

Original Article

A Machine Learning Risk Prediction Model for Gastric Cancer with SHapley Additive exPlanations

Bomi Park¹, Chung Ho Kim¹, Jae Kwan Jun², Mina Suh², Kui Son Choi², Il Ju Choi³, Hyun Jin Oh⁴¹Department of Preventive Medicine, Chung-Ang University College of Medicine, Seoul, ²National Cancer Control Institute, National Cancer Center, Goyang, ³Division of Gastroenterology, Department of Internal Medicine, Center for Gastric Cancer, National Cancer Center, Goyang,⁴Division of Gastroenterology, Department of Internal Medicine, Center for Cancer Prevention and Detection, National Cancer Center, Goyang, Korea

Purpose Gastric cancer (GC) prediction models hold potential for enhancing early detection by enabling the identification of high-risk individuals, facilitating personalized risk-based screening, and optimizing the allocation of healthcare resources.

Materials and Methods In this study, we developed a machine learning-based GC prediction model utilizing data from the Korean National Health Insurance Service, encompassing 10,515,949 adults who had not been diagnosed with GC and underwent GC screening during 2013-2014, with a follow-up period of 5 years. The cohort was divided into training and test datasets at an 8:2 ratio, and class imbalance was mitigated through random oversampling.

Results Among various models, logistic regression demonstrated the highest predictive performance, with an area under the receiver operating characteristic curve (AUC) of 0.708, which was consistent with the AUC obtained in external validation (0.669). Importantly, the outcomes were robust to missing data imputation and variable selection. The SHapley Additive exPlanations (SHAP) algorithm enhanced the explainability of the model, identifying advancing age, being male, *Helicobacter pylori* infection, current smoking, and a family history of GC as key predictors of elevated risk.

Conclusion This predictive model could significantly contribute to the early identification of individuals at elevated risk for GC, thereby enabling the implementation of targeted preventive strategies. Furthermore, the integration of noninvasive and cost-effective predictors enhances the clinical utility of the model, supporting its potential application in routine healthcare settings.

Key words Stomach neoplasms, Prediction model, Machine learning, SHapley Additive exPlanations

Introduction

Gastric cancer (GC) remains a significant global health issue, ranking as the fifth most common malignancy and the fourth leading cause of cancer-related mortality worldwide [1]. Despite a decline in incidence in certain regions, GC persists as a major concern, particularly in Korea, where it was the fourth leading cause of cancer deaths and the second most prevalent cancer in 2020 [2].

Prognostic outcomes for GC are highly variable, contingent upon the clinical stage at diagnosis. Patients with metastatic GC exhibit a poor prognosis, with a 5-year survival rate of approximately 7%, contrasting sharply with the 75% 5-year survival rate observed in patients with localized GC [3]. This disparity underscores the critical importance of early detection in improving patient outcomes. The development of GC typically follows a progressive sequence of precancerous stages, including chronic gastritis, atrophic gastritis, intestinal metaplasia, and dysplasia.

Early-stage GC, however, is often asymptomatic or presents with nonspecific symptoms, leading to delays in diagnosis [4]. As a result, routine screening is imperative for the timely identification of GC, with prior studies demonstrating a correlation between regular screening and reduced GC mortality [5,6]. In Korea, the National Cancer Screening Program (NCSP) offers biennial gastroscopy to individuals over 40 years of age, which has been shown to significantly increase the likelihood of detecting GC at a localized stage compared to those who have never been screened [7]. Additionally, participation in the NCSP has been associated with lower GC mortality [8].

Nevertheless, the widespread implementation of population-based screening programs incurs substantial costs. For example, the biennial gastroscopy program for individuals over 40 in Korea is funded by an annual national budget allocation exceeding 250 million dollars. Moreover, gastric endoscopy, while valuable, is not without risks, including potential complications such as infection, bleeding, and per-

Correspondence: Hyun Jin Oh

Division of Gastroenterology, Department of Internal Medicine, Center for Cancer Prevention and Detection, National Cancer Center, 323 Ilsan-ro, Ilsandong-gu, Goyang 10408, Korea

Tel: 82-31-290-1759 E-mail: hyun.jin.8411@gmail.com

Received August 29, 2024 Accepted December 15, 2024 Published Online December 16, 2024

foration [9]. Therefore, optimizing screening intervals and integrating personalized prevention strategies through risk-based stratification are essential. GC prediction models serve as practical tools for stratifying populations according to GC risk, thereby facilitating the development of more targeted and cost-effective screening and prevention protocols.

While several GC prediction models have been previously proposed [10,11], many fail to incorporate critical risk factors, such as *Helicobacter pylori* infection or histologically diagnosed gastric adenoma. Furthermore, most existing models rely predominantly on traditional methodologies, such as the Cox proportional hazard model (CPHM), with limited utilization of machine learning approaches. This study sought to develop a machine learning-based GC prediction model that integrates key GC risk factors. However, machine learning models are frequently criticized for their lack of transparency and explanation [12]. To address this challenge, we employed a recently proposed explainable artificial intelligence algorithm to enhance the interpretability of our machine learning-based prediction model, ensuring that the reasoning process is more accessible and comprehensible.

Materials and Methods

In Korea, GC screening is provided biennially by the NCSP to individuals aged over 40 years. Approximately 50%-60% of the Korean population participates in this program, and the gastroscopy results are systematically recorded in the National Health Insurance Service (NHIS) database. This study established a comprehensive longitudinal health information dataset, integrating a qualification database (a national medical check-up database encompassing health-related behavior data, family history, and screening results) and a treatment database that detailed disease types. Utilizing this dataset, the study followed adults aged 40-74 years who underwent gastroscopy screening for GC through the NCSP in 2013 and 2014. Individuals diagnosed with GC prior to screening or those who passed away within 5 years of screening were excluded from the analysis. For external validation, the Cancer Screenee Cohort of the National Cancer Center, comprising a dataset of 31,285 individuals, was employed to assess the generalizability and performance of the predictive models on independent datasets.

The outcome variable was defined as binary, indicating the presence or absence of GC. GC cases were identified using the International Classification of Diseases, 10th revision (ICD-10) code C16, either as a primary or secondary diagnosis, along with the special codes V193 or V194. The study population was followed for 5 years post-screening to ascertain the development of cancer, categorizing participants

into GC and non-GC groups based on their diagnosis.

To predict GC occurrence within 5 years, the study incorporated 17 variables previously identified as GC risk factors: age, sex, body mass index (BMI), smoking, alcohol consumption, family history of GC, colorectal cancer, or liver cancer, and history of hypertension, diabetes mellitus, myocardial infarction/angina pectoris, stroke, dyslipidemia, colorectal cancer, liver cancer, *H. pylori* infection, or gastric adenoma. The previous use of an *H. pylori* eradication regimen, indicative of an individual's *H. pylori* history, was determined by analyzing prescription history data in the treatment database. Gastric adenoma diagnosis was confirmed through gastric endoscopy and gastric tissue biopsy results. For sensitivity analysis, feature selection from the aforementioned 17 risk factors was performed using stepwise variable selection or least absolute shrinkage and selection operator (LASSO) regression, aimed at removing potentially redundant factors and mitigating the risk of data overfitting [13].

After excluding observations with missing data for any risk factor, a total of 10,515,949 individuals (men, 4,678,843; women, 5,837,106) with complete data were included in the analysis. To address the substantial amount of missing data, particularly in self-reported lifestyle variables within the NHIS dataset, we conducted a sensitivity analysis using multiple imputations ($m=5$) via the Multivariate Imputation by Chained Equations (MICE) package.

Given the low prevalence of GC in the population, which may result in data imbalance, we implemented Random Over-Sampling Examples (ROSE) on the training dataset to achieve a balanced ratio of 5:5 between individuals with and without GC prior to developing the prediction model. We employed a chi-squared test to compare the characteristics of individuals with GC versus those without GC.

The NHIS dataset was randomly divided into training and test datasets at an 8:2 ratio. We then developed three machine learning-based prediction models: logistic regression, decision tree, and eXtreme Gradient Boosting (XGBoost). Hyperparameter tuning for the logistic regression and decision tree models was conducted through fivefold cross-validation on the training set, while the boosting iteration for the XGBoost model was fixed at 500. The performance of these prediction models was assessed based on accuracy, sensitivity, specificity, and the area under the receiver operating characteristic curve (AUROC).

A significant challenge in applying machine learning models is the opacity of feature-response relationships, which limits the interpretability and practical utility of the models for clinicians. To enhance interpretability, we integrated the SHapley Additive exPlanations (SHAP) algorithm into the top-performing model (i.e., the model with the highest AUROC value on the external validation set). The SHAP

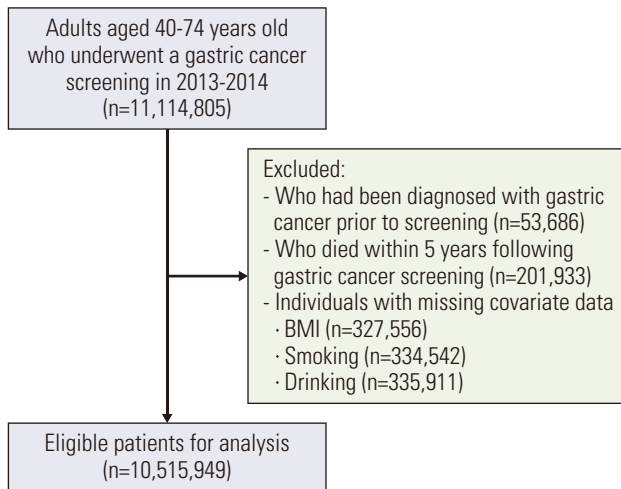


Fig. 1. Flowchart of the study participants. BMI, body mass index.

method, grounded in game theory, offers insights into the contribution of each feature to the model's predictions. All analyses were conducted using R software ver. 3.6.4 (R Foundation for Statistical Computing), and the study's reporting adhered to the STROBE guidelines.

Results

The application of exclusion criteria resulted in a final cohort of 10,515,949 individuals, comprising 65,657 with GC and 10,450,292 without GC, as derived from the NHIS (Fig. 1). The characteristics of the study participants are detailed in Table 1.

We developed three machine learning-based models for GC prediction. The receiver operating characteristic curves generated from the application of these models to the internal validation dataset are presented in Fig. 2. The performance metrics of the models in both internal and external validation datasets are summarized in Table 2. Among the models, logistic regression demonstrated the highest AUROC at 0.708 (95% confidence interval [CI], 0.704 to 0.710) and the greatest sensitivity at 0.670 (95% CI, 0.661 to 0.671). However, it exhibited the lowest accuracy at 0.637 (95% CI, 0.636 to 0.637) and the third highest specificity at 0.643 (95% CI, 0.630 to 0.640). Conversely, the decision tree model outperformed others in terms of accuracy (0.795; 95% CI, 0.795 to 0.796) and specificity (0.797; 95% CI, 0.790 to 0.800). Notably, in the independent external validation dataset, the logistic regression model maintained the highest AUROC at 0.669 (95% CI, 0.580 to 0.710).

Further analysis, including missing data imputation, con-

firmed the robustness of the logistic regression model, which continued to yield the highest AUROC (0.712; 95% CI, 0.708 to 0.717) (S1 Table, S2A Fig.). Stepwise implementation of missing imputation and variable selection did not produce significantly different results (S3 Table, S2B Fig.). Similarly, the integration of missing imputation with variable selection using the LASSO technique did not substantially alter the outcomes (S4 Table, S2C Fig.).

We further analyzed the important features of the 17 variables from the logistic regression model that had the highest AUROC. SHAP values were computed to rank these variables based on their impact on outcome prediction (Fig. 3). The variables are arranged in descending order of importance according to their SHAP values, which are plotted along the y-axis. For each case, the SHAP values are depicted along the x-axis (Fig. 3A), with each point representing a predicted case, and the color indicating the feature's corresponding value. Positive and negative SHAP values signify the prediction of GC presence and absence, respectively. In Fig. 3B, bar lengths correspond to SHAP values, with longer bars indicating a more significant contribution by the variable to GC prediction. The analysis identified age, sex, *H. pylori* infection, smoking status, and family history of GC as the most critical predictors of GC. The likelihood of developing GC increased with advancing age, particularly among males, and was elevated in individuals with a history of *H. pylori* infection, current smokers, and a family history of GC. Additional factors associated with an elevated risk included a higher BMI, the absence of hypertension, frequent alcohol consumption, the presence of gastric adenoma, and a history of diabetes.

Discussion

In this study, we developed and rigorously evaluated machine learning-based models for predicting GC risk. Among the models tested, logistic regression demonstrated superior performance, as indicated by the AUROC values. Given the imbalanced nature of the dataset, AUROC values were employed as the primary metric for model evaluation, in lieu of accuracy metrics [14].

The significance of early screening and diagnosis in reducing GC mortality rates cannot be overstated, as timely intervention markedly improves patient outcomes. The prognosis for patients diagnosed with early-stage GC is notably favorable, with 5-year survival rates approaching 95% [15,16]. While gastric endoscopy remains the gold standard for GC detection, its invasive nature limits its widespread application [14]. Risk prediction models that identify individuals at high risk for GC are crucial in facilitating targeted screen-

Table 1. Baseline characteristics of the study population

Characteristic	Total	Patients with GC	Individuals without GC	p-value
Age (yr)				
40-44	2,178,995 (20.7)	4,256 (6.5)	2,174,739 (20.8)	< 0.001
45-49	1,643,147 (15.6)	5,303 (8.1)	1,637,844 (15.7)	
50-54	1,850,019 (17.6)	9,031 (13.8)	1,840,988 (17.6)	
55-59	1,707,740 (16.2)	11,752 (17.9)	1,695,988 (16.2)	
60-64	1,305,544 (12.4)	12,273 (18.7)	1,293,271 (12.4)	
65-69	1,042,466 (9.9)	12,341 (18.8)	1,030,125 (9.9)	
70-74	788,038 (7.6)	10,701 (16.2)	777,337 (7.4)	
Sex				
Male	4,678,843 (44.5)	41,609 (63.4)	4,637,234 (44.4)	< 0.001
Female	5,837,106 (55.5)	24,048 (36.6)	5,813,058 (55.6)	
BMI (kg/m²)				
< 23	4,054,189 (38.5)	23,245 (35.4)	4,030,944 (38.5)	< 0.001
23-24.9	2,719,298 (25.9)	17,382 (26.5)	2,701,916 (25.9)	
25-29.9	3,321,479 (31.6)	22,640 (34.5)	3,298,839 (31.6)	
≥ 30	420,983 (4.0)	2,390 (3.6)	418,593 (4.0)	
Smoking				
Nonsmoker	6,943,164 (66.0)	34,788 (53.0)	6,908,376 (66.1)	< 0.001
Ex-smoker	1,664,621 (15.8)	15,268 (23.2)	1,649,353 (15.8)	
Smoker	1,908,164 (18.2)	15,601 (23.8)	1,892,563 (18.1)	
Drinking				
Nondrinker	6,191,488 (58.9)	36,282 (55.3)	6,155,206 (58.9)	< 0.001
≤ 3 times/wk	3,851,862 (36.6)	24,288 (37.0)	3,827,574 (36.6)	
≥ 4 times/wk	472,599 (4.5)	5,087 (7.7)	467,512 (4.5)	
Family history of gastric cancer				
No	9,566,290 (91.0)	57,688 (87.9)	9,508,602 (91.0)	< 0.001
Yes	949,659 (9.0)	7,969 (12.1)	941,690 (9.0)	
Family history of colorectal cancer				
No	10,198,797 (97.0)	63,763 (97.1)	10,135,034 (97.0)	0.048
Yes	317,152 (3.0)	1,894 (2.9)	315,258 (3.0)	
Family history of liver cancer				
No	10,092,512 (96.0)	63,050 (96.0)	10,029,462 (96.0)	0.470
Yes	423,437 (4.0)	2,607 (4.0)	420,830 (4.0)	
Hypertension				
No	9,889,820 (94.0)	61,296 (93.4)	9,828,524 (94.1)	< 0.001
Yes	626,129 (6.0)	4,361 (6.6)	621,768 (5.9)	
Diabetes				
No	9,677,054 (92.0)	56,877 (86.6)	9,620,177 (92.1)	< 0.001
Yes	838,895 (8.0)	8,780 (13.4)	830,115 (7.9)	
Myocardial infarction/Angina pectoris				
No	10,381,003 (98.7)	64,212 (97.8)	10,316,791 (98.7)	< 0.001
Yes	134,946 (1.3)	1,445 (2.2)	133,501 (1.3)	
Stroke				
No	10,427,928 (99.2)	64,742 (98.6)	10,363,186 (99.2)	< 0.001
Yes	88,021 (0.8)	915 (1.4)	87,106 (0.8)	
Dyslipidemia				
No	10,026,491 (95.4)	62,009 (94.4)	9,964,482 (95.4)	< 0.001
Yes	489,458 (4.6)	3,648 (5.6)	485,810 (4.6)	

(Continued to the next page)

Table 1. Continued

Characteristic	Total	Patients with GC	Individuals without GC	p-value
Colorectal cancer				
No	10,508,375 (99.9)	65,587 (99.9)	10,442,788 (99.9)	0.001
Yes	7,574 (0.1)	70 (0.1)	7,504 (0.1)	
Liver cancer				
No	10,508,677 (99.9)	65,619 (99.9)	10,443,058 (99.9)	0.303
Yes	7,272 (0.1)	38 (0.1)	7,234 (0.1)	
Helicobacter pylori infection				
No	10,287,130 (97.8)	63,463 (96.7)	10,223,667 (97.8)	< 0.001
Yes	228,819 (2.2)	2,194 (3.3)	226,625 (2.2)	
Gastric adenoma				
No	10,486,163 (99.7)	63,025 (96.7)	10,423,138 (99.7)	< 0.001
Yes	29,786 (0.3)	2,632 (3.3)	27,154 (0.3)	
Incidence of gastric cancer within 5 years				
No	10,450,292 (99.4)	-	-	-
Yes	65,657 (0.6)	-	-	

Values are presented as number (%). BMI, body mass index; GC, gastric cancer.

ing, thereby reducing unnecessary procedures in low-risk groups and optimizing resource allocation for early detection. Additionally, these models enable the personalization of screening schedules based on individual risk profiles, thereby enhancing post-endoscopic health outcomes. Our study presents a machine learning-based prediction model specifically designed to identify individuals at high risk for GC within the general population. This model is distinctive in its reliance on data derived primarily from the national health screening program, which conducts comprehensive screenings on an annual or biannual basis. The dataset incorporates a range of noninvasive, cost-effective, and straightforward variables, including lifestyle factors, family history, and medication history. By utilizing existing data, our model enhances accessibility and ease of use in clinical settings, allowing for practical application without incurring substantial costs or necessitating invasive diagnostics.

Recognizing that the performance of prediction models is contingent upon the selection of appropriate predictors, we conducted a thorough literature review to identify epidemiologically significant GC risk factors. These variables were then incorporated into the model development process, prioritizing domain knowledge over indiscriminate variable inclusion or exclusive reliance on statistical selection methods. Notably, our findings indicate that the AUROC values of classifiers did not significantly improve when employing features selected through stepwise variable selection or LASSO, as compared to the original model, which utilized a comprehensive set of features curated by researchers based on their expertise.

To optimize GC interventions, it is imperative to under-

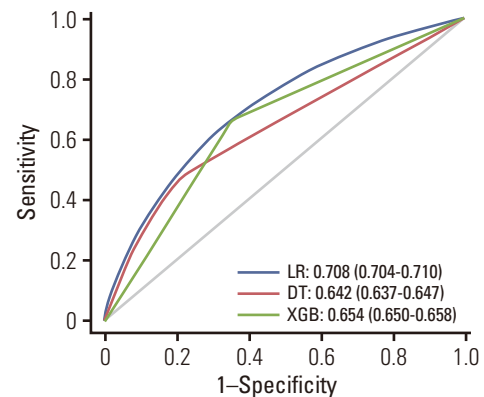


Fig. 2. AUROCs of the prediction models. AUROC, area under the receiver operating characteristic curve; DT, decision tree; LR, logistic regression; XGB, eXtreme Gradient Boosting.

stand the ranked impact of contributing factors, which in turn necessitates prioritizing actions that target preventable elements involved in GC development. To enhance the interpretability of our predictive model, we employed the SHAP algorithm. Through this analysis, we identified age, sex, *H. pylori* infection, smoking, and a family history of GC as the most significant predictors of GC. Incorporating the SHAP algorithm into GC prediction models enables clinicians to deliver personalized preventive interventions, addressing the most impactful factors for each patient. While previous studies have employed machine learning to predict GC occurrence [14,16-22], this is the first study to address the “black box” issue of machine learning-based GC prediction models using an interpretable method such as the SHAP

Table 2. Performance of various machine learning-based models without missing imputation

	Accuracy	Sensitivity	Specificity	AUROC
Internal validation set				
LR	0.637 (0.636-0.637)	0.670 (0.661-0.671)	0.643 (0.630-0.640)	0.708 (0.704-0.710)
DT	0.795 (0.795-0.796)	0.467 (0.460-0.481)	0.797 (0.790-0.800)	0.642 (0.637-0.647)
XGBoost	0.648 (0.647-0.649)	0.659 (0.658-0.680)	0.648 (0.641-0.649)	0.654 (0.650-0.658)
External validation set				
LR	0.676 (0.670-0.680)	0.676 (0.671-0.682)	0.736 (0.690-0.782)	0.669 (0.580-0.710)
DT	0.823 (0.821-0.830)	0.822 (0.817-0.826)	0.666 (0.615-0.718)	0.440 (0.350-0.540)
XGBoost	0.481 (0.480-0.490)	0.481 (0.475-0.487)	0.541 (0.496-0.585)	0.601 (0.510-0.690)

Values in parentheses represent the 95% confidence intervals. AUROC, area under the receiver operating characteristic curve; DT, decision tree; LR, logistic regression; XGBoost, eXtreme Gradient Boosting.

algorithm.

The likelihood of developing GC was observed to increase with advancing age, being male, a history of *H. pylori* infection, current smoking, and a family history of GC. Other factors associated with an increased risk included a higher BMI, the absence of hypertension, frequent alcohol consumption, the presence of gastric adenoma, and a history of diabetes.

Our findings align with established knowledge, confirming that *H. pylori* infection, smoking, and a family history of GC are key risk factors, alongside non-modifiable variables such as age and sex [23-27]. Given that behavioral factors and *H. pylori* infection are modifiable, it is essential to consider these elements when developing effective GC prevention strategies. *H. pylori*, a class I carcinogen, is recognized as the leading cause of GC [25-27]. Considering its high prevalence in Korea [28], anti-*H. pylori* interventions are of paramount importance. Our results underscore the critical role of *H. pylori* in GC pathogenesis and the necessity for targeted interventions in populations with high prevalence rates.

Furthermore, the increased risk of GC among current smokers emphasizes the need for comprehensive smoking cessation programs. Individuals with a family history of GC—a significant risk factor—would benefit from intensified gastroscopy surveillance. While the impact of BMI as a risk factor appears modest, weight management in individuals with a BMI of 30 or higher may still provide preventive benefits. Additionally, reducing the frequency of alcohol consumption could contribute to lowering GC risk.

This study has several limitations. First, the control group consisted of individuals who were not diagnosed with GC within 5 years following GC screening. However, this group may have included individuals who could develop GC in the future, potentially influencing the study outcomes. Additionally, individuals who passed away within 5 years of screening were excluded from the analysis, as some of these individuals may have later been diagnosed with GC. Due to constraints within the closed analytical environ-

ment—where the analysis of national health insurance claims data was restricted to designated computers within a secure network—we were unable to use the version of R required to perform survival machine learning and account for competing risks. This technical limitation hindered the implementation of survival machine learning in this study. Future research should aim to address this limitation by employing survival machine learning approaches. Second, the model's scope was constrained by the limited range of variables included, necessitating cautious interpretation of results based solely on the evaluated variables. Notably, key risk factors for GC, such as chronic atrophic gastritis and salt consumption, were excluded from the analysis due to unavailability of related data. Expanding the range of variables considered could enhance the model's performance. Third, *H. pylori* infection status was inferred from medication history rather than direct diagnosis, which may have led to an underestimation of GC risk in the *H. pylori*-infected cohort, as it included individuals who had undergone treatment that could have eradicated the infection. Fourth, the application of machine learning techniques required careful hyperparameter tuning, which is crucial for influencing classification outcomes. However, due to the aforementioned constraints of the analytical environment, our ability to perform comprehensive hyperparameter tuning was limited. With access to a more advanced computational environment, additional tuning could significantly improve model performance. Finally, there is a potential overlap between individuals in the internal validation set and the external validation dataset. Although this overlap is likely minimal, it may introduce some bias into the validation results by affecting the independence of the datasets.

Despite these limitations, a significant strength of this study lies in the use of a large, nationwide screening dataset, encompassing 10,515,949 individuals, to develop a machine learning-based GC prediction model that focuses on lifestyle, noninvasive characteristics, and major risk factors, includ-

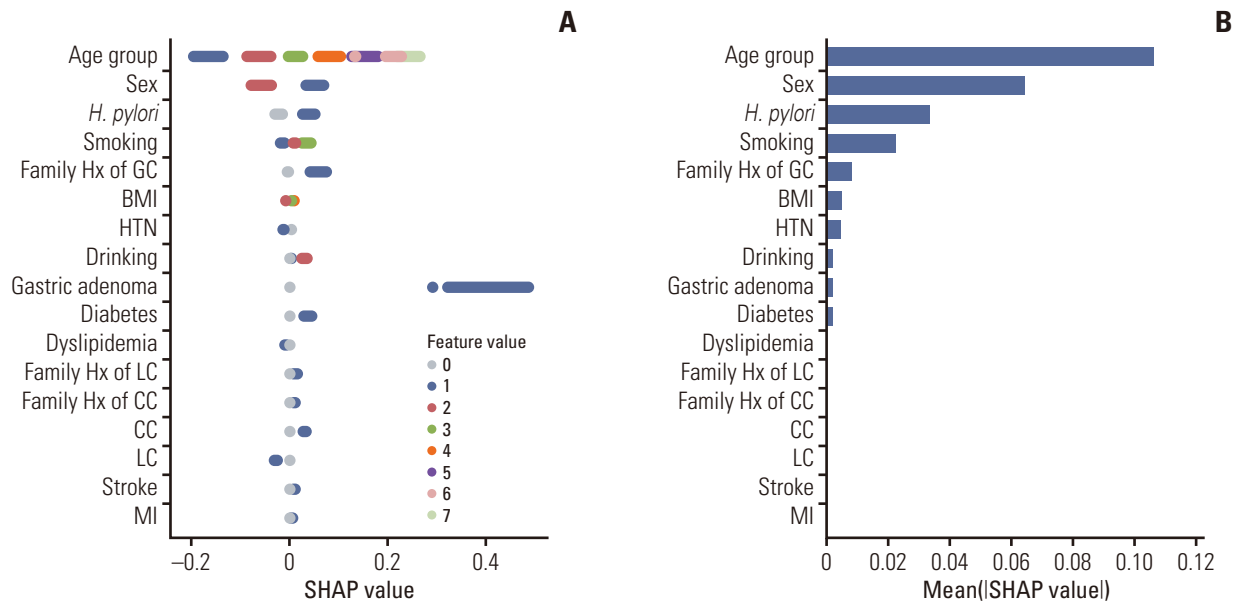


Fig. 3. Summary of SHapley Additive exPlanations (SHAP) values. (A) Each dot represents the impact of a feature on one subject. The dot's color indicates the feature's value, while its position on the x-axis indicates the SHAP value, reflecting the feature's contribution to altering the model's prediction for that individual. Features are plotted on the y-axis and organized in descending order based on mean SHAP values. Variables and coding for analysis: age group (1, 40-44; 2, 45-49; 3, 50-54; 4, 55-59; 5, 60-64; 6, 65-69; 7, 70-74), sex (1, male; 2, female), *Helicobacter pylori* infection (1, infection; 0, no infection), smoking status (1, nonsmoker; 2, former smoker; 3, current smoker), Family history of gastric cancer (1, yes; 0, no), body mass index (BMI; 1, < 23; 2, 23-24.9; 3, 25-29.9; 4, ≥ 30), alcohol consumption (0, no drinking; 1, ≤ 3 times/wk; 2, ≥ 4 times/wk), disease history or family history of disease (1, yes; 0, no). (B) Mean absolute SHAP values. The five most influential features are age, sex, *H. pylori*, smoking, and family history of gastric cancer (GC). CC, colorectal cancer; HTN, hypertension; Hx, history; LC, liver cancer; MI, myocardial Infarction.

ing *H. pylori* infection. The adoption of a machine learning approach offers distinct advantages, such as optimizing feature selection, capturing complex nonlinear relationships, and uncovering hidden patterns within the data, thereby surpassing the predictive accuracy of traditional models like the CPHM [29]. Notably, our findings confirmed the superiority of the machine learning model over the CPHM in predicting GC risk. To evaluate comparative performance, we developed a GC prediction model using the CPHM and conducted a direct comparison with several machine learning algorithms. Our analysis demonstrated that the machine learning algorithms consistently outperformed the CPHM in identifying individuals at risk for gastric cancer. Specifically, the AUROC for the CPHM was 0.516 (95% CI, 0.514 to 0.519), which was significantly lower than the AUROC values obtained from all machine learning models (S5 Table, S6 Fig.). This marked improvement in performance underscores the scientific merit of adopting machine learning in this analysis, particularly for addressing key limitations of the current gold standard, most likely the CPHM. The CPHM estimates the effects of multiple variables on survival outcomes by incorporating time-to-event data and generating hazard

ratios (HRs) for each variable. However, it operates under the assumption of a constant HR over time—a simplification that may not adequately reflect the dynamic nature of disease progression. Additionally, the CPHM's handling of missing data poses challenges, potentially limiting its effectiveness in risk prediction. By contrast, machine learning models address these constraints, achieving greater predictive accuracy through optimized feature selection, the ability to capture intricate nonlinear relationships, and the detection of latent patterns in the dataset [14,29].

Furthermore, the robustness and generalizability of the prediction model were validated through external testing on an independent dataset. Additionally, an important benefit of our machine learning-based approach is its ability to elucidate which variables contribute to an increased risk of GC, as well as whether their impact is positive or negative. This capacity for determining the influence of specific variables enhances the interpretability of the prediction model and its practical utility. Moreover, this approach facilitates precise prevention strategies tailored to each patient's risk profile.

By enhancing GC risk awareness and promoting appropriate gastroscopic screening, these prediction models are

poised to support primary and secondary prevention efforts. Moreover, the integration of the SHAP algorithm increases the model's transparency, thereby improving its clinical applicability and aiding decision-making in clinical practice. The use of readily available data from the national screening program, collected every 1-2 years, further underscores the model's clinical relevance. This prediction model holds promise for reducing GC incidence and mortality by enabling the effective identification and management of high-risk individuals.

Electronic Supplementary Material

Supplementary materials are available at Cancer Research and Treatment website (<https://www.e-crt.org>).

Ethical Statement

The Institutional Review Board of the National Cancer Center approved this study (Ncc2021-0141). Informed consent was waived because the data analyses were performed retrospectively using anonymized data.

Author Contributions

Conceived and designed the analysis: Park B, Jun JK, Suh M, Choi KS, Choi IJ, Oh HJ.

Collected the data: Park B, Oh HJ.

Contributed data or analysis tools: Park B, Kim CH, Oh HJ.

Performed the analysis: Park B, Kim CH, Oh HJ.

Wrote the paper: Park B.

Funding acquisition: Oh HJ.

ORCID iDs

Bomi Park  : <https://orcid.org/0000-0001-5834-9975>

Hyun Jin Oh  : <https://orcid.org/0000-0002-5779-6340>

Conflicts of Interest

Conflict of interest relevant to this article was not reported.

Funding

This work was supported by the National Cancer Center Grant (2111060).

References

- Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2021;71:209-49.
- Kang MJ, Jung KW, Bang SH, Choi SH, Park EH, Yun EH, et al. Cancer statistics in Korea: incidence, mortality, survival, and prevalence in 2020. *Cancer Res Treat*. 2023;55:385-99.
- National Cancer Institute. Stomach cancer survival rates and prognosis [Internet]. National Cancer Institute; 2023 [cited 2024 Aug 10]. Available from: <https://www.cancer.gov/types/stomach/survival>
- Correa P. Gastric cancer: overview. *Gastroenterol Clin North Am*. 2013;42:211-7.
- Miyamoto A, Kuriyama S, Nishino Y, Tsubono Y, Nakaya N, Ohmori K, et al. Lower risk of death from gastric cancer among participants of gastric cancer screening in Japan: a population-based cohort study. *Prev Med*. 2007;44:12-9.
- Lee KJ, Inoue M, Otani T, Iwasaki M, Sasazuki S, Tsugane S, et al. Gastric cancer screening and subsequent risk of gastric cancer: a large-scale population-based cohort study, with a 13-year follow-up in Japan. *Int J Cancer*. 2006;118:2315-21.
- Choi KS, Jun JK, Suh M, Park B, Noh DK, Song SH, et al. Effect of endoscopy screening on stage at gastric cancer diagnosis: results of the National Cancer Screening Programme in Korea. *Br J Cancer*. 2015;112:608-12.
- Jun JK, Choi KS, Lee HY, Suh M, Park B, Song SH, et al. Effectiveness of the Korean National Cancer Screening Program in reducing gastric cancer mortality. *Gastroenterology*. 2017;152:1319-28.
- Neumann H, Meier PN. Complications in gastrointestinal endoscopy. *Dig Endosc*. 2016;28:534-6.
- Eom BW, Joo J, Kim S, Shin A, Yang HR, Park J, et al. Prediction model for gastric cancer incidence in Korean population. *PLoS One*. 2015;10:e0132613.
- Cai Q, Zhu C, Yuan Y, Feng Q, Feng Y, Hao Y, et al. Development and validation of a prediction rule for estimating gastric cancer risk in the Chinese high-risk population: a nationwide multicentre study. *Gut*. 2019;68:1576-87.
- Wu M, Zhao Y, Dong X, Jin Y, Cheng S, Zhang N, et al. Artificial intelligence-based preoperative prediction system for diagnosis and prognosis in epithelial ovarian cancer: a multicenter study. *Front Oncol*. 2022;12:975703.
- Mahmoudian M, Venalainen MS, Klen R, Elo LL. Stable iterative variable selection. *Bioinformatics*. 2021;37:4810-7.
- Jiang S, Gao H, He J, Shi J, Tong Y, Wu J. Machine learning: a non-invasive prediction method for gastric cancer based on a survey of lifestyle behaviors. *Front Artif Intell*. 2022;5:956385.
- Song Z, Wu Y, Yang J, Yang D, Fang X. Progress in the treatment of advanced gastric cancer. *Tumour Biol*. 2017;39:1010428317714626.
- Taninaga J, Nishiyama Y, Fujibayashi K, Gunji T, Sasabe N, Iijima K, et al. Prediction of future gastric cancer risk using a machine learning algorithm and comprehensive medical check-up data: a case-control study. *Sci Rep*. 2019;9:12384.
- Afrash MR, Shafiee M, Kazemi-Arpanahi H. Establishing machine learning models to predict the early risk of gastric cancer based on lifestyle factors. *BMC Gastroenterol*. 2023;23:6.
- Briggs E, de Kamps M, Hamilton W, Johnson O, McInerney CD, Neal RD. Machine learning for risk prediction of

- oesophago-gastric cancer in primary care: comparison with existing risk-assessment tools. *Cancers (Basel)*. 2022;14:5023.
19. Mohammadnezhad K, Sahebi MR, Alatab S, Sadjadi A. Modeling epidemiology data with machine learning technique to detect risk factors for gastric cancer. *J Gastrointest Cancer*. 2024;55:287-96.
 20. Huang RJ, Kwon NS, Tomizawa Y, Choi AY, Hernandez-Boussard T, Hwang JH. A comparison of logistic regression against machine learning algorithms for gastric cancer risk prediction within real-world clinical data streams. *JCO Clin Cancer Inform*. 2022;6:e2200039.
 21. Liu MM, Wen L, Liu YJ, Cai Q, Li LT, Cai YM. Application of data mining methods to improve screening for the risk of early gastric cancer. *BMC Med Inform Decis Mak*. 2018;18:121.
 22. Zhu SL, Dong J, Zhang C, Huang YB, Pan W. Application of machine learning in the diagnosis of gastric cancer based on noninvasive characteristics. *PLoS One*. 2020;15:e0244869.
 23. Poorolajal J, Moradi L, Mohammadi Y, Cheraghi Z, Gohari-Ensaf F. Risk factors for stomach cancer: a systematic review and meta-analysis. *Epidemiol Health*. 2020;42:e2020004.
 24. Shin CM, Kim N, Yang HJ, Cho SI, Lee HS, Kim JS, et al. Stomach cancer risk in gastric cancer relatives: interaction between *Helicobacter pylori* infection and family history of gastric cancer for the risk of stomach cancer. *J Clin Gastroenterol*. 2010;44:e34-9.
 25. Ahn HJ, Lee DS. *Helicobacter pylori* in gastric carcinogenesis. *World J Gastrointest Oncol*. 2015;7:455-65.
 26. Luo X, Li H, He L. Correlation analysis of endoscopic manifestations and eradication effect of *Helicobacter pylori*. *Front Med (Lausanne)*. 2023;10:1259728.
 27. Ko KP. Epidemiology of gastric cancer in Korea. *J Korean Med Assoc*. 2019;62:398-406.
 28. Lim SH, Kwon JW, Kim N, Kim GH, Kang JM, Park MJ, et al. Prevalence and risk factors of *Helicobacter pylori* infection in Korea: nationwide multicenter study over 13 years. *BMC Gastroenterol*. 2013;13:104.
 29. Abdullah Alfayez A, Kunz H, Grace Lai A. Predicting the risk of cancer in adults using supervised machine learning: a scoping review. *BMJ Open*. 2021;11:e047755.