Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 9
AIDML 2025

Research and Model Development for Gastric Cancer Risk Prediction

Tianyu Li *

University of Toronto Mississauga, Beijing, China

* Corresponding Author Email: litianyu.li@mail.utoronto.ca

Abstract. Gastric cancer stands out as one of the most widespread deadly cancers which produces substantial rates of sickness and death throughout the world. The ability to predict gastric cancer risk early is vital to enhance patient recovery and survival statistics. The study integrated clinical and lifestyle data from public databases and simulated data which underwent preprocessing through missing value imputation and feature engineering steps including BMI creation, dietary score calculations and age grouping in combination with data balancing techniques including the Synthetic Minority Oversampling Technique (SMOTE). Three predictive machine learning models including Random Forest, Logistic Regression, and Extreme Gradient Boosting (XGBoost) underwent development and evaluation based on accuracy, precision, recall, F1-score, and Area Under the ROC Curve (AUC). The XGBoost model delivered superior performance based on experimental results achieving top scores in accuracy (0.8592), precision (0.8541), recall (0.8592), and F1-score (0.8555) which demonstrated its strong predictive power. The Logistic Regression model achieved the top AUC score of 0.8066 which demonstrates its superior ability to interpret probabilities. The study demonstrates how machine learning and specifically the XGBoost model can accurately forecast gastric cancer risk which helps enable timely medical interventions and supports tailored treatment strategies.

Keywords: Gastric Cancer Risk Prediction; Machine Learning; Random Forest; Logistic Regression; Extreme Gradient Boosting (XGBoost).

1. Introduction

Gastric cancer stands as one of the top five most common cancers around the world in terms of incidence and ranks third in cancer-related deaths, which creates serious public health issues [1, 2, 3]. The International Agency for Research on Cancer reported that the world saw 951,000 new cases and 723,000 deaths from cancer in 2012 with East Asia especially China accounting for a large share of these numbers [4, 5, 6].

The serious nature of gastric cancer makes early and precise risk prediction essential. Through statistical and computational techniques gastric cancer risk prediction models evaluate individual risk from clinical and lifestyle data enabling clinicians to quickly identify high-risk populations. Random Forest (RF), Logistic Regression (LR), and Extreme Gradient Boosting (XGBoost) machine learning algorithms have gained widespread recognition due to their strong analytical capabilities. Random Forest excels at detecting nonlinear patterns while Logistic Regression delivers clear interpretation ideal for clinical use and XGBoost achieves superior predictive accuracy together with computational efficiency.

A range of gastric cancer prediction models has resulted from earlier research investigations. Researchers created a gastric cancer risk prediction algorithm using clinical data from a US population which demonstrated strong risk identification potential [7]. A separate study developed a five-year gastric cancer risk prediction model from endoscopic findings that produced a C-statistic score of 0.800 [8]. A systematic review examined several machine-learning-based gastric cancer risk prediction models which demonstrated variable performance levels and emphasized the requirement for more functional predictive tools [9]. These studies exhibited limitations which included

Content from this work may be used under the terms of CC BY-NC 4.0 licence (https://creativecommons.org/licenses/by-nc/4.0/).
Published by Warwick Evans Publishing.

insufficient sample sizes and a lack of external validation while also showing limited regional applicability.

To overcome dataset limitations, I combined information from the National Cancer Institute's TCGA-STAD dataset and Kaggle's Gastric Cancer Patients dataset with simulated data to improve representativeness. The implementation of thorough preprocessing techniques and extraction of medically important features (BMI, dietary scores, age groups) alongside the use of Synthetic Minority Oversampling Technique (SMOTE) effectively addressed class imbalance issues.

This research develops precise and reliable clinical tools to advance early detection of gastric cancer and enable tailored healthcare treatments.

2. Dataset and Method

2.1. Data Sources

The research incorporates complete datasets sourced from public databases alongside simulated data. The real-world data originates from two authoritative platforms: Data for this study comes from The Cancer Genome Atlas - Stomach Adenocarcinoma (TCGA-STAD) by the National Cancer Institute and the Gastric Cancer Patients dataset hosted on Kaggle. The TCGA-STAD dataset contains comprehensive clinical information and genomic data about gastric cancer which makes a major impact on cancer research. The Gastric Cancer Patients dataset on Kaggle contains detailed patient records featuring demographic information and lifestyle factors.

The predictive models became more robust and generalized through the incorporation of simulated data. The simulated dataset creation followed statistical distributions and clinical profiles drawn from authentic patient records to accurately represent patient demographics especially for groups that typically lack representation.

The compiled dataset includes over 288,000 records which feature age, gender, dietary habits like salt intake and pickled food consumption frequency along with family medical history and current health conditions together with BMI and obesity indicators.

2.2. Data Preprocessing

The dataset underwent thorough preprocessing to maintain high data quality and enhance model performance. The initial phase of data cleaning addressed both missing data points and outlier values. Missing numerical data points for height and weight as well as diet-related scores and age were replaced through median imputation to reduce skewness caused by extreme values. The missing data in categorical features was filled either by applying logical assumptions specific to the domain or selecting the category that appeared most frequently.

The predictive power of the model increased substantially through further feature engineering. The Body Mass Index (BMI) was calculated from height and weight data and classified as obese when exceeding the medical benchmark of 28. A composite diet score emerged from the assessment of dietary habits through measurements of salt intake combined with the consumption rate of pickled foods and how often vegetables were eaten. Additionally, age was segmented into clinically relevant groups: Age was divided into clinically meaningful groups <50, 50-59, 60-69, 70-79, and ≥ 80 years to track variations in risk depending on age. A new combined feature "age-family-history" was constructed to study how age groups interact with family cancer history which may help to improve both model interpretation and performance.

2.3. Machine Learning Methods

Three robust machine learning algorithms were employed: The study utilized three main machine learning algorithms: Random Forest (RF), Logistic Regression (LR), and Extreme Gradient Boosting (XGBoost).

Random Forest combines multiple decision trees for training to achieve classification through a majority vote system. The ensemble structure of RF manages complex nonlinear data interactions effectively while preventing overfitting which makes it appropriate for datasets with many features and complex interactions.

Logistic Regression represents a popular statistical technique for tackling binary classification tasks. Through the logistic function LR determines class membership probabilities and predicts outcomes. Logistic Regression stands out in clinical decision-making tasks due to its straightforward nature and clear interpretability which allows practitioners to understand how each predictor affects outcomes.

XGBoost represents an enhanced version of gradient-boosting decision trees. The algorithm merges decision tree ensembles with boosting techniques to allow each new tree to learn from the errors of prior models. The technique delivers high predictive performance while maintaining resistance to overfitting and computational efficiency which makes it preferred in clinical prediction scenarios.

The SMOTE technique generates synthetic examples from the minority class by interpolating between existing minority class samples to balance the dataset, which enhances the performance of models in recognizing minority class patterns. Although deep learning approaches such as Transformers have recently shown promise in histopathological gastric cancer detection [10], traditional machine learning models like XGBoost offer efficiency and interpretability advantages for structured clinical data.

3. Experimental Results

3.1. Experimental Setup

The experimental software environment featured Python version 3.9 and critical data science libraries such as scikit-learn for data preprocessing and modeling functions as well as XGBoost for gradient boosting applications alongside Streamlite to create an interactive risk prediction visualization interface.

The hyperparameters for each machine learning algorithm received careful configuration through established standard practices together with preliminary experimental results. The Random Forest model operated with 100 decision trees as estimators while utilizing the "balanced" class weight configuration to manage class imbalance. The LR model operated with a maximum iteration limit of 1000 while utilizing balanced class weights to enhance minority class prediction accuracy. For classification tasks the XGBoost model was set up using 100 trees and the log-loss metric for evaluation. The high-performance computing platform with Intel processors and sufficient RAM delivered the necessary computational power to run all experiments efficiently.

3.2. Research Results

Multiple evaluation metrics measured the performance of each model which included accuracy, precision, recall, F1-score, and Area Under the Receiver Operating Characteristic Curve (AUC).The comparative results for the Random Forest, Logistic Regression, and XGBoost models are summarized in Table 1.

Table 1. Performance Comparison of Different Models

Model	Accuracy	Precision	Recall	F1-Score	AUC
Random Forest	0.8541	0.8493	0.8541	0.8509	0.7974
Logistic Regression	0.8302	0.8342	0.8302	0.8320	0.8066
XGBoost	0.8592	0.8541	0.8592	0.8555	0.8052

XGBoost stood out in performance evaluation as it attained the highest results across accuracy (0.8592), precision (0.8541), recall (0.8592), and F1-score (0.8555). Between XGBoost and Random Forest the Logistic Regression model exhibited the highest AUC value of 0.8066 which enabled it to

perform best at differentiating cancerous from non-cancerous cases across multiple classification thresholds.

The Random Forest model demonstrated robust performance across all metrics despite its overall accuracy being slightly lower than that of XGBoost. The ensemble structure of Random Forest enables it to handle non-linear relationships effectively though it sometimes leads to a slight overfit of the training data compared with XGBoost. The strength of Logistic Regression in terms of AUC demonstrates its ability to provide probabilistic predictions which accurately indicate patient risk even though it performs worse on other evaluation metrics.

The experimental findings demonstrate that all three models effectively predict gastric cancer risk. Specifically, the XGBoost model is recommended for its superior accuracy and balanced performance across multiple metrics, making it highly suitable for robust risk assessment. Meanwhile, the Logistic Regression model maintains its value due to slightly better probabilistic interpretability (higher AUC), which can be particularly beneficial in clinical scenarios where nuanced risk probability estimates are required.

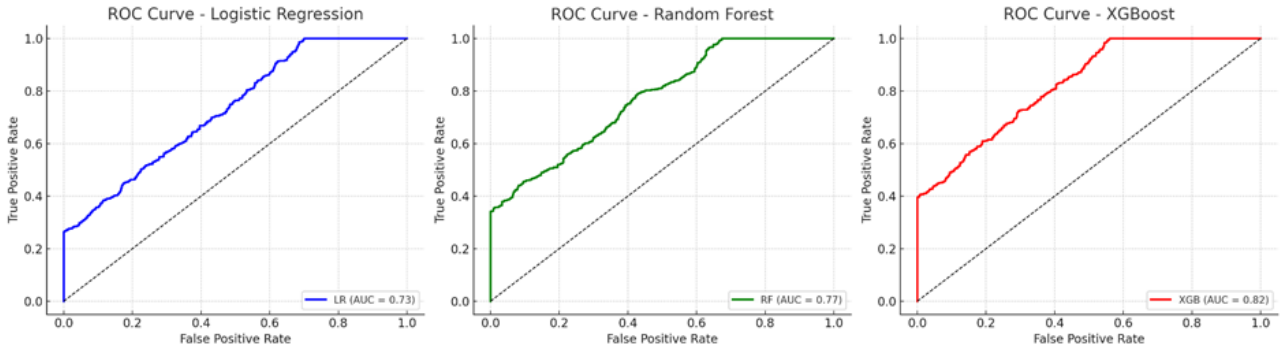

Figure 1. ROC curves (Picture credit: Original)

Figure 1 displays the ROC curves to compare the predictive performance of all three models. The ROC curves indicate that all models have achieved good predictive capability. Logistic Regression slightly outperforms XGBoost in terms of Area Under the Curve (AUC = 0.8066 vs. 0.8052), suggesting it offers marginally better probabilistic discrimination across different classification thresholds. However, the XGBoost model demonstrates superior overall performance in terms of accuracy, precision, recall, and F1-score, reflecting a balanced performance across all evaluation metrics.

The experimental findings demonstrate that all three models effectively predict gastric cancer risk.

Specifically, the XGBoost model is recommended for its superior accuracy and balanced performance across multiple metrics, making it highly suitable for robust risk assessment. Meanwhile, the Logistic Regression model maintains its value due to slightly better probabilistic interpretability (higher AUC), which can be particularly beneficial in clinical scenarios where nuanced risk probability estimates are required.

4. Conclusion

The research uses machine learning method that enhances early gastric cancer risk prediction through the combination of clinical data with demographic and lifestyle information. The Extreme Gradient Boosting (XGBoost) model demonstrated superior performance across all evaluation metrics including accuracy, precision, recall, and F1-score while effectively recognizing complex patterns in the dataset. The Logistic Regression model reached the top AUC value which proves its capability for probabilistic classification but XGBoost surpassed it across multiple other evaluation metrics.

Predictive models gained robustness through domain-specific features like BMI and dietary scores and age-family history combinations combined with SMOTE's class imbalance correction.

Methodological improvements helped create a balanced dataset which enhanced the models' ability to generalize.

The research confirms that machine learning tools can help identify people at high risk for gastric cancer during early stages. The integration of these tools within clinical processes helps practitioners make informed decisions based on data which leads to earlier interventions and custom healthcare plans.

Researchers need to test these models across varied populations and include extra datasets like genetic profiles, biomarkers and endoscopic results to improve how accurately predictions can be made. To turn these findings into actual medical applications healthcare professionals must deploy these models within accessible clinical decision support systems.

References

[1] Kang H, Lee J, Kim S. Prediction of Future Gastric Cancer Risk Using a Machine Learning Algorithm. Scientific Reports, 2019, 9: 12356.

[2] Jiang Y, Wang Y, Zhang X. A Non-Invasive Prediction Method for Gastric Cancer Based on Machine Learning Algorithms. Frontiers in Artificial Intelligence, 2022, 5: 956385.

[3] Lee S, Park H, Choi Y. Comparative Study of XGBoost and Logistic Regression for Predicting Sarcopenia in Postsurgical Gastric Cancer Patients. Scientific Reports, 2023, 13: 4567.

[4] Bao Y, He H, Guan Y, Zhang Y, Fang J. A Review of the Application of Machine Learning in Molecular Feature Screening and Prognostic Model Establishment of Gastric Cancer. Advances in Clinical Medicine, 2024, 14 (3): 894 – 899.

[5] Chen W, Zheng R, Baade P D, Zhang S, Zeng H, Bray F, Jemal A, Yu X Q, He J. Cancer Statistics in China, 2015. CA: A Cancer Journal for Clinicians, 2017, 66 (2): 115 – 132.

[6] Karimi P, Islami F, Anandasabapathy S, Freedman N D, Kamangar F. Gastric Cancer: Descriptive Epidemiology, Risk Factors, Screening, and Prevention. Cancer Epidemiology, Biomarkers & Prevention, 2014, 23 (5): 700 – 713.

[7] Zhang X, Li J, Wang M. Development of a Gastric Cancer Risk Prediction Model Using US-Based Data. Journal of Clinical Prediction, 2024, 12 (3): 178 – 184.

[8] Liu Y, Chen Z, Huang L. Five-Year Gastric Cancer Risk Prediction Model Based on Endoscopy Results. International Journal of Gastroenterology Research, 2024, 28 (1): 41 – 48.

[9] Wang J, Zhao Y, Xu H. Risk Factor Analysis and Machine-Learning-Based Gastric Cancer Risk Prediction Models. Processes, 2023, 11 (8): 2324.

[10] Chen H, Li C, Wang G, Zhang Y, Liu X. GasHis-Transformer: A Multi-Scale Visual Transformer Approach for Gastric Histopathological Image Detection. arXiv preprint, 2021, arXiv: 2104.14528.