to more effective colorectal cancer prevention and treatment strategies[10]. For this reason, many genome-wide studies of colorectal cancer have been carried out, and gene expression data have been examined.

While public health informatics shares many characteristics with other information technology fields, it is distinct in several respects. It is essential to focus on applications of information science and technology that benefit populations rather than individuals, to prevent disease instead of treating it, and to intervene at all vulnerable points in the causal chains that lead to illness, injury, or disability in order to take preventive action, all of which can be performed within a governmental rather than a private context [11].

Machine learning (ML) is a subfield of artificial intelligence that uses data-driven learning to make predictions about new data when exposed to new data. The goal of researchers is to teach computers to detect complex patterns and make data-driven decisions[12]. ML methods have achieved high performance in many situations over the last decade, thanks to the availability of large datasets and increased computing power[13]. ML methods are one of the technologies that have seen widespread use in disease diagnosis and clinical decision support systems in recent years, and they have a wide range of applications. ML methods are typically used to classify disease prediction[14,15]. ML, which has a wide range of applications in the field of health, is the foundation of applications in the determination of genetic diseases, early detection of cancer diseases, and pattern recognition in medical imaging[16]. Extreme Gradient Boosting is an ML algorithm based on gradient augmented method and decision trees, which has become increasingly popular in both data science and remote sensing fields with its high classification performance[17,18]. The main reason for this method's success is its objective function in the learning process. It consists of the objective function, loss function, and regularization terms. The loss function calculates the difference of each predicted class value made by the model from the actual value[19,20]. The hypothesis of this study is whether the proposed the XGboost method can classify

colorectal cancer based on the open-access gene data and identify the fundamental genes associated with the disease.

MATERIALS AND METHODS

Dataset

In the study, the XGboost method, one of the ML methods, was applied to the open-access colorectal cancer gene dataset. The open-access dataset was obtained from https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE4107. The data set includes gene sequencing results of 10 mucosae from healthy controls and colonic mucosa of 12 patients with colorectal cancer.

Feature selection

Featureselection is an essential step in any predictive modeling project. This operation is also referred to as "Feature Selection". One of the most critical steps in developing a statistical model is deciding what data to include. High efficiency can be achieved by identifying a data set's most useful valuable properties before working with huge data sets and models with high computational costs. Feature selection is the process of defining features in a data set that affect the dependent variable. The explanatory variables' high dimensionality can result in both long computation times and the risk of over-learning the data.

Furthermore, the models with many features are difficult to interpret. Before performing statistical modeling, important features should ideally be selected[21]. Most ML and data mining methods may be ineffective when dealing with high-dimensional data. As a result, when the dimensionality is reduced, these methods produce more effective results[22]. Data sets on gene expression are quite large. Because of the large size of gene expression data sets, modeling analyses take a long time, and these data sets may result in computational inefficiency in the analysis. The model's performance may suffer due to the high dimensionality issue. A large number of genes in gene expression data sets can also cause a classification algorithm to overfit the training examples and undergeneralize new samples. In this study, LASSO, one of the feature selection methods, was used to solve these problems. Robert Tibshirani developed the LASSO method in 1996. The LASSO method constrains the sum of the absolute values of the model parameters to be less than a fixed value (upper limit). To accomplish the process, the method employs a throttling process that penalizes the coefficients of the regression variables, causing some of them to drop to zero. It is beneficial when the data set contains few observations and many variables. Furthermore, LASSO improves model interpretability by removing irrelevant variables unrelated to the response variable, thereby eliminating the problem of over-learning[23].

XGBoost

Gradient Boosting is a powerful ML technique introduced by Friedman in 2001. Gradient Boosting is an ML technique for regression and classification problems that produce an ensemble of weak predictive models, usually decision trees, in a prediction model. Gradient Boosting is based on boosting techniques. Since it is based on the boosting method, it aims to construct a large number of weak learners in order and incorporate them into a complex model[24,25].

XGBoost is the abbreviation of the term Extreme Gradient Boosting, and its basic structure is based on gradient boosting and decision tree algorithms. Friedman developed the original version of the XGBoost algorithm in 2002[26]. It became trendy in the world of ML after it was presented as an article by Tianqi Chen and Carlos Guestrin, two researchers at the University of Washington, at the Special Interest Group Association for Information Discovery and Data Mining of Computing Machines 2016 conference[27].

XGBoost is a very popular algorithm used for health, energy, finance, etc., areas in the fields. Compared to other algorithms, it is in a very advantageous position regarding speed and performance. In addition, XGBoost is highly predictive, 10 times faster than different algorithms, and includes several several regularizations that improve overall performance and reduce overfitting or over-learning. Gradient boosting is an ensemble method combining weak classifiers with boosting to construct a strong classifier. The strong learner is trained iteratively, starting with a primary learner. Both gradients boosting and XGBoost follow the same principle. The main differences between them lie in the implementation details. By using different regularization techniques, XGBoost achieves better performance by controlling the complexity of the trees[28].

Modeling

XGBoost, one of the ML methods, was used in the modeling. Analyzes were carried out using the n-fold cross-validation method. In the n-fold cross-validation method, the data is first divided into n parts, and the model used is applied to n parts. One of the n parts is used for testing, while the other n-1 parts are utilized for training the model. The mean of the obtained values is evaluated for the cross-validation method. Accuracy, balanced accuracy, sensitivity, selectivity, positive predictive value, negative predictive value, and F1-score were used as performance evaluation criteria. The methods used in the modeling were made using the R programming language.

Study protocol

This study, which was prepared using the National Center for Biotechnology Information Gene Expression Omnibus open-access dataset involving human participants, was following the ethical standards of the institutional and national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards. Ethical approval was obtained from the Inonu University Institutional Review Board (IRB) for Non-Interventional Clinical Research (2022/3650). STROBE (Strengthening the reporting of observational studies in epidemiology) guideline was utilized to assess the likelihood of bias and overall quality for this study[29].

Statistical analysis

IBM SPSS (Statistical Package for the Social Sciences, Inc, Chicago, IL, USA) Statistics 25.0 program was used in the analysis. The number of patients required for the comparison of two independent groups consisting of patients with and without colorectal cancer was calculated using the G*Power 3.1 package program [alpha= 0.05, power (1-beta) =0.8, effect size=1.40, allocation ratio=1, alternative hypothesis (H1)= two tailed]. As a result of the analysis, the number of patients required in each group was determined as 10. The Shapiro Wilk normality test was used to determine whether the variables had a normal distribution. Data were given as median (minimum-maximum) and mean± standard deviation. The independent t-test was used for the analysis of normally distributed gene expression (MIER1, CYR61, HSPA1B///HSPA1A, DSC2, CYP2B7P///CYP2B6, FOS, FRYL, HOXA10-HOXA9///MIR196B///HOXA9, PTGR1), while the Mann-Whitney U test was used for the analysis of non-normally distributed genes (FOSBNR4A1, FOSB NR4A1, NR4A2, RSRP1, ADAMTS1, SFRP2, CCDC3). P-value <0.05 was considered statistically significant.

RESULTS

The data set consists of 10 male (45.5%) and 12 female (54.5%) patients. The mean age of the patients was 40.27 ± 7.66. There were 5 males and 7 females in the patient group and 5 males and 5 females in the control group. While the mean age of the patient group was 40.91 ± 8.31, the mean age of the control group was 39.5 ± 7.16.

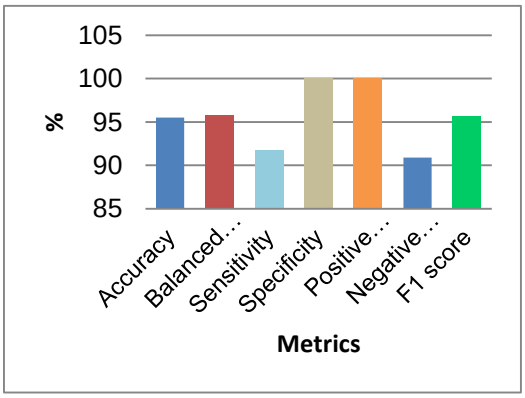

Figure 1. Graph of values for performance criteria obtained from XGboost models

In the current modeling stage, 17 genes remained in the data set obtained by applying the LASSO feature selection method to the data set of 54675 genes. The descriptions of the data set created with these genes and the descriptors of the examined target variable are presented in Table 1. The results of the performance metrics obtained according to the results of the XGboost model are given in Table 2.

Accuracy, balanced accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and F1 score obtained from the XGboost model were 95.5%, 95.8%, 91.7%, 1%, 1%, and 90.9%, and 95.7%, respectively. Figure 1 plots the values of performance criteria obtained from the XGboost model. Figure 2 shows the variable importance values of genes for the XGboost model

Table 1. Descriptive statistics for Input variables

Gene name	Prop number	Groups				p
		Colorectal Cancer		Control		
		Mean± SD	Median (Min-Max)	Mean± SD	Median (Min-Max)	
MIER1	1555105_a_at	341±87	361 (212-534)	666±191	629 (458-1111)	<0.001*
CYR61	201289_at	8645±4450	8332 (3440-18477)	136±59	113 (66-228)	<0.001*
NR4A1	202340_x_at	1656±1201	1305 (460-4350)	104±104	67 (16-355)	<0.001**
HSPA1B///HSPA1A	202581_at	655±259	615 (320-1223)	1995±692	1807 (917-3009)	<0.001*
FOSB	202768_at	3229±4715	1814 (529-17762)	18±10	13 (10-41)	<0.001**
NR4A2	204621_s_at	870±796	565 (347-2921)	137±64	144 (26-232)	<0.001**
DSC2	204750_s_at	532±301	513 (43.7-916)	1542±469	1469 (989-2380)	<0.001*
CYP2B7P///CYP2B6	206754_s_at	397±243	360 (10.7-981)	1905±833	1787 (685-3287)	<0.001*
RSRP1	209006_s_at	512±335	417 (214-1432)	1655±553	1643 (851-2646)	<0.001**
FOS	209189_at	15568±7750	12793 (3883-33539)	422±220	358 (164-788)	<0.001*
FRYL	212548_s_at	4854±1409	5246 (2142-6794)	8108±1676	8182 (5040-11260)	<0.001*
HOXA10-HOXA9///MIR196B///HOXA9	214651_s_at	986±284	914 (436-1474)	2134±437	2162 (1576-2714)	<0.001*
ADAMTS1	222486_s_at	774±624	525 (238-2362)	57±51	25 (10-137)	<0.001**
SFRP2	223122_s_at	1999±2225	1463 (305-8190)	46±58	18 (3-161)	<0.001**
CCDC3	223316_at	913±689	741 (289-2796)	44±56	21 (15-199)	<0.001**
PTGR1	231897_at	1167±528	1028 (507-2197)	2996±671	2818 (2184-4022)	<0.001*
	233932_at	190±84	154 (93-335)	53.55±51	35 (3-136)	<0.001*

*: Independent sample t-test; **: Mann Whitney U test; Min: minimum; Max: maximum; SD: Standard deviation

Table 2. Values for the metrics of the classification performance of the XGboost model

Metric	Value (%)
Accuracy	95.5
Balanced Accuracy	95.8
Sensitivity	91.7
Specificity	100
Positive predictive value	100
Negative predictive value	90.9
F1 score	95.7

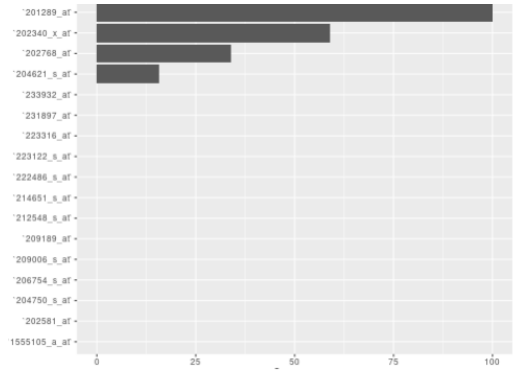

Figure 2. The graphic of variable importance values for the XGboost model

DISCUSSION

Public health informatics is described as the systematic application of information to public health practice through computer science and technology. Surveillance, prevention, preparedness, health promotion, research, and learning are all public health practices that use informatics. As the contemporary era has progressed, computer technology has become an indispensable instrument for improving the usage of public health surveillance. Although informatics has been widely used in many industries, its application in public health has been limited in the literature [30]. The current study presents the classification of colorectal cancer based on gene sequencing data with the Xgboost model regarding informatics systems in public health.

In recent years, the global prevalence of CRC has increased at an alarming rate. In 2020, an estimated 1.93 million new CRC cases will be diagnosed and 0.94 million CRC-caused deaths worldwide, accounting for 10% of global cancer incidence (total 19.29 million new cases) and 9.4% of all cancer-caused deaths (total 9.96 million deaths). According to estimates, CRC is the third highest cause of cancer-related deaths in both men and women worldwide in 2020, with an anticipated 515,637 male and 419,536 female fatalities in 2020 [31]. Great efforts and advances have been made to better understand the pathophysiology of CRC. Endoscopic resection, local surgical excision, targeted therapy, radiation therapy, ablative treatments, chemotherapy, immunotherapy, and genomic studies of the disease have increased the overall survival of the disease [31,32]. The prevalence of CRC varies by country. Hungary, Slovakia, Norway, the Netherlands, and Denmark had the highest age-standardized incidence rates in 2020, with rates of 45.3, 43.9, 41.9, 41.0, and 40.9 cases per 100,000 people, respectively. Guinea, Gambia, Bangladesh, Bhutan, and Burkina Faso had the lowest age-standardized incidence rates in 2020, with 3.3, 3.7, 3.8, and 3.8 cases per 100,000 people, respectively. In 2020, China and the United States had the most significant estimated number of new CRC cases, and the number of new cases is expected to rise steadily over the next 20 years due to demographic reasons. With its increasing incidence, large numbers of CRC cases pose a growing global public health problem [31]. Therefore, it is necessary to develop treatments by fully revealing the physiology of the disease.

The widespread use of next-generation sequencing (NGS) has led to a relatively clear understanding of the genomics of colorectal cancer. However, progress in using molecular biomarkers in standard practice for the disease has been slow. There is currently no approved targeted therapy for CRC based on a positive predictive marker used [32]. It is known that sequential genomic and epigenetic changes cause CRC, and many genomic studies are being conducted to identify biomarkers with the clinical benefit that may be associated with the disease, and more are needed.

In the dataset used in this study, samples taken from the colon mucosa of colorectal cancer patients and the colonic mucosa of the control group were examined in terms of genomics. Gene expression profiles were obtained from the samples obtained from the mucosa. As a result, 54675 genes were obtained for 22 samples (12 colorectal cancers, 10 controls). Gene expression datasets are pretty large, and due to their large size, modeling with these datasets can cause take a long time analysis and computational inefficiency in the analysis. Therefore, before modeling with the existing data set, the most important genes associated with the output variable were selected by the Lasso variable selection method. Seventeen genes were selected by this method. Moreover, these genes were used in modeling. The accuracy, balanced accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and F1 score metrics obtained with the XGboost model used were 95.5%, 95.8%, 91.7%, 100%, 100%, 90.9%, and 95.7%, respectively. According to these results, the disease was classified correctly. CYR61, NR4A, FOSB, and NR4A2 genes can be used as biomarkers for colorectal cancer according to the variable importanceobtained from the XGboost method result.

One study used ML technology to predict the risk of postoperative recurrence in stage IV colorectal cancer patients. The study used logistic regression, decision tree, GradientBoosting, and lightGBM as ML methods. According to the results obtained, GradientBoosting was the most successful method and determined five influential variables associated with the disease[33]. Another study predicted the risk factors associated with recurrence in patients with colorectal cancer by using five ML classification techniques, including support vector machine, random forest, multivariate adaptive regression splines, extreme learning machine, and extreme

gradient boosting[34]. The Cancer Genome Atlas (TCGA) network performed a comprehensive molecular characterization of 224 resected colon and rectal tumors, demonstrating similar patterns of genomic alterations in these tumors, identifying recurrent genomic alterations, and characterizing tumors. This was one of the most significant milestones in the fight against colorectal cancer[32,35]. One study reported that NR4A2 is abnormally expressed in colorectal cancer cells[36]. In another study, high expression of CYR61 was reported to be associated with poor prognosis in colorectal cancer[37]. In another study, Cyr61 expression was evaluated in 251 colorectal cancer patients, and 157 showed strong Cyr61 expression[38]. Another study reported that FOSB was regulated in normal and tumor tissues in colorectal cancer patients[39].

XGBoost model suggested that the CYR61, NR4A, FOSB, and NR4A2 genes were associated with colorectal cancer according to the degree of importance values. The current research had a few limitations. Primarily, the open-access data set was employed for the related classification task. The use of actual data, including more extensive patient data on colorectal cancer, may offer more clinically useful information. Secondly, the genes used for the classification of colorectal cancer in this study are limited to those in the open-source dataset, and examining different genes in future studies may provide more comprehensive clinical implications.

As a result, with this study, genes associated with colorectal cancer were determined, and genomic biomarkers for the disease were revealed. The reliability of the genes obtained with more comprehensive analyses in the future can be tested, treatment approaches can be developed based on these genes, and their usability in clinical practice can be detailed. Thus, the way to make individual-based treatments and immunotherapy approaches more prone to clinical practice will be possible from the clinical point of view.

Yazar Katkıları: Çalışma konsepti/Tasarımı: SA, ZK, CC; Veri toplama: ZK, CC; Veri analizi ve yorumlama: SA, ZK, CC; Yazı taslağı: ZK, SA; İçeriğin eleştirel incelenmesi: SA, ZK, CC; Son onay ve sorumluluk: SA, ZK, CC; Teknik ve malzeme desteği: -; Süpervizyon: SA, CC; Fon sağlama (mevcut ise): yok.
Etik Onay: Ethical committee approval was obtained from the Inonu University Institutional Review Board (IRB) for Non-Interventional Clinical Research (Approval number: 2022/3650).
Hakem Değerlendirmesi: Dış bağımsız.
Çıkar Çatışması: Yazarlar çıkar çatışması beyan etmemişlerdir.
Finansal Destek: Yazarlar bu çalışmanın maddi destek almadığını beyan etmişlerdir.
Yazarın Notu: Mevcut çalışma sırasında analiz edilen veri kümeleri makul talep üzerine ilgili yazardan temin edilebilir.
STROBE bildirimi: Yazarlar, öğelerin STROBE Bildirimi-kontrol listesini okudular ve makale, öğelerin STROBE Bildirimi-kontrol listesine göre hazırlandı ve revize edildi.
Author Contributions: Concept/Design: SA, ZK, CC; Data acquisition: ZK, CC; Data analysis and interpretation: SA, ZK, CC; Drafting manuscript: ZK, SA; Critical revision of manuscript: SA, ZK, CC; Final approval and accountability: SA, ZK, CC; Technical or material support: -; Supervision: SA, CC; Securing funding (if available): n/a.
Ethical Approval: Girişimsel Olmayan Klinik Araştırmalar için İnönü Üniversitesi Kurumsal İnceleme Kurulu'ndan (IRB) etik kurul onayı alınmıştır (Onay numarası: 2022/3650).
Peer-review: Externally peer-reviewed.
Conflict of Interest: Authors declared no conflict of interest.
Financial Disclosure: The authors declared that this study received no financial support.
Acknowledgement: The datasets analyzed during the current study are available from the corresponding author upon reasonable request.
STROBE statement: The authors have read the STROBE Statement-checklist of items, and the manuscript was prepared and revised according to the STROBE Statement-checklist of items.

REFERENCES

1. Günther J, Seyfert H-M. The first line of defence: insights into mechanisms and relevance of phagocytosis in epithelial cells. Semin Immunopathol. 2018;40:555-65.
2. Cao W, Chen HD, Yu YW, Li N, Chen WQ. Changing profiles of cancer burden worldwide and in China: a secondary analysis of the global cancer statistics 2020. Chin Med J (Engl). 2021;134:783-91.
3. Mattiuzzi C, Lippi G. Current cancer epidemiology. J Epidemiol Glob Health. 2019;9:217-22.
4. Sharma R. An examination of colorectal cancer burden by socioeconomic status: evidence from GLOBOCAN 2018. EPMA J. 2020;11:95-117.
5. Abualkhair WH, Zhou M, Ahnen D, Yu Q, Wu XC, Karlitz JJ. Trends in incidence of early-onset colorectal cancer in the United States among those approaching screening age. JAMA Netw Open. 2020;3:e1920407.
6. MacEwan JP, Dennen S, Kee R, Ali F, Shafrin J, Batt K. Changes in mortality associated with cancer drug approvals in the United States from 2000 to 2016. J Med Econ. 2020;23:1558-69.
7. Del Boccio P, Urbani A. Homo sapiens proteomics: clinical perspectives. Ann Ist Super Sanita. 2005;41:479-82.
8. Martin DB, Nelson PS. From genomics to proteomics: techniques and applications in cancer research. Trends Cell Biol. 2001;11:60-5.
9. Gagan J, Van Allen EM. Next-generation sequencing to guide cancer therapy. Genome Med. 2015;7:80.
10. Grady WM, Carethers JM. Genomic and epigenetic instability in colorectal cancer pathogenesis. Gastroenterology. 2008;135:1079-99.
11. Magnuson JA, 'O'Carroll PW. Introduction to public health informatics. In Public Health Informatics and Information Systems (Eds Magnuson JA, Fu PC Jr): 3-18. Springer, London, 2014

12. Polikar R. Ensemble learning. In Ensemble Machine Learning (Eds Zhang C, Ma Y): 1-34. Springer, Boston, MA, 2012
13. Yagin FH, Yagin B, Arslan AK, Colak C. Comparison of performances of associative classification methods for cervical cancer prediction: Observational study. Turkiye Klinikleri J Biostat. 2021;13:266-72.
14. Akman M, Genç Y, Ankarali H. [Random forests methods and an application in health science]. Turkiye Klinikleri J Biostat. 2011;3:36-48.
15. Yılmaz R, Yagin FH. Early detection of coronary heart disease based on machine learning methods. Medical Records. 2022;1:1-6.
16. Witten IH, Frank E. Data mining: practical machine learning tools and techniques with Java implementations. Acm Sigmod Record. 2002;31:76-7.
17. Yagin FH, Cicek IB, Kucukakcali Z. Classification of stroke with gradient boosting tree using smote-based oversampling method. Medicine Science. 2021;10:1510-5.
18. Percin I, Yagin FH, Arslan AK, Colak C. An interactive web tool for classification problems based on machine learning algorithms using java programming language: data classification software. Proceedings of the 2019 3rd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT); 2019.IEEE:1-7.
19. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. Proceedings of the 22nd Acm Sigkdd International Conference On Knowledge Discovery And Data Mining. 2016;785-94.
20. Rumora L, Miler M, Medak D. Impact of various atmospheric corrections on sentinel-2 land cover classification accuracy using machine learning classifiers. SPRS Int J Geo-Inf. 2020;9:277.
21. Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. Bioinformatics. 2007;23:2507-17.
22. Fodor IK. A Survey of Dimension Reduction Techniques. Lawrence Livermore National Lab, CA, 2002.
23. Fonti V. Research Paper in Business Analytics: Feature Selection with LASSO. Amsterdam: VU Amsterdam 2017
24. Wang J, Li P, Ran R, Che Y, Zhou Y. A short-term photovoltaic power prediction model based on the gradient boost decision tree. Appl Sci. 2018;8:689.
25. Z.S P. Evaluating XGBoost for User Classification By Using Behavioral Features Extracted From Smartphone Sensors. KTH Royal Institute of Technology, School of Computer Science and Communication, Sweden, 2018.
26. Dikker J. Master thesis Boosted tree learning for balanced item recommendation in online retail. 2017.
27. Chen T. Guestrin, C. Xgboost: A scalable tree boosting system. Proceedings of the Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (New York, NY, USA, 2016), KDD '16, ACM. 2016:785-94.
28. Salam Patrous Z. Evaluating XGBoost for user classification by using behavioral features extracted from smartphone sensors (Masters thesis). Stockholm, KTH Royal Institute of Technology, 2018.
29. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP; STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Bull World Health Organ. 2007;85:867-72.
30. Amir PN, Sazali MF, Salvaraji L, Dulajis N, Rahim SSSA, Avoi R. Public health informatics in global health surveillance: a review: public health informatics. Borneo Epidemiology J. 2021;2:74-88.
31. Xi Y, Xu P. Global colorectal cancer burden in 2020 and projections to 2040. Transl Oncol. 2021;14:101174.
32. Mondaca S, Yaeger R. Colorectal cancer genomics and designing rational trials. Ann Transl Med. 2018;6:159.
33. Xu Y, Ju L, Tong J, Zhou C-M, Yang JJ. Machine learning algorithms for predicting the recurrence of stage IV colorectal cancer after tumor resection. Sci Rep. 2020;10:2519.
34. Ting WC, Chang HR, Chang CC, Lu CJ. Developing a novel machine learning-based classification scheme for predicting SPCs in colorectal cancer survivors. Appl Sci. 2020;10:1355.
35. Guinney J, Dienstmann R, Wang X, de Reynies A, Schlicker A, Soneson C et al. The consensus molecular subtypes of colorectal cancer. Nat Med. 2015;21:1350-6.
36. Rodriguez-Calvo R, Tajes M, Vazquez-Carrera M. The NR4A subfamily of nuclear receptors: potential new therapeutic targets for the treatment of inflammatory diseases. Expert Opin Ther Targets. 2017;21:291-304.
37. Xie L, Song X, Lin H, Chen Z, Li Q, Guo T et al. Aberrant activation of CYR61 enhancers in colorectal cancer development. J Exp Clin Cancer Res. 2019;38:213.
38. Jeong D, Heo S, Ahn TS, Lee S, Park S, Kim H et al. Cyr61 expression is associated with prognosis in patients with colorectal cancer. BMC Cancer. 2014;14:164.
39. Musella V, Verderio P, Reid JF, Pizzamiglio S, Gariboldi M, Callari M et al. Effects of warm ischemic time on gene expression profiling in colorectal cancer tissues and normal mucosa. PloS One. 2013;8:e53406.