The Scientific Temper (2025) Vol. 16 (spl-1): 131-135
E-ISSN: 2231-6396, ISSN: 0976-8653
Doi: 10.58414/SCIENTIFICTEMPER.2025.16.spl-1.16
https://scientifictemper.com/

RESEARCH ARTICLE

Analysis and prediction of stomach cancer using machine learning

Vaishali Yeole[1*], Rushikesh Yeole[2], Pradheep Manisekaran[3]

Abstract

Cancer prediction and analysis systems offer aid in the management of patients and have been found to provide accurate forecasts for stage and survival prediction. This study presents a cancer prediction system developed using machine learning models and implemented with Streamlit. This system is capable of accurately predicting cancer stage onset *al*ong with chances of the patient's onset of survival based on prior patient information. For predictive purposes, categories such as random forest and XGBoost were employed. The model achieved an effective accuracy of 85% for stage prediction and 97% for predictability of patients' survival. This application includes a simple interface that healthcare professionals can employ to enter patient data and immediately make educated predictions. This paper illustrates the assistance these integrated systems provide clinicians and how they can ameliorate functional healthcare practices. In the future we are hopeful and aim towards further increasing the strength and efficiency of the system by enhancing the dataset used and additional predictive models.

Keywords: Stomach cancer, Prediction system, Cancer, Analysis, Stage prediction, Survival prediction.

Introduction

Gastric cancer is a leading contributor to cancer incidence and mortality globally, posing significant challenges to early diagnosis and effective treatment. Despite advancements in medical science, managing gastric cancer remains complex due to its heterogeneity and late-stage diagnosis in many cases. Recently, artificial intelligence (AI) approaches, particularly machine learning (ML) and deep learning, have demonstrated transformative potential for gastric cancer.

Stomach cancer is one of the most common malignancies in the world. It also poses as one of the most challenging types of cancer to clearly understand as well as effectively diagnose and treat. Although medical science has come a long way in cancer treatment, gastric cancer is still a troublesome and multidimensional problem, especially in cases where the cancer is diagnosed at later stages. Recently, AI methods, in particular machine learning and deep learning, have shown astonishing promise in improving the full spectrum of health care for patients who have gastric cancer.

The cancer prediction system developed by the team employs machine learning techniques to make stage and survival predictions concerning cancer patients using modern tools. The system has been built with an intuitive user interface using Streamlit, enabling input of patient data and receiving predictions instantly. The stage prediction and survival prediction techniques that were utilized in the system include random forest, logistic regression, XGBoost and decision tree classifier. The highest prediction accuracies achieved were 85 and 97%, respectively.

This work also sheds light on how AI-based techniques can refine computer-aided early-stage detection and treatment of gastric cancer.

Literature Survey

Afrash, Shafiee, and Kazemi-Arpanahi (2023) study using machine learning algorithms to predict the early risk of gastric cancer based on lifestyle factors. The study shows that ML algorithms are able to analyze sophisticated

[1]Research Scholar, Department of CSE, Chhatrapati Shivaji Maharaj University, Navi Mumbai, Maharashtra, India.

[2]Department of Information Technology, VES Institute of Technology (VESIT), Navi Mumbai, Maharashtra, India.

[3]Mentor, Dept of CSE, Chhatrapati Shivaji Maharaj University, Panvel, Maharashtra, India.

***Corresponding Author:** Vaishali Yeole, Research Scholar, Department of CSE, Chhatrapati Shivaji Maharaj University, Navi Mumbai, Maharashtra, India, E-Mail: yeole.vaishali@gmail.com

How to cite this article: Yeole, V., Yeole, R., Manisekaran, P. (2025). Analysis and prediction of stomach cancer using machine learning. The Scientific Temper, **16**(spl-1):131-135.

Doi: 10.58414/SCIENTIFICTEMPER.2025.16.spl-1.16

Source of support: Nil

Conflict of interest: None.

© The Scientific Temper. 2025

Received: 17/04/2025 **Accepted:** 15/05/2025 **Published** : 21/05/2025

multidimensional datasets to recognize patterns that exist within them more than traditional statistical approaches. There are two main issues that the study addresses, which are class imbalance and relevant risk factor selection for stomach cancer.

Taninaga *et al*. (2019) claim to develop a machine learning algorithm that predicts the future risk of gastric cancer using extensive medical examination records. This case-control study highlights the benefits of ML based predictions compared to conventional risk evaluation techniques. The research shows how effective ML can be applied to large medical datasets and presents an optimistic solution for the early detection of cancer.

Du *et al*. (2024) studied explainable ML models for early diagnosis of stomach cancer. Their work recognizes the role of machine learning in the AI medical domain and the need for clinicians to comprehend and have confidence in machine learning predictions. The research provides different feature selection methods and machine learning techniques that improve the accuracy of the model while keeping the model explainable.

Jiang and others (2022) created a non-invasive method of predicting stomach cancer that uses machine learning to predict lifestyle factors. The research analyzes the performance of four ML models: XGBoost, decision trees, random forests, and logistic regression. Among them, XGBoost achieves the best performance with an AUC of 89.6%, accuracy of 85.7%, sensitivity of 78.7%, and specificity of 76.9%. The study highlights the role of Helicobacter pylori infection, serum pepsinogen levels, smoking, drinking, food choices, and family history in developing gastric cancer.

The work by Cao *et al*. (2022) examines the use of AI in the diagnosis and treatment of gastric cancer. The study mentions the impact of ML in early diagnostic detection, especially in developing areas where tools such as endoscopy may not be available.

Fan *et al*. (2023) reviewed risk factors and the models of gastric cancer prediction based on ML. The study reviews the application of AI in the diagnosis and even staging of gastric cancer using deep learning models and convolutional neural networks (CNNs). Also, it discusses the use of AI-integrated endoscopic systems and radionics for early diagnosis as new emerging technologies.

Sung *et al*. (2021) have collected global cancer data, which explains the epidemiology of gastric cancer disease and its death prevalence. This study highlights the problem of the lack of accurate diagnostic instruments that enable early detection of the disease and, hence, higher survival rates.

Pimentel-Nunes *et al*. (2015) examined the submucosal endoscopic dissection guidelines by the European Society of Gastrointestinal Endoscopy (ESGE). The study reviews diagnostic and therapeutic approaches to early gastric cancer through minimally invasive procedures.

Chen and Guestrin (2016), in reference to a surgical area, incorporate XGBoost, an efficient and accurate scalable system for tree-boosting that handles complex large datasets exceptionally well. This AI technique serves as a predictive models of gastric cancer.

Bornschein *et al*. (2012), in reference, analyzed the serologic evaluation of gastric mucosal atrophy in regard to its candidacy as a serum biomarker for gastric cancer. The study supports the argument for non-invasive procedures for diagnostic purposes.

Yoshihara *et al*. (2007), in reference, examined the utility of serum pepsinogen level as a screening method to decrease mortality from gastric cancer. Their argument makes relevance to the concept of integrating markers or biomarkers with ML-based systems for cancer prediction models.

Harada *et al*. (2020) in reference discuss the recent progress in the management of esophageal cancer, which has some parallels to the management of gastric cancer. The research underscores the applications of AI in enhancing treatment results.

The Global Burden of Disease Cancer Collaboration (2019), in reference, analyses all the cancer important statistics, such as case number, death case, and corresponding disability life year, which strengthens the argument for increased investment resources into the algorithm-based identification and treatment of cancer tumors.

Miki (2011) introduces the "ABC method" for screening gastric cancer, which uniquely incorporates serum anti-*Helicobacter pylori* IgG antibody and serum pepsinogen levels. The study reinforces the need for combining this methodology with ML models to improve early detection in these methods.

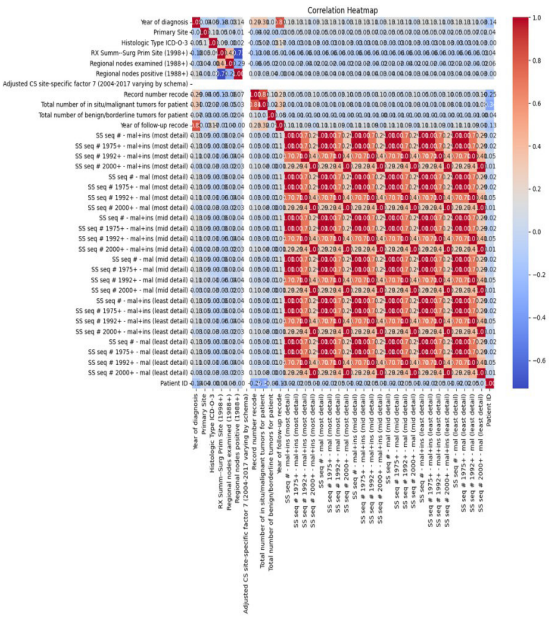
Zhu *et al*. (2020) created an ML model for diagnosing gastric cancer with the incorporation of external non-invasive features. The research proves the ability of ML to lower expenditures and increase the availability of services to the public, especially those in underdeveloped rural areas where healthcare access is scarce.

Methodology

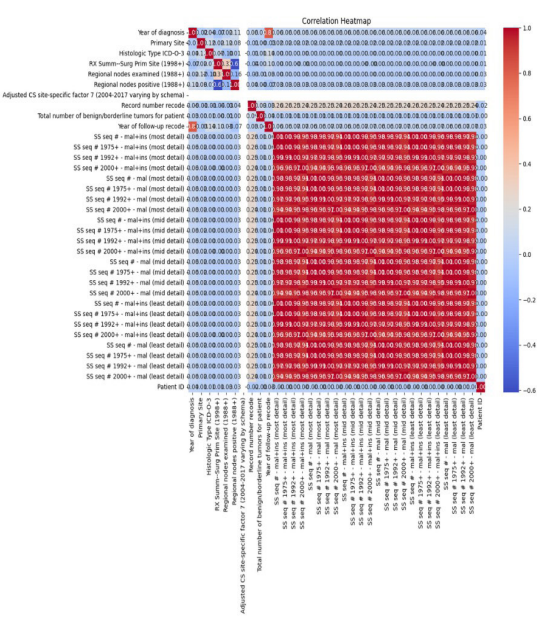
The methodology for this cancer prediction system is designed to harness machine learning (ML) techniques to predict cancer stages and survival outcomes using clinically relevant and non-invasive patient data. Models such as random forest, logistic regression, XGBoost, SVM, and linear regression were employed for their robustness and interpretability, tailored to the specific needs of stage and survival prediction.

Stage Prediction

The objective of this module is to predict the cancer stage using insights derived from the correlation heatmap (Figure 1). The heatmap revealed significant relationships among



Stage Prediction Heatmap



Survival Prediction Heatmap

Figure 1: Stage and survival correlation heatmap

demographic, diagnostic, and tumor-related attributes, which guided the selection of features for stage prediction. Selected attributes include age, sex, year of diagnosis, race, tumor size, tumor extension, lymph node involvement, and metastasis status. These attributes were chosen because they are crucial indicators of cancer progression and staging as defined by clinical guidelines such as the TNM classification. For example, tumor size and extension directly correlate with the extent of local cancer spread, while lymph node involvement and metastasis status signify regional and distant disease spread. Demographic factors like age and sex were included due to their impact on cancer biology and patient prognosis. Linear regression, random forest, and XGBoost were implemented, out of which XGBoost was utilized with an accuracy of 85% for its capability to handle a mix of numerical and categorical variables, robustness against overfitting, and ability to capture non-linear relationships within the data. User inputs are processed through a dynamic form, ensuring seamless alignment with the training data. The final output is presented as the predicted cancer stage, providing critical information for clinical decision-making and treatment planning.

Survival Prediction

Based on the insights from the correlation heatmap (Figure 2), the survival prediction module is designed to estimate a patient's survival status. The selected features include tumor-related characteristic clinical staging variables (e.g., regional nodes examined and positive) and time-related factors (e., year of follow-up). These attributes were chosen

Months from diagnosis to treatment	Delays in treatment can affect the prognosis and outcomes of gastric cancer patients.
TNM 7/CS v0204+ Schema recode	Reclassification to ensure consistency with updated staging systems and protocols.
EOD Schema ID Recode (2010+)	Standardization of staging schema ensures accurate comparison and treatment planning.
Site recode - rare tumors	Specific site and type of gastric tumor can influence treatment and prognosis.
Derived EOD 2018 Stage Group (2018+)	Grouping based on the extent of disease aids in prognosis and treatment decisions.
RX Summ--Scope Reg LN Sur (2003+)	Scope of lymph node surgery can impact survival rates and recurrence of gastric cancer.
EOD Primary Tumor (2018+)	Detailed characterization of the primary tumor for precise staging and treatment.
EOD Regional Nodes (2018+)	Detailed information on regional lymph node involvement, essential for staging.
EOD Mets (2018+)	Detailed information on distant metastasis, critical for staging and treatment planning.
Tumor Size Summary (2016+)	Summarized tumor size, an important factor in staging and prognosis.
Regional nodes positive (1988+)	Number of positive lymph nodes affects staging and prognosis.
Chemotherapy recode (yes, no/unk)	Indicates whether chemotherapy was administered, influencing treatment response and outcomes.
Radiation recode	Indicates whether radiation therapy was administered, affecting treatment response.
RX Summ--Systemic/Sur Seq (2007+)	Sequencing of systemic treatments and surgery influences treatment response and outcomes.
Histologic Type ICD-O-3	Histological type of tumor can guide treatment decisions and impact prognosis.
Grade Recode (thru 2017)	Tumor grade indicates the aggressiveness and likely behavior of the cancer.
Summary stage 2000 (1998-2017)	Overall stage summarizing the extent of cancer spread, critical for treatment planning.
Reason no cancer-directed surgery	Reasons for not performing surgery can impact overall treatment strategy and outcomes.
Regional nodes examined (1988+)	Number of lymph nodes examined during surgery, influencing staging and prognosis.
SEER Combined Mets at DX-lung (2010+)	Information on metastatic spread to lungs, critical for staging and treatment.
Survival months	Duration of survival from diagnosis, an important outcome measure.

Months from diagnosis to treatment	Delays in treatment can affect the prognosis and outcomes of gastric cancer patients.
TNM 7/CS v0204+ Schema recode	Reclassification to ensure consistency with updated staging systems and protocols.
EOD Schema ID Recode (2010+)	Standardization of staging schema ensures accurate comparison and treatment planning.
Site recode - rare tumors	Specific site and type of gastric tumor can influence treatment and prognosis.
Derived EOD 2018 Stage Group (2018+)	Grouping based on the extent of disease aids in prognosis and treatment decisions.
RX Summ--Scope Reg LN Sur (2003+)	Scope of lymph node surgery can impact survival rates and recurrence of gastric cancer.
EOD Primary Tumor (2018+)	Detailed characterization of the primary tumor for precise staging and treatment.
EOD Regional Nodes (2018+)	Detailed information on regional lymph node involvement, essential for staging.
EOD Mets (2018+)	Detailed information on distant metastasis, critical for staging and treatment planning.
Tumor Size Summary (2016+)	Summarized tumor size, an important factor in staging and prognosis.
Regional nodes positive (1988+)	Number of positive lymph nodes affects staging and prognosis.
Chemotherapy recode (yes, no/unk)	Indicates whether chemotherapy was administered, influencing treatment response and outcomes.
Radiation recode	Indicates whether radiation therapy was administered, affecting treatment response.
RX Summ--Systemic/Sur Seq (2007+)	Sequencing of systemic treatments and surgery influences treatment response and outcomes.
Histologic Type ICD-O-3	Histological type of tumor can guide treatment decisions and impact prognosis.
Grade Recode (thru 2017)	Tumor grade indicates the aggressiveness and likely behavior of the cancer.
Summary stage 2000 (1998-2017)	Overall stage summarizing the extent of cancer spread, critical for treatment planning.
Reason no cancer-directed surgery	Reasons for not performing surgery can impact overall treatment strategy and outcomes.
Regional nodes examined (1988+)	Number of lymph nodes examined during surgery, influencing staging and prognosis.
SEER Combined Mets at DX-lung (2010+)	Information on metastatic spread to lungs, critical for staging and treatment.
Survival months	Duration of survival from diagnosis, an important outcome measure.

Figure 2: Selection of features

as they exhibit strong or moderate correlations with survival outcomes in the heatmap, reinforcing their significance in determining patient prognosis. Tumor size and histologic type are critical as they directly impact cancer progression and aggressiveness. Regional node involvement (examined and positive) is justified due to its strong influence on staging and survival probabilities.

Time-related factors, such as the year of follow-up, provide insights into how temporal delays or advancements affect treatment effectiveness and patient outcomes. Logistic regression, random forest, and SVM were tested, out of which random forest, with an accuracy of 97%, was employed for its efficiency in handling binary classification tasks, particularly for modeling survival probabilities.

Discussion

With reference to Figure 2, in this study, a set of critical features were identified and selected to predict the stage and survival outcomes in gastric cancer patients. These attributes were chosen based on their clinical relevance, data availability, and potential impact on patient prognosis. To determine the most critical features, a systematic approach was employed, including a thorough review of existing literature to identify known prognostic and diagnostic factors, followed by exploratory data analysis to assess correlations and distributions. Advanced feature selection techniques, such as univariate analysis and machine learning-based methods, were utilized to evaluate the importance of each attribute. Key factors such as demographic details (age, sex, race), tumor characteristics (size, spread, lymph node involvement, metastasis), and diagnostic staging systems (TNM, EOD) were prioritized for their strong association with staging and survival outcomes.

Selection of Methods

In selecting methods for this study, we employed tailored predictive modeling techniques to achieve high accuracy in predicting stage and survival outcomes in gastric cancer patients. For stage prediction, we utilized XGBoost due to its effectiveness in modeling continuous outcomes, achieving an accuracy of 85%. Random forest was chosen for survival prediction, as it is well-suited for binary classification tasks, yielding an accuracy of 97%. The selection of these methods was guided by the nature of the prediction tasks, the data structure, and the goal of maximizing predictive performance. These results highlight the robustness of the selected models in addressing the specific requirements of stage and survival prediction tasks.

Collection of Data

This study's data was collected through the SEERStat application based on resources provided by the surveillance, epidemiology and end results program. This database is known to have valuable cancer registry data with a full range of demographics, tumor details, and survival metrics. Data was extracted with specific filters relevant to gastric cancer, focusing on the most crucial stage and survival predictors.

Criteria for evaluating the model

The evaluation of the predictive models in this study was conducted using well-established performance metrics to ensure reliability and validity. Accuracy was used as the primary metric to measure the proportion of correct predictions for both stage and survival models, providing an overall sense of model performance. Additionally, for the random forest model used in survival prediction, sensitivity (true positive rate) and specificity (true negative rate) were assessed to evaluate the model's ability to correctly identify positive and negative survival outcomes. Cross-validation techniques were employed to validate the robustness of both models and to prevent overfitting.

Conclusion

To sum up, this research effectively showcases machine learning methodologies that can improve prediction in gastric cancer diagnosis and prognosis. The system was trained on stems from the SEER database, which allowed for accurate estimation of cancer stage and survival rates with accuracy of 85 and 97%, respectively. The models included XGBoost and random forest, which performed best at these specified tasks. It is stressed in this paper that active consideration of data can facilitate clinical decision making and improve patients' health. Further work can deepen the system by including additional datasets and more sophisticated machine learning methods.

Future Scope

The proposed cancer prediction system demonstrates significant potential for improving clinical decision-making and personalized treatment strategies. The following points outline key areas for future development and enhancement:

- Integration with Real-Time Clinical Data
- Development of Multi-Class Models: Generalize the framework to include other cancers, enabling a comprehensive cancer prediction and treatment system.
- Expansion to Other Cancer Types Generalizes the framework to include other cancers, enabling a comprehensive cancer prediction and treatment system.

Acknowledgment

This research project," Analysis and Prediction of Stomach Cancer using Machine Learning," was developed to provide accurate forecasts for stage prediction as well as survival prediction. Throughout this endeavor, our team has made significant progress in combining machine learning techniques with patient feedback, gaining both theoretical insights and practical experience in applying these technologies to healthcare. The achievement of our

project's goals is a direct result of the dedication and effort of each team member. We sincerely appreciate the guidance and unwavering support of our advisor, Prof. Pradheep Manisekaran. His expertise was instrumental in shaping our approach and ensuring the successful execution of this research.

Conflicts of Interest

The authors declare that they have no conflicts of interest related to this research.

References

Afrash, M. R., Shafiee, M., & Kazemi-Arpanahi, H. (2023). Establishing machine learning models to predict the early risk of gastric cancer based on lifestyle factors. *BMC Gastroenterology, 23*(6). https://doi.org/10.1186/s12876-022-02626-x

Bornschein, J., Selgrad, M., Wex, T., Kuester, D., & Malfertheiner, P. (2012). Serological assessment of gastric mucosal atrophy in gastric cancer. BMC Gastroenterology, 12, 10. https://doi.org/10.1186/1471-230X-12-10

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16), 785–794. https://doi.org/10.1145/2939672.2939785

Cao, R., Tang, L., Fang, M., Zhong, L., Wang, S., Gong, L., Li, J., Dong, D., & Tian, J. (2022). Artificial intelligence in gastric cancer: Applications and challenges. Gastroenterology Reports (Oxford), 10, goac064. https://doi.org/10.1093/gastro/goac064

Du, H., Yang, Q., Ge, A., & *et al*. (2024). Explainable machine learning models for early gastric cancer diagnosis. Scientific Reports, 14, 17457. https://doi.org/10.1038/s41598-024-67892-z

Fan, Z., He, Z., Miao, W., & Huang, R. (2023). Critical analysis of risk factors and machine-learning-based gastric cancer risk prediction models: A systematic review. Processes, 11(2324). https://doi.org/10.3390/pr11082324

Global Burden of Disease Cancer Collaboration, Fitzmaurice, C., Abate, D., Abbasi, N., Abbastabar, H., Abd-Allah, F., Abdel-Rahman, O., Abdelalim, A., Abdoli, A., Abdollahpour, I., Abdulle, A. S. M., Abebe, N. D., Abraha, H. N., Abu-Raddad, L. J., Abualhasan, A., Adedeji, I. A., Advani, S. M., Afarideh, M., Afshari, M., ... Murray, C. J. L. (2019). Global, regional, and national cancer incidence, mortality, years of life lost, years lived with disability, and disability-adjusted life-years for 29 cancer groups, 1990 to 2017: A systematic analysis for the Global Burden of Disease Study. JAMA Oncology, 5(12), 1749–1768. https://doi.org/10.1001/jamaoncol.2019.2996

Harada, K., Rogers, J. E., Iwatsuki, M., Yamashita, K., Baba, H., & Ajani, J. A. (2020). Recent advances in treating oesophageal cancer. F1000Research, 9, F1000 Faculty Rev-1189. https://doi.org/10.12688/f1000research.22926.1

Jiang, S., Gao, H., He, J., Shi, J., Tong, Y., & Wu, J. (2022). Machine learning: A non-invasive prediction method for gastric cancer based on a survey of lifestyle behaviors. *Frontiers in Artificial Intelligence, 5*, 956385. https://doi.org/10.3389/frai.2022.956385

Miki, K. (2011). Gastric cancer screening by combined assay for serum anti-Helicobacter pylori IgG antibody and serum pepsinogen levels – "ABC method". Proceedings of the Japan Academy, Series B: Physical and Biological Sciences, 87(7), 405–414. https://doi.org/10.2183/pjab.87.405

Pimentel-Nunes, P., Dinis-Ribeiro, M., Ponchon, T., Repici, A., Vieth, M., De Ceglie, A., Amato, A., Berr, F., Bhandari, P., Bialek, A., Conio, M., Haringsma, J., Langner, C., Meisner, S., Messmann, H., Morino, M., Neuhaus, H., Piessevaux, H., Rugge, M., ... Deprez, P. H. (2015). Endoscopic submucosal dissection: European Society of Gastrointestinal Endoscopy (ESGE) guideline. Endoscopy, 47(9), 829–854. https://doi.org/10.1055/s-0034-1392882

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians, 71(3), 209–249. https://doi.org/10.3322/caac.21660

Taninaga, J., Nishiyama, Y., Fujibayashi, K., & *et al*. (2019). Prediction of future gastric cancer risk using a machine learning algorithm and comprehensive medical check-up data: A case-control study. Scientific Reports, 9. https://doi.org/10.1038/s41598-019-48769-y

Yoshihara, M., Hiyama, T., Yoshida, S., Ito, M., Tanaka, S., Watanabe, Y., & Haruma, K. (2007). Reduction in gastric cancer mortality by screening based on serum pepsinogen concentration: A case-control study. Scandinavian Journal of Gastroenterology, 42(6), 760–764. https://doi.org/10.1080/00365520601097351

Zhu, S. L., Dong, J., Zhang, C., Huang, Y. B., & Pan, W. (2020). Application of machine learning in the diagnosis of gastric cancer based on non-invasive characteristics. PLoS ONE, 15(12), e0244869. https://doi.org/10.1371/journal.pone.0244869