creative commons
COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

의학석사 학위논문

Machine Learning-Based Classification of Proximal and Distal Gastric Cancer in The Cancer Genome Atlas Database

TCGA 데이터베이스를 활용한 근위부 위암과 원위부 위암의 기계학습 기반 분류

2023 년 8 월

서울대학교 대학원

의학과 외과학

이 은 주

Machine Learning-Based Classification of Proximal and Distal Gastric Cancer in The Cancer Genome Atlas Database

지도 교수 박 도 중

이 논문을 의학과 석사 학위논문으로 제출함

2023년 4월

서울대학교 대학원

의학과 외과학

이 은 주

이은주의 의학석사 학위논문을 인준함

2023년 7월

위 원 장 이혜승 (인)

부위원장 박도중 (인)

위 원 서윤석 (인)

Abstract

Machine Learning-Based Classification of Proximal and Distal Gastric Cancer in The Cancer Genome Atlas Database

Eunju Lee

College of Medicine, Department of Surgery

The Graduate School

Seoul National University

Background: Gastric cancer is a major global health concern, with different classifications based on its histological subtypes or anatomical location. Proximal gastric cancer (PGC) and distal gastric cancer (DGC) are two anatomically distinct subtypes with different risk factors, and understanding their clinicopathological and genetic characteristics is important for accurate diagnosis and treatment. This study investigated the genetic differences between PGC and DGC using machine learning (ML) approaches and data from The Cancer Genome Atlas (TCGA) program, and focused on identifying differences in DNA copy number variation and RNAseq.

Methods: The TCGA-Stomach Adenocarcinoma (STAD) dataset was used to investigate genetic differences between PGC and DGC. The study conducted classical bioinformatic approaches to distinguish PGC and DGC using a volcano plot and heap map from the selected features. To apply ML algorithms, data preprocessing was conducted by utilizing statistical tests to select noteworthy features, and false discovery rate correction was used to address the multiple testing problem. The study used 10-fold cross-validation for the ML algorithms to predict the location of gastric cancers using the selected features.

The validation was performed on subsets of the data, where different approaches were taken for handling the Fundus/Body data: In Group 1, the analysis excluded the

Fundus/Body data; in Group 2, the Fundus/Body data was classified as proximal gastric cancer for analysis; and in Group 3, the Fundus/Body data was classified and analyzed as a separate new group. The best algorithm was then chosen and used to interpret the results with the top 30 features of importance and EnrichR analysis.

Results: The study utilized ML techniques to identify potential genetic features in copy number variation and RNAseq to classify PGC and DGC within the TCGA-STAD dataset. Among the ML algorithms, gradient-boosting algorithms such as CatBoost and LightGBM consistently achieved high performances based on the Area Under the Curve (AUC), regardless of the differences in datasets. When classifying the Fundus/Body as PGC (Group 2), the AUC of the ROC curve was 0.75. However, when analyzing the data excluding the Fundus/Body as PGC (Group 1), the AUC of the ROC curve improved to 0.89. Furthermore, we identified the top 30 important features of CatBoost for classifying the tumor location, including LRRC8D and GULP1, and used them to perform EnrichR analysis, which provided information regarding their relationship with gastric cancer.

Conclusion: By applying ML to the TCGA-STAD database, this study identified potential genetic distinguishing features between PGC and DGC, indicating potential differences in their genetic profiles.

Keywords : Gastric cancer, TCGA, Genetics
Student Number : 2021-20463

Table of Contents

List of Tables

List of Figures

Introduction

Gastric cancer is the fourth most commonly diagnosed cancer in Korea, and the fifth most commonly diagnosed cancer worldwide, according to the 2020 GLOBOCAN data.[1] The classification of gastric cancer can be based on that of The World Health Organization or Lauren's classification, which categorizes gastric cancer based on its histological subtypes.[2,3] Alternatively, gastric cancer can be classified based on its anatomical location, as proximal gastric cancer (PGC) or distal gastric cancer (DGC).[4] DGC commonly occurs in the antrum and pylorus, while PGC occurs in the cardia and fundus, with the latter being more common in Western countries.[5,6] The risk factors for DGC include *H. pylori* infection, while gastroesophageal reflux disease and obesity are associated with PGC.[7-9]

Many research studies have examined the different clinicopathologic characteristics between PGC and DGC.[10] However, the molecular mechanisms behind these differences are still under investigation.[11] Therefore, understanding the clinicopathological and genetic characteristics of PGC and DGC is crucial for the accurate understanding of gastric cancer and its treatment.[12] Recently, machine learning (ML), a subset of artificial intelligence,[13] has been increasingly adopted in various fields, including medicine, to advance research and enhance outcomes. ML automatically acquires knowledge from data to accomplish a given object, so it may capture important and complex patterns which are undetectable to humans. Therefore, this study aims to investigate the genetic differences between PGC and DGC using ML techniques and The Cancer Genome Atlas (TCGA) database. The study also aimed to assess whether there were differences in copy number variation and gene expression profiles, derived from RNAseq data, between the two groups.

Methods

1. Data Preprocessing

The TCGA project is a collaboration between the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI), which began in 2005.[14] Its objective is to better understand the genomic and epigenomic alterations that occur in different types of cancer by collecting tissue samples and subjecting them to various genomic and epigenomic analyses. The data generated from this project has been made publicly available for researchers worldwide to

develop new therapeutic targets, diagnostic tools, and improve cancer biology. The TCGA project has also served as a model for other large-scale genomic initiatives such as the Genotype-Tissue Expression project and the International Cancer Genome Consortium.

The study utilized a subset of the TCGA database called TCGA-STAD, which focuses on STomach ADenocarcinoma, to analyze the genetic characteristics of proximal and distal gastric cancer. The formation of TCGA-STAD involved the collection of tumor samples from over 400 patients with stomach adenocarcinoma. These samples were subjected to various genomic and epigenomic analyses such as whole genome sequencing, RNAseq, methylation analysis, and proteomic analysis.

In detail, the study uses three types of data. The first is the TCGA-STAD gene-level copy number variation (CNV), estimated (n=441 with 24,777 identifiers) using the GISTIC2 method.[15] The GISTIC module identifies regions of the genome that are significantly amplified or deleted across a set of samples.[16] Each aberration is assigned a G-score that considers the amplitude of the aberration as well as the frequency of its occurrence across samples.

Several studies have identified specific CNVs that are associated with STAD.[17] For example, amplification of the HER2 (ERBB2) gene, involved in cell growth and division, is commonly observed in STAD and is associated with a poorer prognosis.[17] The copy number profile was measured experimentally using a whole genome microarray at a TCGA genome characterization center. Subsequently, the TCGA FIREHOSE pipeline applied the GISTIC2 method to produce segmented CNV data, mapped to genes to produce gene-level estimates. Genes are mapped onto the human genome coordinates using the University of California Santa Cruz(UCSC) Xena HUGO probe map.[18]

The second type of data is TCGA-STAD gene expression by RNAseq (n=417 with 26,541 identifiers), which was mean-normalized (per gene) across all TCGA cohorts. Indeed, many studies have used RNAseq to investigate the molecular mechanisms underlying STAD and to identify potential biomarkers and therapeutic targets.[19] For example, RNAseq analysis has identified genes that are differentially expressed in STAD compared to normal stomach tissue, including genes involved in cell cycle regulation, cell signaling, and immune function.[20] In addition, RNAseq analysis has been used to identify gene expression signatures that can predict patient

outcomes and responses to therapy.[21] For example, a gene expression signature that is associated with a better response to chemotherapy in patients with STAD. All RNAseq values are generated at UCSC by combining the "gene expression RNAseq" values of all TCGA cohorts. Values are mean-centered per gene, then the converted data from the cohort is extracted. The last data type is phenotypes (n=580 with 108 identifiers), which includes gender, age, anatomic neoplasm subdivision, and so on.[18] To combine the above data types, we match the sample ids for each data and then concatenate it. As a result, 336 cases with 51,323 identifiers are available.

Since one of the study objectives is to observe the difference in genetic characteristics between PGC and DGC, we index the category values in the anatomic neoplasm subdivisions, as shown in Table 1. Since TCGA does not separate the fundus and body, the study considered different groups for the dataset: the fundus and body category values are removed (Group 1), or are set as PGC (Group 2), or are regarded as a new classification class (Group 3), which should be predicted (Table 1).

Category	Number	Index
Gastroesophageal Junction	37	0 (Proximal)
Cardia/Proximal	45	0 (Proximal)
Fundus/Body	124	Group 1: Drop (Exclusion) Group 2: 0 (Proximal) Group 3: 2 (Fundus/Body)
Antrum/Distal	130	1 (Distal)
Other	3	Drop (Exclusion)
Stomach (Not otherwise specified)	5	
Discrepancy	1	

Table 1. The distribution of the anatomic neoplasm subdivisions in gastric cancer. (In Group 1, the analysis was conducted excluding the Fundus/Body data, while in Group 2, the Fundus/Body data was classified as proximal gastric cancer for analysis, and in Group 3, it was classified and analyzed as a separate new group.)

2. Statistical analysis

The study performed several statistical tests to identify the noteworthy features. A normality test for both distributions, by indexing the anatomic neoplasm subdivisions, was performed to determine if both distributions passed the normality test. If both distributions passed the normality test, an equality test was performed, and t-tests with or without equality were performed accordingly. In addition, if normality was not passed, the Mann-Whitney U test and the chi-square test were performed for continuous and categorical values, respectively. If there is a genetic difference between PGC and DGC, and the *P* value was less than 0.05, that was regarded as statistically significant.

However, we remark that multiple testing correction is necessary because it provides a way to control the rate at which false positives occur when performing multiple statistical tests simultaneously.[22] When conducting multiple tests, it is possible to obtain meaningful results by chance alone, even when there is no true effect present. This can lead to false positive results, which can lead to wrong conclusions. Some commonly used methods for addressing the multiple testing problem when conducting multiple statistical tests simultaneously, each with its advantages and disadvantages, include the Bonferroni correction, false discovery rate (FDR), Benjamini-Hochberg correction, and so on. This study used the FDR since it is less conservative with high statistical power, and it minimizes the risk of missing important results.[23] The alpha parameter was set as 0.05, which is a threshold that controls the overall FDR. This value represents the significance level of the statistical test. The value is compared to the *P* value, which represents the probability of obtaining the observed results by chance alone. Using the *P* values obtained with the FDR, we finally chose 2,109 candidates from the 7,784 features whose *P* values are less than 0.05 among more than 50,000 identifiers, which may result in PGC or DGC. For statistical analysis, the study utilized Python libraries including SciPy, which is an open-source software for mathematics, science, and engineering.

3. Standard Approach

Before utilizing ML, feature distributions were compared between PGC and DGC patients, where each feature has a *P* value smaller than 0.05. Next, volcano and heatmap plots were produced to check whether some of the RNA and DNA data are separable between PGC and DGC patients.

4. **Machine Learning Approach**

To apply ML algorithms, data was split after FDR into training and test datasets, where *P* values for each feature distribution are smaller than 0.05. Then, 10-fold cross-validation was performed for various ML algorithms. The study validated various tree generation methods: from the vanilla decision tree method,[24] some methods generated new trees independently, which were used for voting to improve the performance (e.g., Random Forest[25] and Extra Trees[26]). Gradient boosting classifiers are a type of machine learning algorithm, where the generation of a decision tree depends on the previous trees and focuses on their errors (e.g., vanilla Gradient Boosting,[27] Light Gradient Boosting Machine,[28] CatBoost,[29] Extreme Gradient Boosting,[30] Ada Boost[31]). In addition, other standard methods were considered: logistic regression, linear discriminant,[32] quadratic discriminant,[33] ridge,[34] Naïve Bayes,[35] support vector machine,[36] the *K*-Nearest Neighbor algorithm,[37] and Dummy Classifier, which ignore features and determine a class randomly. Among them, the best algorithm is selected to analyze which features are significant for predicting gastric cancer locations.

Finally, the study conducted EnrichR analysis for the top 30 important features to confirm whether they are indeed related to gastric cancers or not.[38-40] Note that the study objective is to determine which features are significant, so we do not consider an ensemble of ML models to improve performance (e.g., blending machine learning models).

Results

In the TCGA data, the location of the gastric cancer and the number of corresponding patients were 'Antrum/Distal': 130, 'Cardia/Proximal': 45, 'Fundus/Body': 124, 'Gastroesophageal Junction': 37, 'Other': 3, 'Stomach (NOS)': 5, and 'Discrepancy': 1.

The study first extracted the top 10 genes with respect to copy numbers and RNA seq, sorted by *P* values, as shown in Table 2 and Table 3. PTEN is a tumor suppressor gene that is involved in various cellular processes, including cell growth, differentiation, and survival.[41] Loss of PTEN function has been implicated in the development and progression of several types of cancer, including gastric cancer.

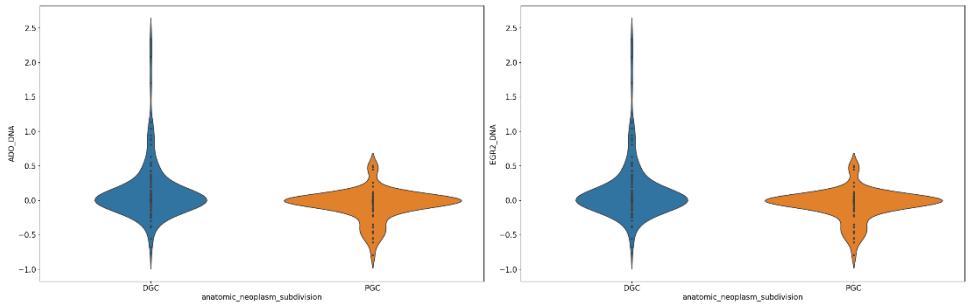
Although there are some high-ranking genes, e.g., HMX3, to the best of our knowledge, there is little evidence linking these genes to gastric cancer. However, it is noteworthy that they have different distributions for proximal and distal cancers. 1 shows the distribution of the DNAs which have the smallest *P* values. Figure 2 shows that there are also some differences between the two different cancer positions. However, analyzing features independently is inefficient due to the large number of features, and some feature distributions between PGC and DGC are barely distinguishable (Figure 1). Therefore, it is unclear which combination of genes or samples are suitable for characterizing the gastric cancer position. Indeed, the study compared the copy number variation and expression of RNAseq. Among the genes that are commonly expressed, 62 showed significant differences in both copy number and expression of RNAseq. Furthermore, 1,659 genes showed significant differences in copy number but not in the expression of RNAseq, while there were no genes that showed significant expression differences in RNAseq but no differences in copy number. This result suggests that copy number may be a more important expression phenotype for determining the location of gastric cancer. However, as we observed, there are many features with high correlations (Figure 1), so it is naïve to conclude that the combination of low *P* value features is effective. Besides, even by using a heatmap (Figure 3), the study could not observe some differences between PGC and DGC. Accordingly, a new approach that can capture multiple features simultaneously is necessary.

Proximal dominant		Distal dominant	
Feature	*P* value	Feature	*P* value
NKAP	0.005362	RTKN2	0.0000895
FGF13	0.006486	ADO	0.000117
KIAA1210	0.007785	EGR	0.000117
LONRF3	0.007966	RN7SL591P	0.000145
WDR44	0.0086	ZNF365	0.000145
MIR1277	0.0086	JMJD1C	0.00017
IL13RA1	0.009243	NRBF2	0.000186
DOCK11	0.009243	MIR1296	0.000213
ZCCHC12	0.009965	REEP3	0.000228
SMIM10	0.010270047	PTEN	0.000466

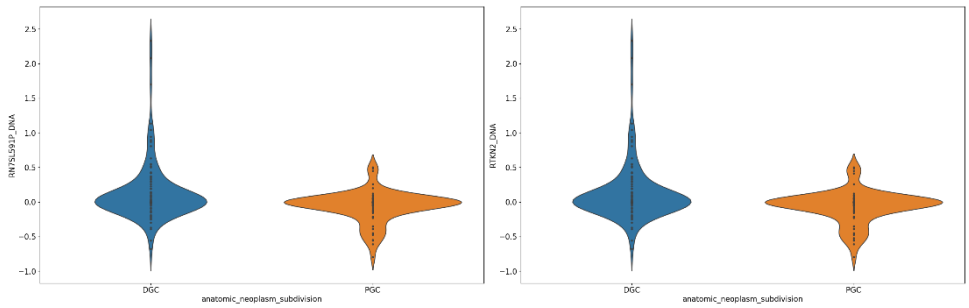
Table 2. Top 10 important copy numbers (gene level) with respect to the copy level sorted by *P* values.

Proximal dominant		Distal dominant	
Feature	*P* value	Feature	*P* value
MYBPH	0.000051	MPP3	0.000156
ZAR1L	0.000066	IRX3	0.001313
LOC100131257	0.000236	ARNT2	0.001515
SNORA70D	0.000251	HOMER2	0.002174
CACNA2D4	0.000536	RELA	0.00237
GPR128	0.000546	EYA2	0.00256
ALKBH1	0.000551	IRS4	0.002954
SPAM1	0.000691	ADRB1	0.003034
ZNF18	0.000853	WWTR1-AS1	0.00324
DQ594410	0.000946	FAM102A	0.003635

Table 3. Top 10 important features with respect to RNAseq sorted by *P* values.



(a) ADO_DNA (b) EGR2_DNA



(c) RN7SL591P_DNA (d) RTKN2_DNA

Figure 1. Distribution of gene values across different types. The distributions of individual genes between distal gastric cancer (DGC) and proximal gastric cancer (PGC) are markedly different. However, the distributions for DNA features are almost identical between the two groups.

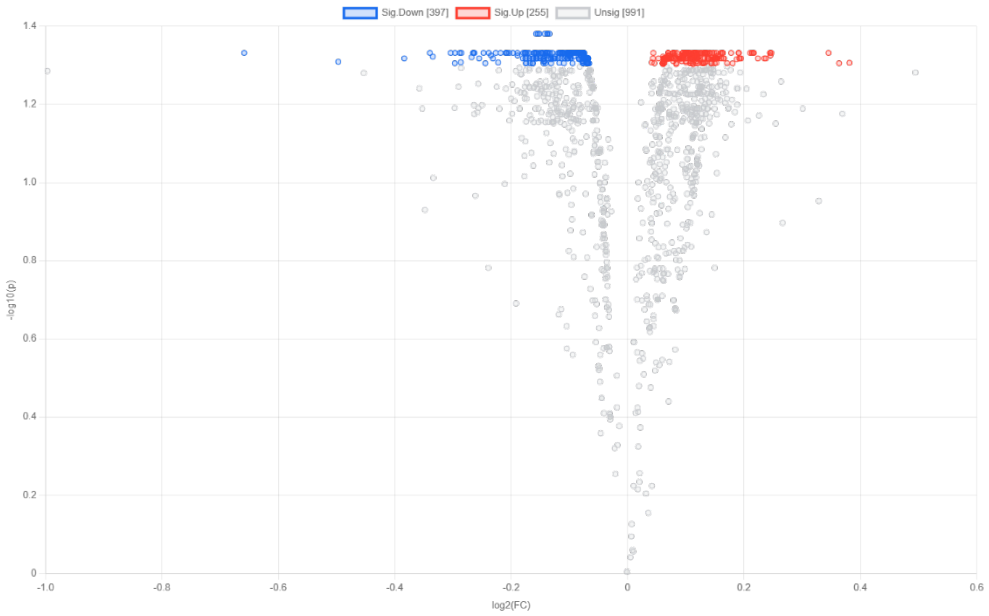


Figure 2. Volcano plot with respect to RNA and DNA features after false discovery rate correction. The blue and red points denote features having different distributions with a large fold change.

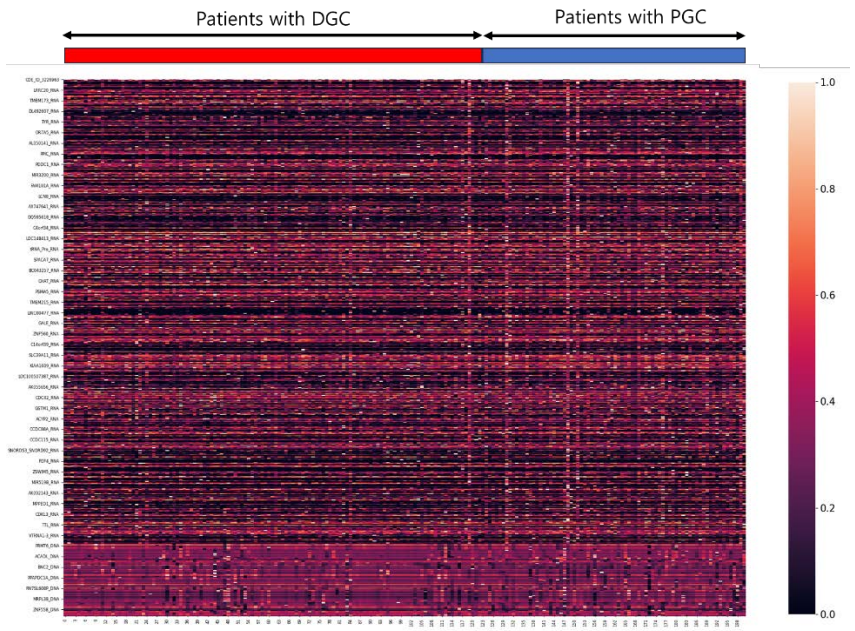


Figure 3. Heatmap Plot. The x-axis represents patients, while the y-axis represents individual features, including RNA sequences and DNA copy numbers.

Since the determination of the ideal combinations of RNA and DNA features for gastric cancer positions is complex, the study utilized ML algorithms to predict the locations of gastric cancer because well-trained models can identify which features are crucial for determining the location of cancers. The various ML algorithms require validation, and it is then possible to choose the best algorithm. As illustrated in Table 1, the study considered three types of datasets (Group 1: the fundus and body category values are removed, Group 2: the fundus and body category values belong to PGC, and Group 3: the fundus and body category values are in a new class). To validate the methods, 10-fold cross-validations were used. To perform this, the training and test datasets were randomly split into 75% and 25% portions, respectively.

Model	Accuracy	AUC[a]	Recall	Precision	F1 score	Kappa	MCC[b]
Gradient Boosting Classifier	0.7571	0.905	0.85	0.787	0.8032	0.4789	0.5187
Light Gradient Boosting Machine	0.7929	0.889	0.8611	0.8188	0.8317	0.5593	0.5817
CatBoost Classifier	0.7929	0.8758	0.8958	0.7968	0.8361	0.5508	0.5809
Extreme Gradient Boosting	0.7571	0.8568	0.8486	0.7809	0.8054	0.4784	0.5048
Extra Trees Classifier	0.7786	0.8508	0.8722	0.7952	0.8235	0.5232	0.5536
Random Forest Classifier	0.7714	0.841	0.8611	0.7918	0.8182	0.5035	0.522
Ada Boost Classifier	0.7786	0.8246	0.8264	0.8231	0.8174	0.5319	0.5499
Logistic Regression	0.7786	0.8197	0.8597	0.7995	0.8241	0.5263	0.539
Naive Bayes	0.7	0.7469	0.7486	0.7782	0.7506	0.3789	0.4022
K-Nearest Neighbor algorithm	0.7	0.7446	0.8236	0.7358	0.7703	0.3333	0.3504
Decision Tree Classifier	0.6857	0.6668	0.7569	0.7417	0.7381	0.339	0.3581
Linear Discriminant Analysis	0.6857	0.6666	0.7542	0.7374	0.742	0.3328	0.3388
Quadratic Discriminant Analysis	0.4929	0.5032	0.5097	0.6249	0.5317	0.0036	0.0031
Dummy Classifier	0.6143	0.5	1	0.6143	0.7605	0	0
Ridge Classifier	0.7143	0	0.7806	0.7814	0.7685	0.3988	0.4144
SVM - Linear Kernel	0.6571	0	0.6417	0.7049	0.6218	0.3373	0.3885

Table 4. Results of 10-fold cross-validation using various machine learning algorithms with Group 1. (In Group 1, the analysis was conducted excluding the Fundus/Body data.)

[a] AUC : Area under curve

[b] MCC : Matthews Correlation Coefficient

Model	Accuracy	AUC[a]	Recall	Precision	F1 score	Kappa	MCC[b]
Logistic Regression	0.7786	0.8136	0.8611	0.799	0.8248	0.5215	0.5374
CatBoost Classifier	0.7643	0.8018	0.9069	0.7709	0.8276	0.455	0.4837
Random Forest Classifier	0.7571	0.7974	0.8611	0.7715	0.8076	0.4715	0.4977
Gradient Boosting Classifier	0.75	0.779	0.8625	0.7717	0.8083	0.4445	0.4668
Extra Trees Classifier	0.7429	0.7812	0.825	0.779	0.7948	0.4427	0.4637
Ridge Classifier	0.7357	0	0.85	0.7623	0.7983	0.4153	0.4381
Ada Boost Classifier	0.7286	0.7614	0.8278	0.7678	0.7899	0.3995	0.4169
K-Nearest Neighbor algorithm	0.7214	0.7099	0.8708	0.7326	0.7924	0.3722	0.4014
Light Gradient Boosting Machine	0.7143	0.7939	0.8153	0.7561	0.7756	0.3689	0.3854
SVM - Linear Kernel	0.7	0	0.7028	0.8051	0.7103	0.4102	0.4642
Naive Bayes	0.6714	0.7033	0.725	0.7369	0.7251	0.303	0.3006
Decision Tree Classifier	0.6571	0.6483	0.7	0.7382	0.7142	0.2907	0.2945
Dummy Classifier	0.6143	0.5	1	0.6143	0.7605	0	0
Linear Discriminant Analysis	0.5929	0.6189	0.6056	0.703	0.6363	0.1731	0.1875
Quadratic Discriminant Analysis	0.5571	0.5693	0.4986	0.7103	0.5607	0.1245	0.1428

Table 5. Results of 10-fold cross-validation using various machine learning algorithms with Group 2. (In Group 2, the Fundus/Body data was classified as proximal gastric cancer for analysis.)

[a] AUC : Area under curve

[b] MCC : Matthews Correlation Coefficient

Model	Accuracy	AUC[a]	Recall	Precision	F1 score	Kappa	MCC[b]
CatBoost Classifier	0.7281	0.6853	0.1595	0.5333	0.2352	0.1441	0.1836
Light Gradient Boosting Machine	0.7235	0.6743	0.3167	0.4338	0.3575	0.2179	0.2201
Random Forest Classifier	0.7233	0.65	0.1571	0.51	0.2203	0.1277	0.1645
Gradient Boosting Classifier	0.7194	0.6485	0.3	0.4625	0.3421	0.2054	0.2227
Dummy Classifier	0.7146	0.5	0	0	0	0	0
Extra Trees Classifier	0.7142	0.6179	0.1238	0.5333	0.1952	0.0946	0.1373
Ada Boost Classifier	0.7012	0.6774	0.3476	0.5079	0.3969	0.2113	0.226
Logistic Regression	0.6885	0.6781	0.4714	0.469	0.4556	0.2432	0.2513
K-Nearest Neighbor algorithm	0.6826	0.5348	0.1071	0.35	0.1583	0.0247	0.0368
Linear Discriminant Analysis	0.6557	0.6208	0.481	0.4035	0.4323	0.1898	0.1963
Ridge Classifier	0.652	0	0.4833	0.4016	0.4255	0.1838	0.1934
Decision Tree Classifier	0.6344	0.5493	0.3548	0.3549	0.3421	0.0978	0.0994
SVM - Linear Kernel	0.6338	0	0.4167	0.3224	0.3221	0.1186	0.1482
Quadratic Discriminant Analysis	0.5986	0.5375	0.4	0.3254	0.3507	0.0671	0.0702
Naive Bayes	0.5937	0.5926	0.4881	0.3626	0.4111	0.1201	0.1191

Table 6. Results of 10-fold cross-validation using various machine learning algorithms with Group 3. (In Group 3, the Fundus/Body data was classified and analyzed as a separate new group.)

[a] AUC : Area under curve

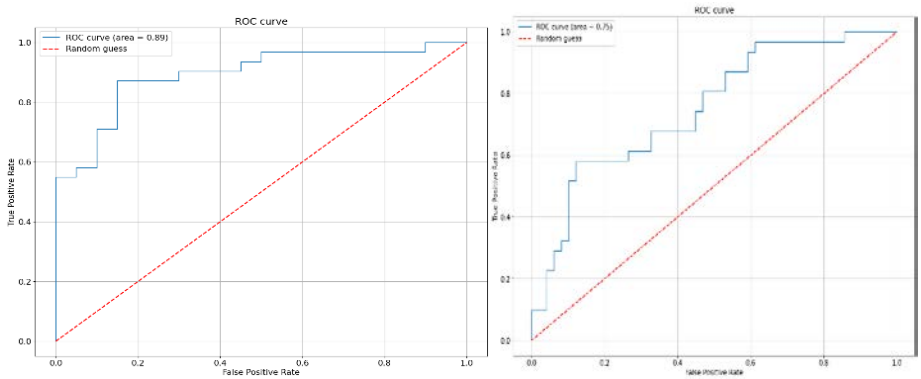
[b] MCC : Matthews Correlation Coefficient

The results demonstrated that the gradient boosting methods were superior to the other methods. Especially, CatBoost belongs to the top three ML algorithms in each of Group 1, Group 2, and Group 3, based on the AUC (Table 4~6). The performance in Group 3 degrades significantly because there is a large gap between the binary classification and multiple classification with many features.

Therefore, the study focused on Group 1 and Group 2, but not Group 3, and selected CatBoost to analyze the results, since it is known as one of the best gradient boosting methods due to its ability to handle large datasets with high-dimensional features and its efficient memory usage. It also has several built-in features to prevent overfitting, such as the ability to perform early stopping and to use a custom loss

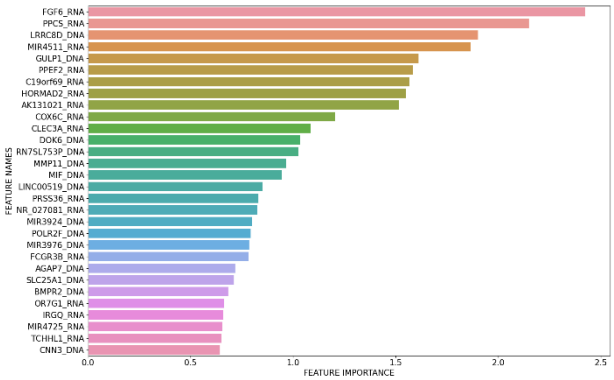
function. Additionally, it has a robust implementation of feature importance calculation, which can help users understand which features are most important in their models.

For Group 1, the study observed an AUC of the ROC curve of 0.89 when using the test dataset, as shown in Figure 4. (a). This demonstrates a remarkable performance. For Group 2, the study obtained an AUC of 0.75 for ROC curve when using the test dataset, which means that the prediction performance is fairly good, but there is a gap between Group 1 and Group 2, as shown in Figure 4. (b). Commonly, adding uncertain data for training and testing may degrade ML algorithms. Moreover, we computed the feature importance calculated by CatBoost. Figure 5. (a)-(b) shows which features are important in determining the position of gastric cancers for Group 1 and Group 2, respectively.

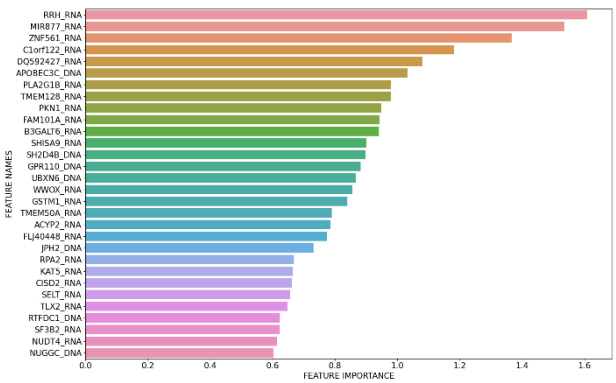


(a) Group 1 (b) Group 2

Figure 4. ROC curves with AUCs of 0.89 and 0.75 for Group 1 and Group 2, respectively, after training with CatBoost. (In Group 1, the analysis was conducted excluding the Fundus/Body data, while in Group 2, the Fundus/Body data was classified as proximal gastric cancer for analysis.)



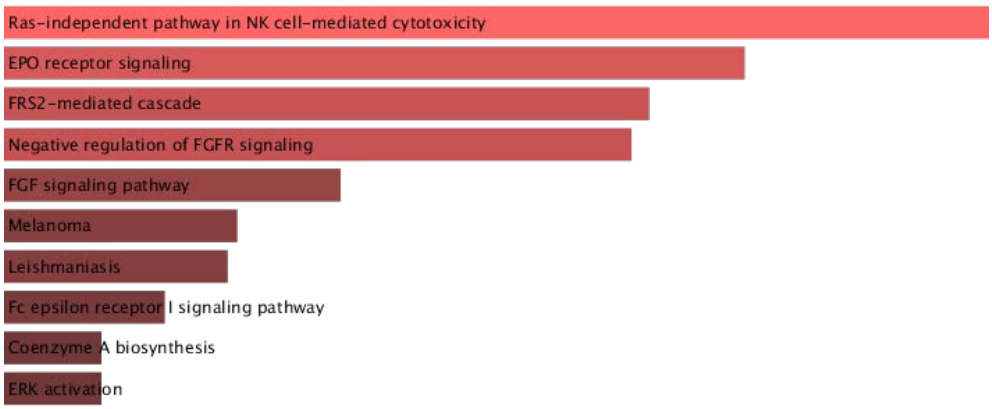
(a) Group 1, in which the analysis was conducted excluding the Fundus/Body data



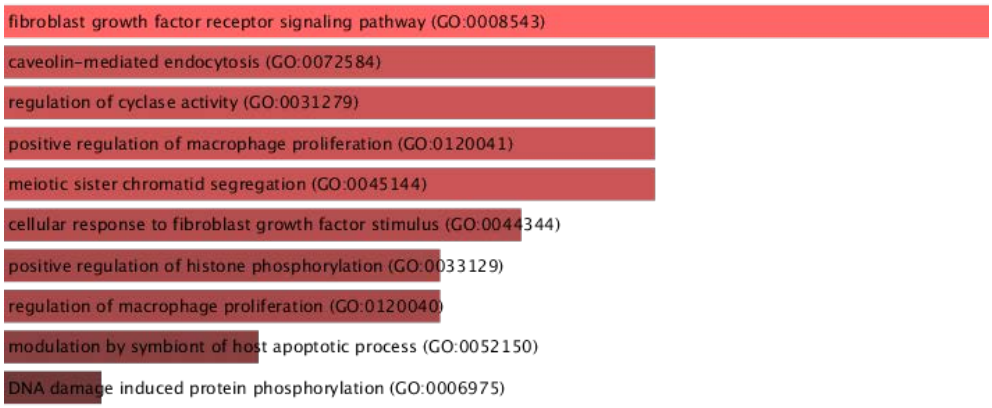
(b) Group 2, in which the Fundus/Body data was classified as proximal gastric cancer for analysis

Figure 5. Top 30 feature importance by CatBoost in Group 1 and Group 2.
(Group 1 excludes the Fundus/Body data in the analysis, whereas Group 2 classifies the Fundus/Body data as proximal gastric cancer for analysis.)

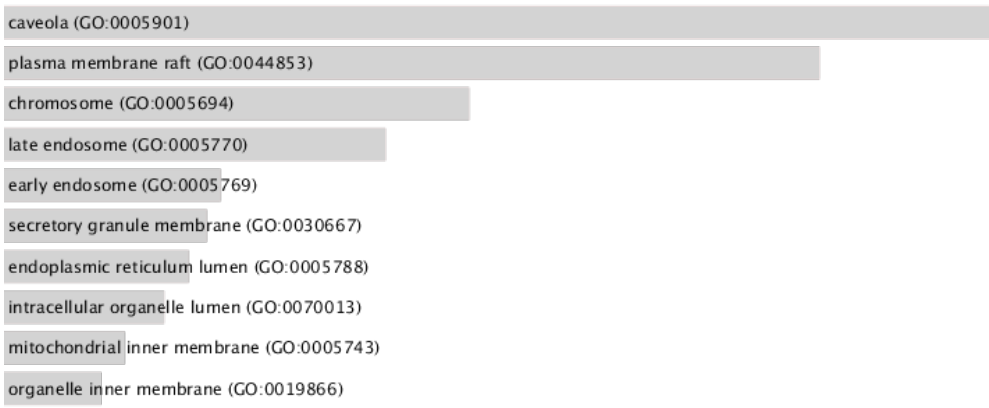
Figure 6 shows the EnrichR analysis of the top 30 important features for CatBoost's predictions based on *P* values in Group 1, for which performance is indicated by an AUC of 0.89. These features are highly related to gastric cancers. For instance, studies have shown that FGFR1, FGFR2, and FGFR4 are overexpressed in gastric cancer tissues and that the aberrant activation of FGFR signaling is involved in the development and progression of gastric cancer.[42] Additionally, some preclinical studies have suggested that targeting FGFR signaling may have therapeutic potential in treating gastric cancer.[43]



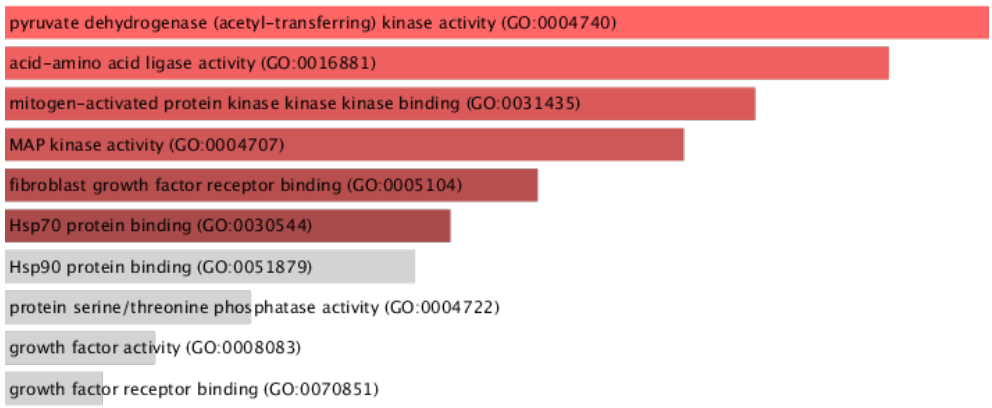
(a) Bioplanet_2019



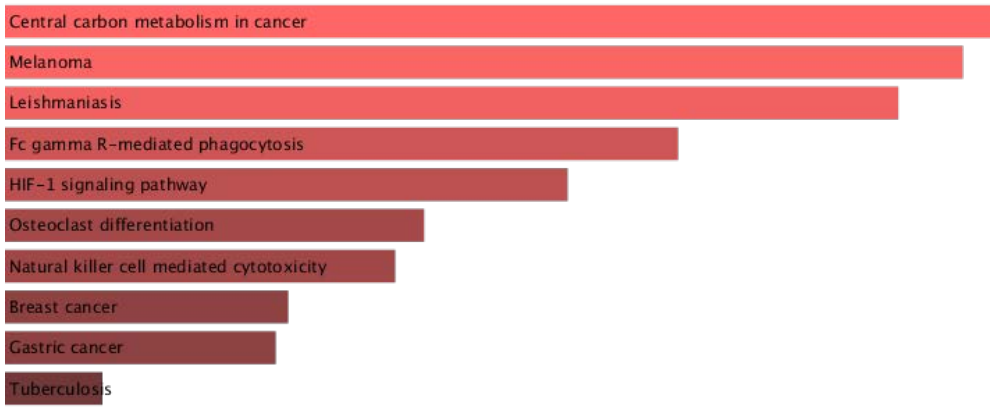
(b) GO_Biological_Process_2021



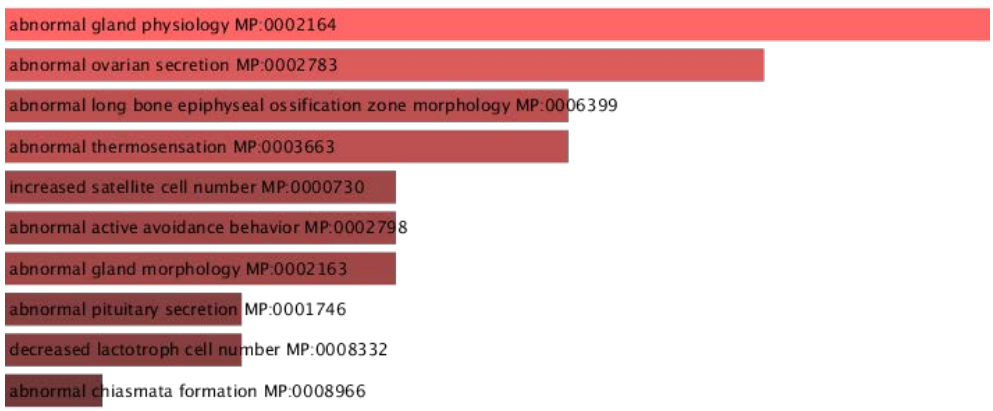
(c) GO_Cellular_Component_2021



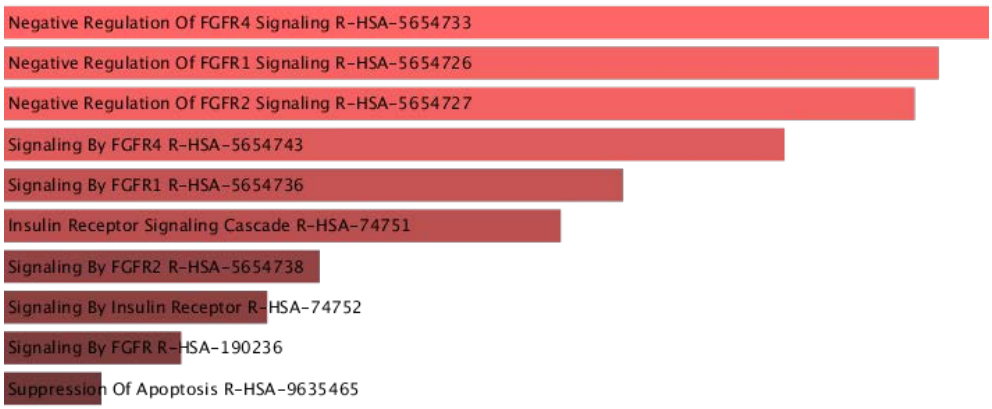
(d) GO_Molecular_Function_2021



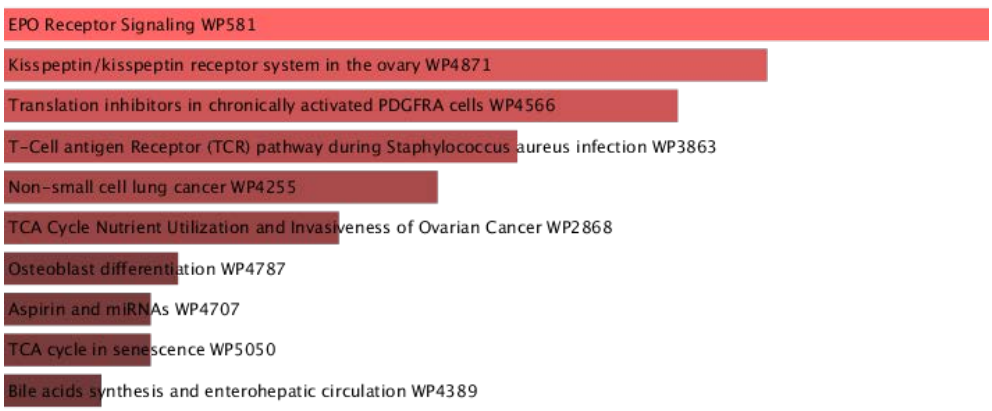
(e) KEGG_2021_Human



(f) MGI_Mammalian_Phenotype_Level_4_2021



(g) Reactome_2022



(h) WikiPathway_2021_Human_bar_graph

Figure 6. EnrichR analysis of the top 30 important features for CatBoost's predictions based on *P* values in Group 1. (Group 1 excludes the Fundus/Body data in the analysis.)

Discussion

This study utilized the ML-based classification of PGC and DGC through the TCGA database. The study findings offer valuable insights that could contribute to the development of personalized treatment strategies and targeted therapies. The clinical importance of differentiating between PGC and DGC is particularly emphasized when tumors are located in regions such as the mid to high body of the stomach. In these cases, function-preserving surgeries such as proximal gastrectomy or pylorus-preserving gastrectomy become viable surgical options in addition to total gastrectomy. This differentiation is crucial in order to tailor the surgical approach to

the specific characteristics and location of the tumor, ultimately aiming to maximize patient outcomes and preserve gastric function

Overall, the study highlights the importance of considering anatomical and genetic differences when classifying gastric cancer and developing personalized treatment strategies. By suggesting possible genetic characteristics associated with PGC and DGC, the findings may contribute to more effective diagnosis, treatment, and prognosis for gastric cancer patients.

The TCGA database grouped the fundus and body together. Based on the Japanese classification of gastric carcinoma, the stomach can be divided into three sections — the upper, middle, and lower parts — using lines that connect the trisected points on the lesser and greater curvatures.[4] Even though mid-body cancer and lower-body cancer of the stomach are anatomically part of the middle third of the stomach, body cancer was grouped into the PGC group. This grouping may have obscured the unique genetic characteristics of PGC and made it challenging to differentiate between PGC and DGC, particularly for middle-third stomach body cancer, which is classified as PGC. When analyzing the data with the exclusion of the Fundus/Body, the performance of ML prediction increased from an AUC of 0.75 to 0.89. Future research should consider more detailed analyses using databases that include comprehensive clinicopathologic and genetic information. Larger patient cohorts and more homogeneous patient populations, such as Korean patients, could provide a more accurate understanding of the molecular mechanisms driving these subtypes of gastric cancer.

This study had some limitations related to the heterogeneous patient population in the TCGA database, which comprises individuals of various racial backgrounds. This may have introduced confounding factors when evaluating cancer subtype differences across different ethnicities. As East Asian and Western gastric cancer exhibit distinct characteristics, the genetic differences observed in this study may not be wholly generalizable to all racial groups. Another limitation of this study is the inability to compare normal and tumor data in the RNAseq data obtained from the TCGA database. While RNAseq has shown remarkable performance in accurately distinguishing PGC and DGC using ML, it may not be directly applicable or specific to tumor samples.

Conclusion

Through the utilization of ML and the TCGA-STAD database, the study introduced possible genetic distinguishing points between PGC and DGC, suggesting potential differences in their genetic profiles.

References

1. Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians 2021;71(3):209-249.
2. Nagtegaal ID, Odze RD, Klimstra D, et al. The 2019 WHO classification of tumours of the digestive system. Histopathology 2020;76(2):182.
3. Lauren P. The two histological main types of gastric carcinoma: diffuse and so-called intestinal-type carcinoma: an attempt at a histo-clinical classification. Acta Pathologica Microbiologica Scandinavica 1965;64(1):31-49.
4. jp JGCAjkk-ma. Japanese classification of gastric carcinoma: 3rd English edition. Gastric cancer 2011;14(2):101-112.
5. Smyth EC, Nilsson M, Grabsch HI, van Grieken NC, Lordick F. Gastric cancer. The Lancet 2020;396(10251):635-648.
6. Karimi P, Islami F, Anandasabapathy S, Freedman ND, Kamangar F. Gastric Cancer: Descriptive Epidemiology, Risk Factors, Screening, and PreventionGastric Cancer. Cancer epidemiology, biomarkers & prevention 2014;23(5):700-713.
7. Kuipers E. Exploring the link between Helicobacter pylori and gastric cancer. Alimentary pharmacology & therapeutics 1999;13:3-11.
8. Crew KD, Neugut AI. Epidemiology of gastric cancer. World journal of gastroenterology: WJG 2006;12(3):354.
9. Chen Y, Liu L, Wang X, et al. Body mass index and risk of gastric cancer: a meta-analysis of a population with more than ten million from 24 prospective studies. Cancer epidemiology, biomarkers & prevention 2013;22(8):1395-1408.
10. Piso P, Werner U, Lang H, Mirena P, Klempnauer J. Proximal versus distal gastric carcinoma—what are the differences? Annals of surgical oncology 2000;7:520-525.
11. Zhang Y, Zhang P-S, Rong Z-Y, Huang C. One stomach, two subtypes of carcinoma—the differences between distal and proximal gastric cancer. Gastroenterology Report 2021;9(6):489-504.
12. Cristescu R, Lee J, Nebozhyn M, et al. Molecular analysis of gastric cancer identifies subtypes associated with distinct clinical outcomes. Nature medicine 2015;21(5):449-456.
13. Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. Lancet 2019;393(10181):1577-1579. DOI: 10.1016/S0140-6736(19)30037-6.
14. 53 DCCBRJMAKAPTPDWY, 68 TSSLDA. The cancer genome atlas pan-cancer analysis project. Nature genetics 2013;45(10):1113-1120.
15. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G. GISTIC2. 0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome biology 2011;12:1-14.
16. Beroukhim R, Getz G, Nghiemphu L, et al. Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. Proceedings of the National Academy of Sciences 2007;104(50):20007-20012.

17. Gravalos C, Jimeno A. HER2 in gastric cancer: a new prognostic factor and a novel therapeutic target. Annals of oncology 2008;19(9):1523-1529.
18. Goldman MJ, Craft B, Hastie M, et al. Visualizing and interpreting cancer genomics data via the Xena platform. Nature biotechnology 2020;38(6):675-678.
19. Zhou L, Huang W, Yu H-F, Feng Y-J, Teng X. Exploring TCGA database for identification of potential prognostic genes in stomach adenocarcinoma. Cancer Cell International 2020;20(1):1-12.
20. Sun J, Jiang Q, Chen H, et al. Genomic instability-associated lncRNA signature predicts prognosis and distinct immune landscape in gastric cancer. Annals of Translational Medicine 2021;9(16).
21. Van't Veer LJ, Dai H, Van De Vijver MJ, et al. Gene expression profiling predicts clinical outcome of breast cancer. nature 2002;415(6871):530-536.
22. Noble WS. How does multiple testing correction work? Nature biotechnology 2009;27(12):1135-1137.
23. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological) 1995;57(1):289-300.
24. Quinlan JR. Learning decision tree classifiers. ACM Computing Surveys (CSUR) 1996;28(1):71-72.
25. Breiman L. Random forests. Machine learning 2001;45:5-32.
26. Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. Machine learning 2006;63:3-42.
27. Friedman JH. Stochastic gradient boosting. Computational statistics & data analysis 2002;38(4):367-378.
28. Ke G, Meng Q, Finley T, et al. Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems 2017;30.
29. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features. Advances in neural information processing systems 2018;31.
30. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining2016:785-794.
31. Hastie T, Rosset S, Zhu J, Zou H. Multi-class adaboost. Statistics and its Interface 2009;2(3):349-360.
32. Balakrishnama S, Ganapathiraju A. Linear discriminant analysis-a brief tutorial. Institute for Signal and information Processing 1998;18(1998):1-8.
33. Tharwat A. Linear vs. quadratic discriminant analysis classifier: a tutorial. International Journal of Applied Pattern Recognition 2016;3(2):145-180.
34. Hoerl AE, Kennard RW. Ridge regression: applications to nonorthogonal problems. Technometrics 1970;12(1):69-82.
35. Webb GI, Keogh E, Miikkulainen R. Naïve Bayes. Encyclopedia of machine learning 2010;15(1):713-714.
36. Hearst MA, Dumais ST, Osuna E, Platt J, Scholkopf B. Support vector machines. IEEE Intelligent Systems and their applications 1998;13(4):18-28.
37. Mladenović N, Hansen P. Variable neighborhood search. Computers & operations research 1997;24(11):1097-1100.

38. Chen EY, Tan CM, Kou Y, et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC bioinformatics 2013;14(1):1-14.
39. Kuleshov MV, Jones MR, Rouillard AD, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic acids research 2016;44(W1):W90-W97.
40. Xie Z, Bailey A, Kuleshov MV, et al. Gene set knowledge discovery with Enrichr. Current protocols 2021;1(3):e90.
41. Kim B, Kang SY, Kim D, Heo YJ, Kim K-M. PTEN protein loss and loss-of-function mutations in gastric cancers: the relationship with microsatellite instability, EBV, HER2, and PD-L1 expression. Cancers 2020;12(7):1724.
42. Futami T, Kawase T, Mori K, et al. Identification of a novel oncogenic mutation of FGFR4 in gastric cancer. Scientific Reports 2019;9(1):14627.
43. Yashiro M, Matsuoka T. Fibroblast growth factor receptor signaling as therapeutic targets in gastric cancer. World journal of gastroenterology 2016;22(8):2415.

초 록

TCGA 데이터베이스를 활용한 근위부 위암과 원위부 위암의 기계학습 기반 분류

이은주

의학과 외과학 전공

서울대학교

서론: 위암의 위치에 따라 근위부 위암과 원위부 위암으로 분류할 수 있으며, 서로 다른 위험 요인을 가지며, 임상병리학적 특성이 다르다는 것이 알려져 있다. 본 연구에서는 기계학습 기반으로 The Cancer Genome Atlas (TCGA) 데이터베이스를 활용하여 근위부 위암과 원위부 위암의 DNA copy number variation 및 RNAseq의 차이를 비교해보고자 하였다.

방법: TCGA-STAD 데이터셋을 전처리하여 근위부 위암과 원위부 위암에서의 유전적 차이를 조사 및 분석하였다. 기계 학습 기반의 Grandient Boosting 알고리즘을 이용하여 위암의 위치를 DNA copy number variation 및 RNAseq 정보를 이용하여 예측 및 예측에 쓰인 인자를 확인하였다. 상위 30개 특성을 추출하여 EnrichR 분석을 수행하였다.

결과: 근위부 위암과 원위부 위암에 관련된 잠재적 유전자들을 확인하였으며, 기계학습 알고리즘 중 하나인 CatBoost를 활용하여 근위부 위암과 원위부 위암의 구분하는 특징에 대해서 확인하였다.

위저부/위체부 데이터를 근위부 위암으로 분류하여 분석시에 CatBoost의 테스트 데이터에 대한 ROC 곡선의 AUC는 0.75였으나, 제외하고 분석시에는 ROC 곡선의 AUC는 0.89로 향상되었다.

결론: 본 연구에서는 기계학습 기반으로 TCGA-STAD 데이터베이스를 활용하여 근위부 위암과 원위부 위암을 구분하는 가능성 있는 유전적 차이에 대해 확인해보았다.

주요어 : 위암, 유전학
학 번 : 2021-20463