THE JOURNAL OF COGNITIVE SYSTEMS
Journal Homepage: https://dergipark.org.tr/en/pub/jcs
Vol.6, No.2, 2021

# Biomarkers for predicting diabetes in gastric cancer patients with machine learning methods based on proteomic data

[1]Şeyma YAŞAR, [2]Büşra Nur FINDIK

[1] Inonu University Department of Biostatistics and Medical Informatics, Faculty of Medicine, Malatya, Turkey. (e-mail: seyma.yasar@inonu.edu.tr),
[2]Nevsehir Haci Bektas Veli University, Department of Therapy and Rehabilitation, Kozakli Vocational School, Nevsehir, Turkey. (e-mail: busranurfindik@nevsehir.edu.tr).

**ARTICLE INFO**

**Received:**
**Revised:**
**Accepted:**

**Keywords:**
Diabetes mellitus
Gastric Cancer
Stochastic Gradient Boosting
Bagged CART
Classification

**Corresponding author:** Şeyma YAŞAR
✉ seyma.yasar@inonu.edu.tr
☎ +90 507 639 71 21

**ISSN:** 2548-0650

**DOI:**

**ABSTRACT**

Gastric cancer is a type of cancer that occurs when cells in the stomach tissue grow and multiply abnormally. Gastric cancer usually starts in the inner layer of the stomach wall and can spread to other layers over time. This type of cancer is most common in people over the age of 50, but it can also occur in younger people. Symptoms of gastric cancer include indigestion and stomach pain, nausea and vomiting, loss of appetite and weight loss, bloody stools, fatigue and weakness. Although the exact cause of stomach cancer is not known, several risk factors have been identified. These risk factors include infection with the bacterium Helicobacter pylori, a family history of stomach cancer, consumption of excessively salty foods, smoking, heavy alcohol use and some genetic factors. Diabetes, on the other hand, is a hormonal disorder that regulates the body's blood sugar levels. Normally, an organ called the pancreas controls blood sugar by producing a hormone called insulin. Insulin helps glucose (sugar) enter the cells so that they can make energy. In diabetes, this regulation is disrupted, which can lead to high blood sugar and various health problems. The relationship between stomach cancer and diabetes is not yet fully understood. In this study, machine learning models (Stochastic Gradient Boosting, Bagged Classification and Regression Trees) based on proteomic data were used to predict the diabetes risk of 40 gastric cancer patients, 21 with DM and 19 with non-DM. Performance metrics for the optimal model (Stochastic Gradient Boosting) the accuracy, balanced accuracy, sensitivity, specificity, positive predictive value, negative predictive value and F1-score values are 0.86, 0.83, 0.67, 1.00, 1.00, 0.80, 0.80, respectively. According to the variable importance values obtained as a result of the model, Mucin-13 protein has a positive predictive value in predicting the diabetes risk of gastric cancer patients in the clinic.

## 1. INTRODUCTION

Gastric cancer is a heterogeneous and aggressive malignant tumor. It is the fifth most common cancer worldwide and the third most common cause of cancer-related deaths . Gastric cancer is a disease state with different incidence and mortality rates across continents [1]. Gastric cancer, which is a multifactorial disease when its etiology is analyzed, has many risk factors, both genetic and environmental. The main risk factors for gastric cancer are age, gender and family history; alcohol consumption, smoking and diet are other important risk factors. In addition, pathogens such as Helicobacter pylori and Epstein-Barr virus (EBV) are also important risk factors for the development of gastric cancer [2]. Symptoms of stomach cancer may include weight loss, loss of appetite, indigestion, stomach pain, stomach bleeding. Treatment can usually include surgery, chemotherapy and radiotherapy. However, early diagnosis is important because treatment of early-stage stomach cancer can be more successful.

Diabetes mellitus is a chronic disease characterized by insulin deficiency or defects in insulin secretion, in which the metabolism is unable to utilize carbohydrates, fats and proteins sufficiently and requires continuous medical care [3]. According to the World Health Organization (WHO), the prevalence of diabetes mellitus is increasing worldwide and it is estimated that there will be 300 million people with diabetes in 2025 [4]. On the other hand, diabetes mellitus is known to be associated with several types of cancer, including prostate, breast and colorectal cancer. DM may be linked to cancers of the liver, uterus, and colon, according to recent findings by several studies [5-7]. Studies on the connection between diabetes mellitus (DM) and the onset of gastric cancer are, nevertheless, few [8, 9].

Proteomics is a science that performs a comprehensive analysis of all proteins that cells or organisms express at any one time. This discipline aims to go beyond genomics to understand which proteins are expressed, how they are

 

regulated and how they interact at the cellular level. Proteomics provides important insights, especially for the understanding and treatment of complex diseases, such as cancer [10]. Proteomics is a comprehensive analysis method used to understand the structure, function and quantity of proteins that result from the processing of genetic information. However, effectively understanding and interpreting these large data sets is becoming increasingly complex. This is where machine learning methods come into play in the analysis of proteomics data and bring a new perspective to the world of science. Machine learning is a sub-branch of artificial intelligence that enables algorithms to learn and improve a specific task. Since proteomics data is often large, complex and multidimensional, traditional statistical methods may be insufficient to extract meaningful information from it. This is where machine learning algorithms come in, performing complex analyses such as data mining, classification and prediction [11]. The combination of proteomics and machine learning is playing an important role in biomedical research, becoming a powerful tool for extracting meaningful information from complex protein data. This combination has great potential in areas such as disease diagnostics, treatment strategies and the discovery of new therapeutic targets.

The aim of this study was to classify the proteomic data of 40 gastric cancer samples (19 with DM and 21 without DM) with two different machine learning models based on 37 proteins with regulation differences (up/down) between the two groups and to identify possible DM-related protein biomarkers based on variable significance for the model with the best classification performance.

## 2. MATERIAL and METHODS

### 2.1. Dataset

The dataset used in the study consists of 5982 proteins from 19 DM and 21 non-DM gastric cancer patients. Among these proteins, only 37 proteins had expression differences between DM and non-DM groups. Therefore, the current study was performed using these 37 proteins. Descriptive statistics of the subjects forming the data set are given in Table 1.

TABLE I
DESCRIPTIVE STATISTICS ON GASTRIC CANCER PATIENTS WITH DM AND NON- DM

| | | Group | |
|---|---|---|---|
| | | **DM** | **Non-DM** |
| | | **Median (Min-Max)** | **Median (Min-Max)** |
| **Age** | | 75 (57-85) | 75 (58-87) |
| | | **Count (%)** | **Count (%)** |
| **Gender** | **Female** | 14 (67,7%) | 12 (63,16%) |
| | **Male** | 7 (33,3%) | 7 (36,84%) |

### 2.2. Stochastic Gradient Boosting

Stochastic Gradient Boosting is a variation of the traditional Gradient Boosting algorithm that uses stochastic learning principles to build forecasting models. Gradient Boosting is an ensemble learning method based on combining weak learners (usually decision trees) to form a strong learner. Basically, each learner is added with a focus on correcting the errors of the previous learners. Stochastic Gradient Boosting applies a stochastic learning process, using random samples to train each learner. This means that each tree is trained on a random subset of the data instead of the full dataset. This can increase the generalizability of the model and reduce overfitting [12].

### 2.3. Bagged Classification and Regression Trees (Bagged CART)

Bagged Classification and Regression Trees (Bagged CART)" machine learning method is an ensemble model that builds a model using classification and regression trees (CART) as part of the "bagging" technique. Bootstrap Aggregating (Bagging) is an ensemble learning technique that aims to improve overall performance by bringing together many weak learners. Each learner (model) is trained with a different subset of data. These subsets are "bootstrap" samples generated by random samples from the original dataset. Classification and Regression Trees (CART) is a tree-based learning algorithm for classification and regression. Each tree contains a set of decision nodes and leaf nodes that can classify or regress the dataset based on features. Bagged CART combines these two concepts by performing the following steps. a. First, it creates bootstrap samples from the dataset. Then, it trains a CART tree using each bootstrap sample. These trees can often be deep and prone to overfitting as they are only trained on a subset. Finally, aggregate the prediction of each tree. In the case of classification, a voting method is usually used (e.g. the class with the most votes). In the case of regression, the outputs of the trees can be averaged. Bagged CART can help make the model more general and robust by training each tree on its own subset of data and then aggregating the predictions of these trees. At the same time, this method can reduce problems such as overfitting by a single tree [13].

### 2.3. Biostatistical Analyses, Data Preprocessing and Performance Evaluation of the Models

Data are summarized by median (IQR). Compliance with normal distribution was performed using the Shapiro-Wilk test. Statistically significant differences between two groups were analyzed by Mann-Whitney U test. $p<0.05$ was considered statistically significant. IBM SPSS Statistics 26.0 program was used in the analyses. Elastic net, one of the variable selection methods applied to improve model performance, was used. Elastic Net is a statistical method used for feature selection and regression analysis. Elastic Net combines L1 regularization (LASSO) and L2 regularization (Ridge regression) techniques. This method is often used for efficient variable selection in multicollinear and high dimensional data sets. The elimination of missing values in the data using missing value assignment methods is very important in terms of improving the performance of machine learning algorithms and obtaining robust results. In this study, Random Forest algorithm is used as a missing value assignment method. Afterwards, 80% of the dataset is divided into 80% for training and 20% for testing. A 5-fold cross validation was applied to the training dataset. Accuracy, Balanced accuracy, Sensitivity, Specificity, Positive

predictive value, Negative predictive value, Matthews correlation coefficient (MCC), G-mean and F1-Score metrics in the performance evaluation of machine learning models created to identify candidate biomarkers that can be used in the diagnosis and follow-up of DM and non-DM with gastric cancer.

## 3. RESULTS

After the Elastic Net variable selection method applied to the data set used in the study, 7 out of 37 proteins were included in the model. The results examining the differences between groups in terms of these proteins are given in Table II.

TABLE II
DESCRIPTIVE STATISTICS FOR INPUT PROTEIN VARIABLES

| | Group | | p-value |
|---|---|---|---|
| | DM (n=21) | Non-DM (n=19) | |
| | Median (IQR) | Median (IQR) | |
| **P11678** | 2582308,76 (2221522,06) | 4118604,53 (7415906,97) | **0,045*** |
| **P21980** | 54632822,24(34847203,55) | 38805758,73(22090063,39) | **0,029*** |
| **Q99685** | 1718870,68(2466964,45) | 848188,04(1145769,56) | **0,022*** |
| **P02745** | 353078,21(369858,59) | 207676,19(303358,69) | **0,025*** |
| **Q8WVV4** | 11530092,31(8930673,75) | 3496090,54(5951723,45) | **<0,001*** |
| **A0A2U3TZL5** | 2272312,55(4246602,68) | 1069462,815(1028267,79) | **0,015*** |
| **Q9H3R2** | 7586416,54(7195764,94) | 1791968,52(3319458,48) | **<0,001*** |

*: Mann-Whitney U test.

When the p values in Table II are taken into consideration, the difference between the groups in terms of the taken into the model proteins P11678, P21980, Q99685, P02745, Q8WVV4, A0A2U3TZL5, and Q9H3R2 is statistically significant. The performance metrics of the training and testing phase of the Stochastic Gradient Boosting and Bagged CART machine learning models created with these 7 proteins to classify diabetes in gastric cancer patients are given in Table III and Table IV, respectively.

TABLE III
TEST AND TRAINING MODEL PERFORMANCE METRICS FOR STOCHASTING GRADIENT BOOSTING

| METRICS | TESTING | TRAINING |
|---|---|---|
| | VALUE | VALUE |
| ACCURACY | 0.99 | 0.86 |
| BALANCED ACCURACY | 0.99 | 0.83 |
| SENSITIVITY | 0.99 | 0.67 |
| SPECIFICITY | 0.99 | 1.00 |
| POSITIVE PREDICTIVE VALUE | 0.99 | 1.00 |
| NEGATIVE PREDICTIVE VALUE | 0.99 | 0.80 |
| F1-SCORE | 0.99 | 0.80 |

TABLE IV
TEST AND TRAINING MODEL PERFORMANCE METRICS FOR BAGGED CART

| METRICS | TESTING | TRAINING |
|---|---|---|
| | VALUE | VALUE |
| ACCURACY | 0.99 | 0.71 |
| BALANCED ACCURACY | 0.99 | 0.68 |
| SENSITIVITY | 0.99 | 1.00 |
| SPECIFICITY | 0.99 | 0.33 |
| POSITIVE PREDICTIVE VALUE | 0.99 | 0.67 |
| NEGATIVE PREDICTIVE VALUE | 0.99 | 1.00 |
| F1-SCORE | 0.99 | 0.80 |

Considering Table III and Table IV, where the metrics for the classification performance of the models are given, the best classification model is Stochastic Gradient Boosting. Therefore, Table III shows that for Stochastic Gradient Boosting, the accuracy, balanced accuracy, sensitivity, specificity, positive predictive value, negative predictive value and F1-score values are 0.86, 0.83, 0.67, 1.00, 1.00, 0.80, 0.80, respectively. The importance ranks of the proteins obtained from the best performing model, which can be used as possible protein biomarkers for DM/non-DM classification with gastric cancer patients, are given in Table V and Figure I.

TABLE V
VARIABLE IMPORTANCE OBTAINED FROM THE STOCHASTIC GRADIENT BOOSTING MODEL

| Accession | Protein Name | Importance |
|---|---|---|
| Q9H3R2 | Mucin-13 | 100 |
| P21980 | Protein-glutamine gamma-glutamyltransferase 2 | 25.653 |
| P02745 | Complement C1q subcomponent subunit A | 24.602 |
| Q8WVV4 | Protein POF1B | 19.729 |
| A0A2U3TZL5 | CD59 molecule | 15.375 |
| P11678 | Eosinophil peroxidase | 10.622 |
| Q99685 | Monoglyceride lipase | 6.404 |

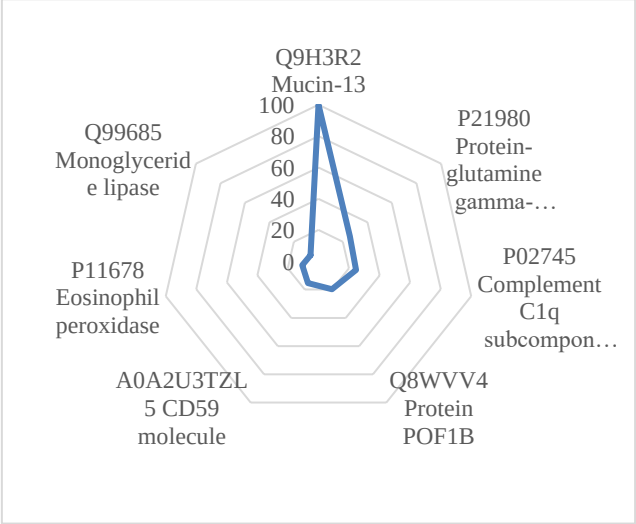


Fig.1. Radar plot of variable importance obtained from the Stochastic Gradient Boosting model

## 3. DISCUSSION

Gastric cancer is a major health problem with high mortality rates worldwide. Although it is more common in certain geographical regions, genetic and environmental factors may also be effective in gastric cancer, which is a type of cancer that usually affects middle-aged and older adults.

 ISSN: 2548-0650 http://dergipark.gov.tr/jcs

Genetic mutations, inflammation and cellular damage play a role in the pathogenesis of gastric cancer. Adenocarcinoma is the most common histologic type of gastric cancer. Tumor size, degree of invasion and lymph node metastases are critical in the staging of the disease. Risk factors that play a role in the development of gastric cancer include genetic predisposition, family history, Helicobacter pylori infection, tobacco and alcohol consumption and dietary habits. Diabetes is a common health problem worldwide and is characterized by high levels of sugar in the blood due to metabolic disorders. Diabetes has an increasing prevalence and usually increases with age. Risk factors involved in the development of diabetes include genetic predisposition, obesity, sedentary lifestyle, poor eating habits and age. In addition, conditions such as gestational diabetes, hypertension and metabolic syndrome may also increase the risk of diabetes. Although the relationship between stomach cancer and diabetes has not been fully elucidated, it is an issue that needs to be addressed. Lin et al. observed an increased risk of gastric cancer in diabetic patients regardless of their body mass index and reported that hyperglycemia is a possible risk factor for energy/metabolism imbalance and that impaired immune system is a cause of gastric cancer [14]. In a more recent study, a meta-analysis by Mansory et al. observed a positive association between Helicobacter pylori infection and DM [15].

In this study, two different machine learning methods (Staochasting Gradient Boosting, Bagged CART) were applied to classify DM (n=21) and non-DM (n=19) with proteomic data obtained as a result of label-free proteomic analysis applied to samples of 40 gastric cancer patients. Stochastic Gradient Boosting and Bagged CART machine learning methods used for classification, Stochastic Gradient Boosting has the best performance metrics with the accuracy, balanced accuracy, sensitivity, specificity, positive predictive value, negative predictive value and F1-score values are 0.86, 0.83, 0.67, 1.00, 1.00, 0.80, 0.80, respectively. According to the optimal model, the top most important protein for DM and non-DM classification are Q9H3R2. Mucin-13 protein, coded Q9H3R2, which may be associated with diabetes by the optimal model and has the highest significance, is a transmembrane glycoprotein secreted in the digestive tract and was reported to be overexpressed in intestinal-type gastric cancer in the 2005 study by Shimamura et al [16]. On the other hand, another study in 2023 suggested that overexpression of mucin family proteins can trigger apoptosis in placental tissues of women with gestational diabetes [17]. Therefore, it is thought that further studies on the protein in question may be instructive in explaining the risk of diabetes in gastric cancer patients.

In conclusion, these two proteins proposed as possible biomarkers for the diagnosis of diabetes in gastric cancer patients based on the Stochastic gradient boosting model may be very useful in the clinic. In addition, as far as we know, there is no study in the literature that classifies diabetes and identifies possible biomarkers with the help of machine learning in gastric cancer patients. It is thought that this study will contribute to the literature in this sense.

## REFERENCES


[1] M.H.S. Jong, S.S. Gisbertz, M. I. Berge Henegouwen, W. A. Draaisma (2023). Prevalence of nodal metastases in the individual lymph node stations for different T-stages in gastric cancer: a systematic review. Updates in Surgery, 75(2), 281-290.

[2] J. Machlowska, J. Baj, M. Sitarz, R. Maciejewski, R. Sitarz (2020) Gastric cancer: epidemiology, risk factors, classification, genomic characteristics and treatment strategies. International journal of molecular sciences, vol. 21, p. 4012.

[3] D. Tomic, J. E. Shaw, and D. J. Magliano (2022) The burden and risks of emerging complications of diabetes mellitus. Nature Reviews Endocrinology, vol. 18, pp. 525-539.

[4] I. Satman, T. Yilmaz, A. Sengul, S. Salman, F. Salman, S. Uygur, et al. (2002) Population-based study of diabetes and risk characteristics in Turkey: results of the turkish diabetes epidemiology study (TURDEP). Diabetes care, vol. 25, pp. 1551-1556.

[5] H. B. El–Serag, H. Hampel, and F. Javadi (2006) The association between diabetes and hepatocellular carcinoma: a systematic review of epidemiologic evidence. Clinical Gastroenterology and Hepatology, vol. 4, pp. 369-380.

[6] E. Friberg, N. Orsini, C. Mantzoros, and A. Wolk (2007) Diabetes mellitus and risk of endometrial cancer: a meta-analysis. Diabetologia, vol. 50, pp. 1365-1374.

[7] S. C. Larsson, N. Orsini, and A. Wolk (2005) Diabetes mellitus and risk of colorectal cancer: a meta-analysis. Journal of the National Cancer Institute, vol. 97, pp. 1679-1687.

[8] A. Sekikawa, H. Fukui, T. Maruo, T. Tsumura, Y. Okabe, and Y. Osaki (2014) Diabetes mellitus increases the risk of early gastric cancer development. European journal of cancer, vol. 50, pp. 2065-2071.

[9] H.-J. Yang, D. Kang, Y. Chang, J. Ahn, S. Ryu, J. Cho, et al. (2020) Diabetes mellitus is associated with an increased risk of gastric cancer: a cohort study. Gastric Cancer, vol. 23, pp. 382-390.

[10] Y. W. Kwon, H.-S. Jo, S. Bae, Y. Seo, P. Song, M. Song, et al. (2021) Application of proteomics in cancer: recent trends and approaches for biomarkers discovery. Frontiers in Medicine, vol. 8, p. 747333.

[11] H. Desaire, E. P. Go, and D. Hua (2022) Advances, obstacles, and opportunities for machine learning in proteomics. Cell Reports Physical Science, vol. 3.

[12] E. A. Freeman, G. G. Moisen, J. W. Coulston, and B. T. Wilson (2016) Random forests and stochastic gradient boosting for predicting tree canopy cover: comparing tuning processes and model performance. Canadian Journal of Forest Research, vol. 46, pp. 323-339.

[13] H. Duan, Z. Deng, F. Deng, and D. Wang (2016) Assessment of groundwater potential based on multicriteria decision making model and decision tree algorithms. Mathematical Problems in Engineering, vol. 2016, pp. 1-11.

[14] S.-W. Lin, N. D. Freedman, A. R. Hollenbeck, A. Schatzkin, and C. C. Abnet (2011) Prospective study of self-reported diabetes and risk of upper gastrointestinal cancers. Cancer epidemiology, biomarkers & prevention, vol. 20, pp. 954-961.

[15] K. Mansori, Y. Moradi, S. Naderpour, R. Rashti, A. B. Moghaddam, L. Saed, et al. (2020) Helicobacter pylori infection as a risk factor for diabetes: a meta-analysis of case-control studies. BMC gastroenterology, vol. 20, pp. 1-14.

[16] T. Shimamura, H. Ito, J. Shibahara, A. Watanabe, Y. Hippo, H. Taniguchi, et al. (2005) Overexpression of MUC13 is associated with intestinal-type gastric cancer. Cancer science, vol. 96, pp. 265-273.

[17] S.-S. Cui, P. Zhang, L. Sun, Y.-L.-L. Yuan, J. Wang, F.-X. Zhang, et al. (2023) Mucin1 induced trophoblast dysfunction in gestational diabetes mellitus via Wnt/β-catenin pathway. Biological Research, vol. 56, p. 48.


## BIOGRAPHIES

**Şeyma YAŞAR** obtained her BSc. degree in mathematics from GaziosmanPaşa University in 2009. She received MSc. degree in biostatistics and medical informatics from the Inonu University in 2018. She currently continues Ph.D. degrees in biostatistics and medical informatics from the Inonu University. In 2014, she joined the Department of Biostatistics and Medical Informatics at Inonu University as a researcher assistant. Her research interests are cognitive systems, data mining, machine learning, deep learning, proteomics, bioinformatics.

**Büşra Nur FINDIK** obtained her BSc. degree in Phsiotherapy and Rehabilitation from Istanbul Bilgi University in 2012. She received MSc. degree in Norological Rehabilitation from the Marmara University in 2020. She currently continues Ph.D. degrees in Phsiotherapy and Rehabilitation from the Sağlık Bilimleri University. In 2022, she joined the Department of Therapy and Rehabilitation at Nevsehir Haci Bektas Veli University, Kozakli Vocational School as a lecturer. Her research interests are physical activity, chest phsiotherapy, cardiac rehabilitation.

 ISSN: 2548-0650 http://dergipark.gov.tr/jcs