DOI: 10.37990/medr.1077024 Med Records 2022;4(2):196-202

MEDICAL RECORDS-International Medical Journal

Research Article

Artificial Intelligence-based Colon Cancer Prediction by Identifying Genomic Biomarkers

Genomik Biyobelirteçleri Belirleyerek Yapay Zeka Tabanlı Kolon Kanseri Tahmini

Nur Paksoy, Fatma Hilal Yagin

Malatya Fahri Kayahan Healthcare Center, Department of Family Medicine Physician, Malatya, Turkey
Inonu University, Faculty of Medicine, Department of Biostatistics and Medical Informatics, Malatya, Turkey

Abstract

Aim: Colon cancer is the third most common type of cancer worldwide. Because of the poor prognosis and unclear preoperative staging, genetic biomarkers have become more important in the diagnosis and treatment of the disease. In this study, we aimed to determine the biomarker candidate genes for colon cancer and to develop a model that can predict colon cancer based on these genes.

Material and Methods: In the study, a dataset containing the expression levels of 2000 genes from 62 different samples (22 healthy and 40 tumor tissues) obtained by the Princeton University Gene Expression Project and shared in the figshare database was used. Data were summarized as mean ± standard deviation. Independent Samples T-Test was used for statistical analysis. The SMOTE method was applied before the feature selection to eliminate the class imbalance problem in the dataset. The 13 most important genes that may be associated with colon cancer were selected with the LASSO feature selection method. Random Forest (RF), Decision Tree (DT), and Gaussian Naive Bayes methods were used in the modeling phase.

Results: All 13 genes selected by LASSO had a statistically significant difference between normal and tumor samples. In the model created with RF, all the accuracy, specificity, f1-score, sensitivity, negative and positive predictive values were calculated as 1. The RF method offered the highest performance when compared to DT and Gaussian Naive Bayes.

Conclusion: In the study, we identified the genomic biomarkers of colon cancer and classified the disease with a high-performance model. According to our results, it can be recommended to use the LASSO+RF approach when modeling high-dimensional microarray data.

Keywords: Colon cancer, microarray, genomics, LASSO, random forest, decision tree, gaussian naive bayes

Öz

Amaç: Kolon kanseri dünya genelinde en sık görülen üçüncü kanser türüdür. Kötü prognoz ve net olmayan preoperatif evreleme nedeniyle, hastalığın tanı ve tedavisinde genetik biyobelirteçler daha önemli hale gelmiştir. Bu çalışmada kolon kanseri için biyobelirteç adayı genlerin belirlenmesi ve bu genlere dayalı olarak kolon kanserini başarılı bir şekilde tahmin eden bir modelin geliştirilmesi amaçlanmıştır.

Materyal ve Metot: Çalışmada, Princeton Üniversitesi Gen Ekspresyon Projesi ile elde edilen ve figshare veri tabanında paylaşılan 62 farklı örnekten (22 sağlıklı ve 40 tümör dokusu) 2000 genin ekspresyon düzeylerini içeren bir veri seti kullanıldı. Veriler ortalama ± standart sapma olarak özetlendi. İstatistiksel analizler için bağımsız örneklerde T-testi kullanıldı. Veri setindeki sınıf dengesizliği sorununu ortadan kaldırmak için öznitelik seçiminden önce SMOTE yöntemi uygulandı. Kolon kanseri ile ilişkili olabilecek en önemli 13 gen, LASSO öznitelik seçim yöntemi ile seçildi. Modelleme aşamasında Rastgele Orman (RF), Karar Ağacı (DT) ve Gauss naive Bayes yöntemleri kullanıldı.

Bulgular: LASSO tarafından seçilen 13 genin tümü, normal ve tümör numuneleri arasında istatistiksel olarak anlamlı bir farka sahipti. RF ile oluşturulan modelde doğruluk, seçicilik, f1-skor, duyarlılık, negatif ve pozitif prediktif değerlerinin tümü 1 olarak hesaplanmıştır. DT ve Gaussian Naive Bayes ile karşılaştırıldığında RF yöntemi en yüksek performansı vermiştir.

Sonuç: Çalışmada kolon kanserinin genomik biyobelirteçlerini belirledik ve hastalığı yüksek performanslı bir model ile sınıflandırdık. Elde ettiğimiz sonuçlara göre, yüksek boyutlu mikrodizi verilerinin modellenmesinde LASSO+RF yaklaşımının kullanılması önerilebilir.

Anahtar Kelimeler: Kolon kanseri, mikrodizi, genomik, LASSO, rastgele orman, karar ağacı, gaussian naive bayes

Received: 22.02.2022 **Accepted:** 28.03.2022

Corresponding Author: Fatma Hilal Yagin, Inonu University, Faculty of Medicine, Department of Biostatistics and Medical Informatics, Malatya, Turkey, **E-mail**: hilal.yagin@inonu.edu.tr

INTRODUCTION

According to the World Health Organization, cancer is the second leading cause of death after cardiovascular disease. Colon cancer ranks 3rd in the world in terms of incidence and is the 4th most common cancer. With the introduction of screening programs in the USA in the last 30 years, an improvement in cancer prognosis has been detected thanks to early diagnosis, and this screening program has been implemented in our country since 2009 (1,2).

Being able to perform postoperative staging and determining prognosis by staging alone emphasizes biomarkers and genetic evaluation in colon cancer. For this reason, examination of colon cancer based on genetic biomarkers is very important in the diagnosis and treatment of the disease (3).

Microarray technology has allowed the simultaneous measurement of thousands of gene expressions. Identifying disease-related biomarker candidate genes using microarray gene expression datasets and distinguishing (classifying) disease samples from non-disease samples has been an important research topic in biomedicine and medicine. However, the resulting large-scale datasets created many barriers to computational techniques. The high dimensionality problem affects most microarray gene expression datasets where dimensionality is high (up to tens of thousands of genes) and sample size is small (normally up to hundreds). Also, the high noise-to-variability ratio of microarray trials adds to the difficulties (4).

Machine learning methods are frequently used to overcome current challenges. Machine learning; it can be defined as obtaining previously unknown, valid and applicable information from data stacks through a dynamic process. In this process, many techniques such as clustering, data summarization, learning classification rules, finding dependency networks, developing predictive models, variability analysis and anomaly detection are used. With machine learning, confidential information is retrieved in database systems comprising large data stacks. This process is done using statistics, mathematical disciplines, modeling techniques, database technology and various computer programs (5,6).

Before constructing classification models in machine learning with high-dimensional microarray datasets, it is an important step to remove disease-related genes from the dataset using trait (gene) selection methods. In this way, both biomarker candidate genes can be selected and the performance of the classification models to be created will be improved (7).

In this study, we aimed to determine biomarker candidate genes for colon cancer by using gene expression dataset and to develop a classification model that can provide clinical decision support to healthcare professionals.

MATERIAL AND METHOD

Dataset

In this study, an open source colon cancer gene expression dataset obtained by Princeton University Gene Expression Project and shared in figshare database (https://figshare.com/articles/dataset/The_microarray_dataset_of_colon_cancer_in_csv_format_/13658790/1) was used (8). The dataset includes expression levels of 2000 genes from 62 different samples (22 healthy and 40 tumor tissues).

Statistical Evaluation

Data were summarized as mean ± standard deviation. Compliance with the normal distribution was done with the Kolmogorov-Smirnov test. Independent Samples T Test was used for statistical analysis. Statistical tests with a p value of less than 5% were considered significant. All statistical analyzes were performed using IBM SPSS Statistics for Windows version 26.0 (New York, USA).

Data Preprocessing and Modeling

In datasets with class imbalance problem, most machine learning techniques ignore minority class performance and therefore underperform in minority class. One approach to these datasets is to oversample the minority class and is called the Synthetic Minority Oversampling Technique, or SMOTE for short (9). In order to eliminate the class imbalance problem in the colon cancer gene expression dataset (22 normal and 40 tumor tissues), the SMOTE method was applied before feature selection. In this way, the number of samples in the groups, 40 normal and 40 tumor tissues, was equalized.

Afterwards, the 13 most important genes that may be associated with colon cancer were selected with the LASSO feature selection method. For the generalizability of the model, 80% of the data set is divided as the training set and 20% as the test set. Random Forest, Decision Trees and Gaussian Naive Bayes classification methods were used to predict colon cancer based on selected genes. The performance of the models was evaluated with accuracy, specificity, sensitivity, f1-score, negative predictive value and positive predictive value.

LASSO Feature Selection

In 1996 the LASSO method was first used by Robert Tibshirani. Regularization and property selection are the two main tasks of the method. The LASSO method puts a constraint on the sum of the absolute values of model parameters; the sum must be less than a fixed value (upper bound). To do this, the method implements a narrowing (regularization) process in which regression variables punish their coefficients, some of which reduce them to zero. During the property selection process, variables that still have a coefficient of zero after collapse are selected for the model. This operation minimizes the prediction error. In practice, the parameter that controls the power of punishment, is of great importance. When large enough, the dimensionality is can be reduced in this manner. The larger the parameter, the more coefficients are reduced to zero. There are many advantages to using the LASSO method. First, it can provide very good forecast accuracy, since the reduction and removal of coefficients can reduce variance without a significant increase in deviation. It is especially useful when there are few observations and many variables in the data set. LASSO also helps to improve the interpretability of the model by eliminating irrelevant variables that are not associated with the response variable, so that the problem of overlearning can also be addressed (10,11).

Random Forest

The Random Forests algorithm, a community learning method, aims to increase the classification value by generating multiple decision trees during the classification process. Because it includes random sampling and improved properties of techniques in community methods, the RF method offers better generalizations and makes more valid predictions than conventional machine learning methods. The reasons for the precise estimates of the RF method are that it gives low deviation and low correlation between trees. The low amount of deviation is obtained as a result of the creation of rather large trees. By creating as many different trees as possible, a low correlation structure is achieved. Individually created classification and regression decision trees come together to form the decision forest community. The decision trees here are randomly selected subsets from the data set to which they are connected. The results obtained during the formation of the decision forest are combined to make the latest prediction. For classification, trees each leaf node is created to contain only members of one class. For regression, trees continue to divide until a small number of units remain in the leaf node (12).

Decision Trees

Decision trees (DT) consist of root nodes, branches and leaves. The leaves in the decision trees are the places where the classification occurs and the branches refer to the result. The tree is created by the division variation method from the root node to the leaf nodes. A decision node can contain one or more branches. A decision tree can consist of both categorical and numerical data. The decision tree contains two basic process steps. These operations are splitting and pruning operations (13). The most important step when creating a DT is to decide which attribute values to base it on and which branching to create. In the knowledge gain and gain ratio approach that includes entropy rules, all attributes at hand are tested subjectically and the attribute with the highest knowledge gain is selected for branching. DT are a classification method that creates a model in the form of a tree structure consisting of decision nodes and leaf nodes by classification, property, and target. The decision tree algorithm is developed by dividing the data set into smaller pieces (14,15).

Gaussian Naive Bayes

A simple structured classification based on conditional probability, which is assumed to be equal and independent of each other in the classification of all attributes based on conditional probability. The classification process is done by combining the effects of different attributes on the result. Naive Bayes classifies using statistical methods and is an important algorithm in terms of performance. The importance of qualifications is considered equal in all. The Gaussian Naive Bayes (GNB) classifier is the Naive Bayes method, which is created by assuming that the class label is a Gaussian distribution on the given property values. GNB assigns all data to the closest location. However, instead of using Euclidean distance to calculate the distance between them, it calculates by taking into account the distance from the average and the class variance (16).

RESULTS

Table I contains descriptive statistics for 13 genes selected by LASSO trait selection. When Table I is examined; all 13 genes selected by LASSO had a statistically significant difference between normal and tumor samples. Hsa.8125, Hsa.2710, Hsa.8147, Hsa.36689, Hsa.31933, Hsa.1387 and Hsa.865 were expressed lower in tumor samples, while Hsa.3306, Hsa.22762, Hsa.3016, Hsa.5392, Hsa.1410 and Hsa.2928 were expressed higher in tumor samples.

Table 1. Descriptive statistics for selected genes

Gene Name	Normal (Mean ± SD)	Tumor (Mean ± SD)	t value	p-value
Hsa.8125	2.144 ± 0.496	1.444 ± 0.442	6.87	<0.001
Hsa.2710	1.289 ± 0.392	0.89 ± 0.359	5.3	<0.001
Hsa.8147	2.092 ± 0.799	0.725 ± 0.637	9.97	<0.001
Hsa.36689	0.741 ± 0.42	-0.01 ± 0.318	9.83	<0.001
Hsa.3306	0.289 ± 0.504	1.138 ± 0.482	-8.07	<0.001
Hsa.22762	-0.242 ± 0.564	0.337 ± 0.759	-4.05	0.003
Hsa.31933	-0.107 ± 0.263	-0.475 ± 0.377	5.24	<0.001
Hsa.3016	-0.222 ± 1.074	0.962 ± 1.049	-5.16	<0.001
Hsa.5392	-1.064 ± 0.655	-0.486 ± 0.486	-4.82	<0.001
Hsa.1410	-0.794 ± 0.73	0.002 ± 0.694	-5.53	<0.001
Hsa.2928	-1.312 ± 0.563	-0.526 ± 0.518	-6.98	<0.001
Hsa.1387	0.827 ± 0.648	0.017 ± 0.779	5.4	<0.001
Hsa.865	0.45 ± 0.393	0.06 ± 0.568	3.78	0.006

SD: Standard deviation

Table II presents the results of the performance measures of the RF, DT, and GNB classification models. Specificity, accuracy, f1-score, sensitivity, negative and positive predictive value criteria obtained from the RF model were all calculated as 1. That is, the RF model correctly predicted all samples in the test set. From the DT model, all performance measures were obtained as 0.9. Finally, in the model created with the GNB method, the performance measures were found to be accuracy 0.95, specificity 1, f1-score 0.95, sensitivity 0.9, negative predictive value 0.9091, and positive predictive value 1. The RF method offered the highest performance compared to DT and GNB.

Table 2. Performance measures results for classification models

Metric	Random Forest	Gaussian Naive Bayes	Decision Trees
Accuracy	1	0.95	0.9
Sensitivity	1	0.9	0.9
Specificity	1	1	0.9
PPV	1	1	0.9
NPV	1	0.9091	0.9
F1 score	1	0.95	0.9

PPV: Positive predictive value, NPV: Negative predictive value

DISCUSSION

Since knowing the biological functions of genes is useful for knowing the origin, causes and treatment of many diseases, studies in the field of genomics have been on the agenda of the scientific world for years. In addition to their biological functions, the detection and relationships of genes in the same biological pathway bring microarray studies to the fore. Thanks to the detection of possibly related genes, the detection and treatment of diseases has become easier with the identification of gene clusters (17). Based on this information, in the current study, we developed a model that can predict the disease by identifying the genes associated with colon cancer to provide clinical decision support to physicians.

In this study, we used the LASSO feature selection method to identify colon cancer-related genes. With the LASSO method, Hsa.8125, Hsa.36689, Hsa.3306, Hsa.3016, Hsa.8147, Hsa.2710, Hsa.22762, Hsa.31933, Hsa.5392, Hsa.1410, Hsa.2928, Hsa.1387 and Hsa.865 genes may be associated with colon cancer. Some of the biomarker candidate genes we identified were in agreement with the literature. Shaik et al. showed differential expression of Hsa.8125, Hsa.36689 and Hsa.3306 genes in colon cancer (genes1). In another study, Hsa.8125 and Hsa.3306 were among 100 genes associated with colon cancer (18). Hsa.8125; it is a gene that activates RNA binding activity, is involved in nucleocytoplasmic transport, is located in the endoplasmic reticulum, nucleus and perinuclear region of the cytoplasm. Yan et al. showed that this gene, also known as ANP32A, is overexpressed in colorectal cancer patients and ANP32A levels are higher in poorly differentiated tumors (19). Velmurugan et al. reported that this gene is associated with lymph node metastasis (20).

When the relationship between the Hsa.36689 gene, whose main task is guanylate cyclase activation in the colon, and colon cancer was examined, Yang et al. identified this gene among the top 5 most related genes (21). The Hsa.3016 and Hsa.8147 genes that we detected were also detected as the other genes with the highest frequency in this study.

The Hsa.3306 gene is a gene that plays a role in cell proliferation and is increased in cancer. In another study examining the colon gene data set in the literature, it was identified as one of the ten most closely related genes among 2000 genes due to its association with colon cancer. Among the genes detected in this study, Hsa.8125, one of the genes we detected, is also included. It has been shown that this gene, whose functions are important in the construction of intestinal villi, increases in normal cells and decreases in colon cancer cells (22).

Hsa.8147, also known as the desmin gene, is the gene responsible for the production of desmin, a smooth muscle-type intermediate filament protein expressed by smooth muscle cells, but also in fibrotic tissue in wound healing and tumor ‘desmoplastic’ stroma. Desmin also surrounds the vasculature by being produced by pericytes during angiogenesis in capillaries. It also plays a role in angiogenesis in cancer tissue. Studies have shown an increase in desmin expression in advanced cancer patients (23). In a study conducted in patients with gallbladder cancer, down-regulation of the desmin gene was detected (24).

The Hsa.3016 gene, which we have observed to be strongly associated with colon cancer, is one of the genes responsible for coding the S-100P protein. S100 proteins are involved in many events such as regulation of calcium homeostasis, cell proliferation, apoptosis, and cell migration. The S100 protein family plays a role in many stages of cancer formation and progression. S-100P acts as an inducer of metastasis, overexpression of S-100P increases the expression of S-100A6 and Cathepsin D, which are involved in cellular invasion. Furthermore, S100P promotes transendothelial migration of tumor cells (25).

The Hsa.2710 gene is one of the genes responsible for making Fibulin-1, a secreted glycoprotein that is included in the fibrillar extracellular matrix. It is involved in cell adhesion and migration along protein fibers within the extracellular matrix (ECM). Considered to have a role in cellular transformation and tumor invasion, it acts as a tumor suppressor (26). In the study of Xu et al. , it was shown that fibulin downregulation is associated with colorectal cancer (27).

Nucleolin; it is a multifunctional protein that is also found in the nucleolus, nucleoplasm, and cytoplasm. Hsa.22762 is one of the genes involved in the synthesis of nucleolin.

It is involved in the regulation of translation and stability of oncogenic mRNAs in the nucleoplasm. In our study, the presence of this gene was found to be significantly related in colon cancer patients. It has also been shown in other studies that nucleolin is overexpressed in many cancer types such as stomach, pancreatic, breast, cervix, prostate cancers, leukemias, melanomas and colorectal cancers (28).

The Hsa.31933 gene, which we detected in our study, is one of the genes that helps Autographa californica multiple nuclear polyhedrosis virus (AcMNPV), which is from the Baculovirus family, to successfully initiate the expression of viral genes by preparing the host environment and controlling the subsequent viral gene expression like other DNA viruses to infect their hosts. Viral genes, which are expressed immediately after infection, play a critical role in the early infection process; Hsa.31933 (Immediate-Early Regulatory Protein IE-N gene) is one of these genes. AcMNPV has been studied as a gene therapy vector. In a study by Ono et al., they determined AcMNPV induces antitumor acquired immunity; they showed AcMNPV can act as an effective immune-inducing virus and eukaryotic expression vector for gene carrier and has the potential to be a tumor therapy agent (29).

In another study, recombinant DNA obtained with this virus enabled the production of a natural antigen associated with carcinoma in mice (30). Although there are no studies related to this virus DNA in colon cancer yet, the data in our study showed that there is a strong relationship between colon cancer and this gene. We think that meaningful results can be obtained as a result of the use of AcMNPV as a vector with more comprehensive studies on the treatment of colon cancer.

The Hsa. 5392 gene is also known as ribosomal protein L24 (RPL24). It is one of the genes responsible for the expression of ribosomal proteins. It encodes the ribosomal protein L24, a homolog of the cytosolic RPL24 found in higher eukaryotes. Studies have been conducted on the overexpression of a number of ribosomal protein genes in human tumors and their contribution to tumorigenesis (31).

Hsa.1410 is the gene responsible for the synthesis of the eukaryotic translation initiation factor eIF-2. The role of protein synthesis changes is important in cancer development and progression. Studies show that ribosomal protein synthesis plays a direct role during tumor initiation. The translation initiation process is the rate-limiting step of protein synthesis in eukaryotes, and a group of eukaryotic translation initiation factors (eIFs) are involved. In previous studies, it has been shown that a significant increase in eIF3 subunits, eIF3A, eIF3B and eIF3M overexpression, which is one of the translation initiation factors, in colorectal cancer patients, and eIF4 subunits, of which eIF3C is an oncogene, are also increased in cancer cells (32).

In studies, eIF2a expression was described as transiently increased in normal cells, whereas constitutive overexpression indicated tumor initiation and progression. Golob-Schwarzl et al., they also showed that eIF2 is overexpressed in colorectal cancers (32).

Among the genes we determined, Hsa.2928 is the mRNA gene responsible for the expression of P-cadherin. Cadherins are calcium-dependent cell adhesion proteins that provide cell architecture and integrity, and their degradation is often associated with human cancer (33). Neo-expression or up-regulation of placental cadherin (P-cadherin) has been reported in a variety of carcinomas, including colorectal and bladder carcinomas (34).

The Hsa.1387 Human 11 beta-hydroxysteroid dehydrogenase type II mRNA gene is a gene that has a strong association with colon cancer and has been found to be associated with colon carcinomas. 11 beta-Hydroxysteroid dehydrogenase type II enzyme (11 beta HSD2), which is also located in the colon, which has an important role in water and electrolyte homeostasis, gives specificity to the mineralocorticoid receptor (35).

MAP kinases, also known as (ERKs) encoded by the Hsa.865 (ERK-1, M84490) gene, are regulated by extracellular signaling and act in a signal cascade that regulates various cellular processes such as proliferation, differentiation and cell cycle through the action of extracellular signals. The tumor suppressor pathway is stimulated by ERK-1 phosphorylation (36). The relationship between colon cancer and ERK-1 has been shown in many studies (37-39). In our study, we showed its relationship with colon cancer.

In a similar study using the same data set in the literature, PCA and PLS feature extraction methods were applied and then they classified colon cancer with the support vector machine method with an accuracy of 0.9516 (40). In another study, they found that the combined use of PSO and SVM outperformed the model created with only the SVM algorithm in terms of accuracy (0.94) and performance, and was faster in terms of time analysis (41). In the current study, three models were created using RF, DT and GNB classifiers based on biomarker candidate genes determined by LASSO feature selection method. According to the performance criteria obtained, the LASSO + RF model showed the best performance by correctly classifying all samples.

CONCLUSION

In conclusion, this study identified genomic biomarkers of colon cancer and classified the disease with a high-performance model. According to the results obtained, the LASSO method gave results compatible with the literature while determining the genomic biomarkers. For this reason, genes selected with LASSO can provide clinical decision support to physicians in the diagnosis and treatment of colon cancer. In addition, it can be suggested that the LASSO+RF approach be used in modeling high-dimensional data in medicine.

Financial disclosures: *The authors declared that this study hasn't received no financial support.*

***Conflict of Interest:** The authors declare that they have no competing interest.*

***Ethical approval:** Ethics committee approval is not required in this study.*

REFERENCES

1. Globocan W. Estimated cancer incidence, mortality and prevalence worldwide in 2012. Int Agency Res Cancer. 2012.
2. Labianca R, Beretta G, Gatta G, et al. Colon cancer. Critical Reviews Oncology Hematology. 2004;51:145-70.
3. Loboda A, Nebozhyn MV, Watters JW, et al. EMT is the dominant program in human colon cancer. BMC Med Genomics. 2011;4:1-10.
4. Xu C, Meng LB, Duan YC, et al. Screening and identification of biomarkers for systemic sclerosis via microarray technology. Int J Molecular Med. 2019;44:1753-70.
5. Ahmad MA, Eckert C, Teredesai A. Interpretable machine learning in healthcare. Proceedings of the 2018 ACM international conference on bioinformatics, Computational Biology Health Informatics. 2018
6. Yagin FH, Yagin B, Arslan AK, Çolak C. Comparison of Performances of Associative Classification Methods for Cervical Cancer Prediction: Observational Study. Turkey Clinics J Biostatistics. 2021;13:13:266-72.
7. Khaire UM, Dhanalakshmi R. High-dimensional microarray dataset classification using an improved adam optimizer (iAdam). J Ambient Intelligence Humanized Computing. 2020;11:5187-204.
8. Hameed SS, Hassan R, Hassan WH, et al. HDG-select: A novel GUI based application for gene selection and classification in high dimensional datasets. PloS One. 2021;16:e0246039.
9. Mulla GA, Demir Y, Hassan M. Combination of PCA with SMOTE Oversampling for Classification of High-Dimensional Imbalanced Data. Bitlis Eren University Science and Technology Journal. 2021;10:858-69.
10. Güçkiran K, Cantürk İ, Özyilmaz L. DNA microarray gene expression data classification using SVM, MLP, and RF with feature selection methods relief and LASSO. Journal of Suleyman Demirel University Institute of Science and Technology. 2019;23:126-32.
11. Akyol K, Bayir Ş, Baha Ş. Importance of Attribute Selection for Parkinson Disease. Academic Platform J Engineering Sci. 2020;8:175-80.
12. Yilmaz R, Yagin FH. Early detection of coronary heart disease based on machine learning methods. Med Records. 2022;4:1-6.
13. Secgin Y, Oner Z, Turan MK, Oner S. Gender prediction with parameters obtained from pelvis computed tomography images and decision tree algorithm. Med Science. 2021;10:356-61
14. Doğan Ş, Türkoğlu İ. Hypothyroidi and hyperthyroidi detection from thyroid hormone parameters by using decision trees. Fırat University Journal of Oriental Studies. 2007;5:163-9.
15. Pulat M, Kocakoç ID. Machine Learning and Decision in Turkey. Bibliometric Analysis of Published Theses in the Field of Trees. Journal of Management and Economics. 2021;28:287-308.
16. Kamel H, Abdulah D, Al-Tuwaijari JM. Cancer classification using gaussian naive bayes algorithm. 2019 Int Engineering Conference (IEC); 2019:36:165-5.
17. Quackenbush J. Microarray analysis and tumor classification. New England J Med. 2006;354:2463-72.
18. Jose A. Gene selection by 1-d discrete wavelet transform for classifying cancer samples using dna microarray date. Ph.D. thesis, University of Akron, 2009.
19. Yan W, Bai Z, Wang J, et al. ANP32A modulates cell growth by regulating p38 and Akt activity in colorectal cancer. Oncology Reports. 2017;38:1605-12.
20. Velmurugan BK, Yeh K-T, Lee C-H, et al. Acidic leucine-rich nuclear phosphoprotein-32A (ANP32A) association with lymph node metastasis predicts poor survival in oral squamous cell carcinoma patients. Oncotarget. 2016;7:10879.
21. Liu Q, Tan Y, Huang T, et al. TF-centered downstream gene set enrichment analysis: Inference of causal regulators by integrating TF-DNA interactions and protein post-translational modifications information. BMC Bioinformatics. 2010;11:1-17.
22. Mora JAM, Ordoñez FM, Bonilla DA. Improvement of k-means clustering algorithm performance in gene expression data analysis through pre-processing with principal component analysis and boosting. 2017;3:53-9.
23. Arentz G, Chataway T, Price TJ, et al. Desmin expression in colorectal cancer stroma correlates with advanced stage disease and marks angiogenic microvessels. Clinical Proteomics. 2011;8:1-13.
24. Bhunia S, Barbhuiya MA, Gupta S, et al. Epigenetic downregulation of desmin in gall bladder cancer reveals its potential role in disease progression. Indian J Med Research. 2020;151:311.
25. Chen H, Xu C, Qing'e Jin ZL. S100 protein family in human cancer. Am J Cancer Res. 2014;4:89.
26. Twal WO, Czirok A, Hegedus B, et al. Fibulin-1 suppression of fibronectin-regulated cell adhesion and motility. J Cell Sci. 2001;114:4587-98.
27. Xu Z, Chen H, Liu D, Huo J. Fibulin-1 is downregulated through promoter hypermethylation in colorectal cancer: a CONSORT study. Med (Baltimore). 2015;94.e663
28. Tong X, Mirzoeva S, Veliceasa D, et al. Chemopreventive apigenin controls UVB-induced cutaneous proliferation and angiogenesis through HuR and thrombospondin-1. Oncotarget. 2014;5:11413.
29. Ono C, Sato M, Taka H, et al. Tightly regulated expression of Autographa californica multicapsid nucleopolyhedrovirus immediate early genes emerges from their interactions and possible collective behaviors. Plos One. 2015;10:e0119580.
30. Strassburg CP, Kasai Y, Seng BA, et al. Baculovirus recombinant expressing a secreted form of a transmembrane carcinoma-associated antigen. Cancer Res. 1992;52:815-21.
31. Loging WT, Reisman D. Elevated expression of ribosomal

protein genes L37, RPP-1, and S2 in the presence of mutant p53. Cancer Epidemiology and Prevention Biomarkers. 1999;8:1011-6.

32. Golob-Schwarzl N, Schweiger C, Koller C, et al. Separation of low and high grade colon and rectum carcinoma by eukaryotic translation initiation factors 1, 5 and 6. Oncotarget. 2017;8:101224.

33. Oliveira P, Sanges R, Huntsman D, et al. Characterization of the intronic portion of cadherin superfamily members, common cancer orchestrators. European J Human Genetics. 2012;20:878-83.

34. Van Marck V, Stove C, Jacobs K, et al. Pcadherin in adhesion and invasion: Opposite roles in colon and bladder carcinoma. Int J Cancer. 2011;128:1031-44.

35. Takahashi K, Sasano H, Fukushima K, et al. 11 beta-hydroxysteroid dehydrogenase type II in human colon: a new marker of fetal development and differentiation in neoplasms. Anticancer Res. 1998;18:3381-8.

36. Baba Y, Nosho K, Shima K, et al. Prognostic significance of AMP-activated protein kinase expression and modifying effect of MAPK3/1 in colorectl cancer. British J Cancer. 2010;103:1025-33.

37. Esteve-Puig R, Canals F, Colome N, et al. Uncoupling of the LKB1-AMPKα energy sensor pathway by growth factors and oncogenic BRAFV600E. PloS One. 2009;4:e4771.

38. Zheng B, Jeong JH, Asara JM, et al. Oncogenic B-RAF negatively regulates the tumor suppressor LKB1 to promote melanoma cell proliferation. Molecular Cell. 2009;33:237-47.

39. Kim MJ, Park IJ, Yun H, et al. AMP-activated protein kinase antagonizes pro-apoptotic extracellular signal-regulated kinase activation by inducing dual-specificity protein phosphatases in response to glucose deprivation in HCT116 carcinoma. J Bio Chemistry. 2010;285:14617-27.

40. Arowolo MO, Isiaka RM, Abdulsalam SO, et al. A comparative analysis of feature extraction methods for classifying colon cancer microarray data. EAI Endorsed Transactions Scalable Information Systems. 2017;4:1-6.

41. Al Rajab M, Lu J, Xu Q. Examining applying high performance genetic data feature selection and classification algorithms for colon cancer diagnosis. Computer Methods Programs Bio Med. 2017;146:11-24.