

78896_Auto_Edited-check.docx

Name of Journal: *World Journal of Gastroenterology*

Manuscript NO: 78896

Manuscript Type: ORIGINAL ARTICLE

Retrospective Cohort Study

Machine learning-based gray-level co-occurrence matrix signature for predicting lymph node metastasis in undifferentiated-type early gastric cancer

Wei X *et al.* UEGC LNM

Abstract

BACKGROUND

The most important consideration in determining treatment strategies for undifferentiated early gastric cancer (UEGC) is the risk of lymph node metastasis (LNM). Therefore, identifying a potential biomarker that predicts LNM is quite useful in determining treatment.

AIM

To develop a machine learning (ML)-based integral procedure to construct the LNM gray-level co-occurrence matrix (GLCM) prediction model.

METHODS

We retrospectively selected 526 cases of UEGC confirmed through pathological examination after radical gastrectomy without endoscopic treatment in four tertiary hospitals between January 2015 to December 2021. We extracted GLCM-based features from grayscale images and applied ML to the classification of candidate predictive variables. The robustness and clinical utility of each model were evaluated based on the following factors: Receiver operating characteristic curve (ROC), decision curve analysis, and clinical impact curve.

RESULTS

GLCM-based feature extraction significantly correlated with LNM. The top 7 GLCM-based factors included inertia value 0° (IV_0), inertia value 45° (IV_45), inverse gap 0° (IG_0), inverse gap 45° (IG_45), inverse gap full angle (IG_all), Haralick 30° (Haralick_30), Haralick full angle (Haralick_all), and Entropy. The areas under the ROC curve (AUCs) of the random forest classifier (RFC) model, support vector machine, eXtreme gradient boosting, artificial neural network, and decision tree ranged from 0.805 [95% confidence interval (CI): 0.258-1.352] to 0.925 (95%CI: 0.378-1.472) in the training set and from 0.794 (95%CI: 0.237-1.351) to 0.912 (95%CI: 0.355-1.469) in the

testing set, respectively. The RFC (training set: AUC: 0.925, 95%CI: 0.378-1.472; testing set: AUC: 0.912, 95%CI: 0.355-1.469) model that incorporates Entropy, Haralick_all, Haralick_30, IG_all, IG_45, IG_0, and IV_45 had the highest predictive accuracy.

CONCLUSION

The evaluation results indicate that the method of selecting radiological and textural features becomes more effective in the LNM discrimination against UEGC patients. Additionally, the ML-based prediction model developed using the RFC can be used to derive treatment options and identify LNM, which can hence improve clinical outcomes.

Key Words: Undifferentiated early gastric cancer; Machine learning; Lymph node metastasis; Gray-level co-occurrence matrix; Feature selection; Prediction

Wei X, Yan XJ, Guo YY, Zhang J, Wang GR, Fayyaz A, Yu J. Machine learning-based gray-level co-occurrence matrix signature for predicting lymph node metastasis in undifferentiated-type early gastric cancer. *World J Gastroenterol* 2022; In press

Core Tip: Gray-level co-occurrence matrix-based feature extraction can be a robust and promising tool to improve the efficiency in predicting lymph node metastasis of individual undifferentiated early gastric cancer patients. Additionally, machine learning adopts more optimized algorithms and more clear feature extraction. Models developed using random forest classifier have the highest predictive accuracy in terms of Entropy, Haralick_all, Haralick_30, IG_all, IG_45, IG_0, and IV_45. Further research is required to develop these models for clinical practice.

7

INTRODUCTION

Gastric cancer (GC) is one of the most common and fatal malignancies worldwide and is an important part of the global cancer burden^[1,2]. In GC, undifferentiated early GC

(UEGC) differs from differentiated-type GC in terms of clinical features and disease state, and their treatment and prognosis vary^[3]. Therefore, UEGC should be identified and diagnosed early.

The incidence of lymphatic vessel invasion and risk of lymph node metastasis (LNM) in UEGC are high in surgical specimens of GC^[4,5]. Endoscopic resection (ER), including endoscopic mucosal resection (EMR) and endoscopic submucosal dissection (ESD), has been considered a minimally invasive treatment option for early GC with negligible risk of LNM^[6,7]. Nevertheless, indication or curability evaluation has not been conducted for ESD of undifferentiated GC (e.g., poorly differentiated adenocarcinoma, signet ring cell carcinoma, or mucinous adenocarcinoma) due to the potential risk of LNM. Although ER can be used as painless treatment, the LNM incidence after non-curative ER can be as low as 5.1% and as high as 12.2%^[8-10]. Additionally, ESD is only applicable to intramucosal cancer with a tumor diameter of ≤ 20 mm and without ulcer lesions; thus, treating lesions that meet the ESD indications through surgery is unnecessary^[11,12]. That is, when resection beyond the expanded standard is considered ineffective, the potential risk of LNM cannot be ignored. Hence, additional surgical resection and lymph node dissection should be performed. Unlike differentiated early GC, ER indications of UEGC are limited. Therefore, to address this challenging problem, a precise tool that can predict LNM must be explored.

Previous studies have mainly focused on risk factors for LNM or distant metastasis of differentiated-type early GC^[13-15]. However, for UEGC, LNM has different risk factors. Thus, objective and universal evaluation indicators for evaluating its risk are lacking. In this study, we clarified the LNM risk factors of patients with UEGC who underwent surgical resection. Subsequently, we analyzed clinical-pathological factors by introducing gray-level co-occurrence matrix (GLCM) image feature extraction mining to classify LNM risk groups according to the combination of risk factors. This study aims to provide a reference for clinical diagnosis and treatment.

MATERIALS AND METHODS

Patient selection

The clinical records of 526 patients who were diagnosed with UEGC were confirmed through pathological examination after radical gastrectomy without endoscopic treatment at four tertiary hospitals. These hospitals are Shaanxi Provincial People's Hospital, Shaanxi Provincial Tumour Hospital, the First Affiliated Hospital of Xi'an Jiaotong University, and the Second Affiliated Hospital of Xi'an Jiaotong University. The clinical records were between January 2015 to December 2021 and were retrospectively reviewed. The following were the inclusion criteria: (1) Imaging examination was performed; (2) Patients have a complete set of medical data; (3) Primary lesion was resected either *via* open surgery or laparoscopic surgery and not *via* EMR or ESD; and (4) The status of infiltrating lymph nodes was assessed through routine hematoxylin-eosin staining. To minimize the confounding effect of unnecessary variables, the following were the exclusion criteria: (1) Sufficient information cannot be extracted or mismatched clinical data of patients; and (2) Patients without complete magnetic resonance imaging (MRI) plain scan or the MRI image quality being unacceptable. This study complies with the provisions of the Helsinki Declaration (revised in 2013) and was approved by the Institutional Review Committee of Shaanxi Provincial People's Hospital (2021-Y024). Figure 1 presents in detail the patient screening steps and modeling process.

Construction strategy of the GLCM

All texture parameter post-processing was conducted on Omni dynamics software (GE pharmaceuticals, Shanghai). Two radiologists who have vast experience in gastrointestinal diagnosis referred to the MRI images to sketch the lesions on the ADC map. First, they manually sketch the entire area with cancer on each layer of the map, avoiding the gas in the intestine, until the whole tumor volume was cut out. Second, the software automatically generates the texture features. In this study, the following are the selected texture parameters of the GLCM: Total frequency, energy value, entropy,

inertia value, correlation coefficient, inverse moment, cluster shadow, and cluster prominence.

Data extraction and quality assessment

For variables with missing values (often this missing value is less than 10%), the variable's mean value should be filled. If $\geq 10\%$ of the given variables are missing, this value is excluded from the variable screening of the final model. Similarly, this study adopted unit feature interpolation for the missing values that meet the interpolation requirements. That is, the missing values can be interpolated using the constant values provided or using the statistical data of each column where these missing values are located (e.g., average value, median value, or the most frequently occurring value)^[16,17].

Construction and effectiveness evaluation of the LNM model

Based on the machine learning (ML) algorithm, the commonly used iterative algorithm models are included: Random forest classifier (RFC), decision tree (DT), support vector machine (SVM), eXtreme gradient boosting (XGBoost), and artificial neural network (ANN). The RFC is an integrated method that forms a cumulative effect by integrating multiple relatively simple evaluators. Random forest is an integrated learning tool based on DT. The SVM is a type of a generalized linear classifier that categorizes data binary through supervised learning. The ANN is a nonlinear equation transformation output algorithm comprising input, hidden, and output layers. Finally, XGBoost is an additive model. In each iteration, only the sub-models in the current step are optimized. In this study, we refer to the guide proposed by Luo *et al*^[18] for the best use of prediction models in biomedical research, that is, the Delphi method, which is used to generate the list of reported items.

For the screening of candidate variables, we mainly rely on the principle of “bag repeatedly put back and extract”, sort according to variables' weight, and finally obtain the final predictor of the prediction model from the top 10 variables^[19]. For the effectiveness evaluation of the prediction model, the receiver operating characteristic

(ROC) curve is used to evaluate the accuracy of the model. Meanwhile, the decision curve analysis and clinical impact curve (CIC) were used to evaluate the model's robustness and differentiation, respectively.

Statistical analysis

The measurement and counting data in this study are expressed by interquartile spacing (25%, 75%) and percentage (%), respectively. For the comparison between groups, the continuous variables adopt the *t*-test or Mann-Whitney *U* test of independent samples (provided that it does not conform to the normal distribution). The counting data adopt the chi-square goodness-of-fit test. Values of Bonferroni corrected probability are used to compare the qualitative data^[20]. The prediction model visualization and other data analysis are performed using R software (version 4.0.4, <http://www.r-project.org/>). For the comparison between groups, *P* value < 0.05 is considered statistically significant and vice versa.

RESULTS

Comparison of baseline data between LNM and non-LNM queues

Table 1 summarizes the baseline characteristics of 526 hospitalized patients with UEGC. For internal validation, the patients were randomly divided into two sets using the caret package: Training set (*n* = 368, 70%) and validation set (*n* = 158, 30%). Regarding the LNM rate, the training and validation cohorts were 62 (16.85%) and 29 (18.35%), respectively. In addition to the previously reported clinical-related indicators (*e.g.*, tumor size, infiltration depth, vascular_invasion, and vascular tumor thrombus), significant differences exist between the LNM and non-LNM groups. We found that GLCM-based texture acquisition features also have significant statistical differences between the two groups.

Feature correlation and potential predictors

We conducted a correlation analysis on the variables with significant differences based on the statistical difference analysis of baseline data. As shown in Figure 2A, the correlation matrix (based on Pearson correlation analysis) indicates that the characteristic variables in the GLCM and LNM had a strong correlation degree ($r > 0.6$). For example, Entropy, Haralick full angle (Haralick_all), Haralick 30° (Haralick_30), Inverse gap full angle (IG_all), Inverse gap 45° (IG_45), Inverse gap 0° (IG_0), *etc.* were highly correlated with LNM. This suggests that these potential candidate variables can be used as LNM predictors and for the construction of subsequent models. Interestingly, in the subsequent models developed based on ML algorithms, we found that Entropy, Haralick_all, Haralick_30, IG_all, IG_45, IG_0, and Inertia value 45° (IV_45) occupied high weights as the top 7 GLCM-based factors (Figure 2B). Specifically, Entropy has the largest weight among these factors.

Establishment and performance evaluation of the LNM prediction model

When constructing the RFC model [training set: Areas under the ROC curve (AUC): 0.925, 95% confidence interval (CI): 0.378-1.472; testing set: AUC: 0.912, 95%CI: 0.355-1.469], we repeatedly randomly selected N samples from the original training set N to generate the new training set DT and then generate M DTs to form a random forest according to the above steps. As shown in Figure 3A and Supplementary Table 2, the smallest Gini index after splitting was selected, including that for Entropy, Haralick_all, Haralick_30, IG_all, IG_45, IG_0, and IV_45. Similarly, Haralick_30 and IG_all served as important weight at DT branches (training set: AUC: 0.856, 95%CI: 0.309-1.403; testing set: AUC: 0.813, 95%CI: 0.256-1.370) (Figure 3B). In the ANN model (Figure 4), the accuracy of the prediction model developed using the prediction variables in the GLCM can also reach 0.887 (95%CI: 0.340-1.434) and 0.837 (95%CI: 0.280-1.394) in the training and verification sets, respectively. Although this accuracy is slightly inferior to that of the RFC model, it is better than those of other prediction models (*i.e.*, DT, XGBoost, and SVM). Table 2, Supplementary Table 2, and Figure 5 summarize the predictive performance of ML-based models. In general, the prediction model constructed by

using any ML algorithm was better than the logistic regression algorithm in predicting LNM, further confirming the superiority of ML algorithm, especially the robustness of the RFC.

3

Internal validation of the optimal RFC predictive model

The prediction efficiency of the RFC model was the best in the process of precise stratification of LNM patients. To further evaluate the “stratification effect” of the RFC, results of CIC analysis indicate that high-risk LNM was accurately distinguished using the RFC model, and “cross-linking” did not occur in the stratification process. The results of this model for the validation and training sets were consistent (Supplementary Figure 1), implying that the robustness and LNM discrimination of the RFC model were satisfactory.

DISCUSSION

The standard treatment for early GC is surgery. However, recently, ER has become the standard local treatment for some patients with early GC without LNM^[21]. For a long time, it has been used to treat differentiated-type early GC limited to the mucosa, with a diameter of < 2 cm^[22,23]. Recent studies have shown that ER indications have been expanded in many studies, even including UEGC and ≤ 2 cm diameter, without ulcer or lymphatic vessel invasion^[24]. However, whether UEGC can accept the standard treatment of ER remains a subject of debate. That is, additional surgery should be performed if curability is considered questionable. Given this situation, the risk factors of LNM or distant metastasis and mortality after non-curative ER of UEGC should be investigated. Previous studies have also shown that patients with two or more risk factors (*e.g.*, ulcer, submucosal invasion, and positive vertical margin) benefit greatly from surgical resection after ER that cannot be cured by UEGC^[14,25]. However, due to the heterogeneity of clinical characteristics, risk stratification based on these predictions provides a simple prediction, which is challenging to apply in clinical practice.

The potential application of the GLCM in the prediction of LNM of UEGC has not been systematically explored thus far. In this study, GLCM-based features were extracted from underlying grayscale images collected through MRI. We developed an LNM risk prediction model for patients with UEGC using an ML-based algorithm. The following are the two important findings of our study. First, the accurate risk stratification of UEGC patients who should undergo additional surgery depends on the added value of the GLCM. Second, a new ML-based prediction model was used to identify patients and whether they have LNM. According to previous studies^[26], texture analysis can quantify the spatial differences of pixels and the subtle differences reflected in gray values, which is consistent with the conclusion of this study. To some extent, we used GLCM features to gather spatial information and reduced the overfitting effect by replacing the softmax layer with the ML-based algorithm.

In this study, we created five types of ML-based models (*i.e.*, RFC, ANN, DT, XGBoost, and SVM), which used GLCM features to predict LNM. Interestingly, there were differences in the prediction efficiency obtained by ML-based models of different algorithms. For example, the RFC model had the highest predictive accuracy, which was achieved by incorporating Entropy, Haralick_all, Haralick_30, IG_all, IG_45, IG_0, and IV_45. Meanwhile, the ANN, DT, XGBoost, and SVM exhibited an inferior performance compared with the RFC. This suggests that the accuracy of the RFC in predicting LNM is superior to that of the ML model. A previous study indicated that a random forest algorithm is more efficient in processing classification problems, which is consistent with the results of this study^[27]. Meanwhile, DT is not as good as the RFC in terms of fitting, and the low prediction ability of the ANN model indicates that an “overfitting” phenomenon may occur. In general, different ML models show consistent accuracy, indicating that the prediction performance of ML can be improved through data processing.

Our results confirm a GLCM-based LNM classification, which has an ideal predictive effect on the diagnosis and treatment of patients with UEGC. However, the following problems were inevitably encountered in this study. First, because this study involved a

retrospective analysis, the case inclusion criteria may have a certain bias on the results, which remains to be confirmed by a large sample of prospective studies in the future. Second, there were relatively few selected cases in this study, and only some parameters of the GLCM were extracted. Thus, the results of its prediction model should be verified by external data. Third, when data from multi-center and large sample studies are available in the future, it is crucial to predict the presence or absence of LNM. Additionally, the GLCM is an important imaging sequence of UEGC, and hence we will further perform other image texture analyses in subsequent research.

CONCLUSION

GLCM-based feature extraction could, in general, serve as a robust and promising tool to improve predictive efficiency for LNM in individual UEGC patients. ML adopts the algorithm of “classification and pruning” and clearer feature extraction, leading to better data fitting than the conventional prediction model. ⁴ The model constructed using the RFC had the highest predictive accuracy, with the following being the most important predictors: Entropy, Haralick_all, Haralick_30, IG_all, IG_45, IG_0, and IV_45. In the future, we are still required to validate and optimize these prediction models using datasets of various scenarios to better apply them to clinical practice.

ARTICLE HIGHLIGHTS

Research background

Gray-level co-occurrence matrix (GLCM) based feature extraction could serve as a robust and promising tool to improve the predictive efficiency for lymph node metastasis (LNM) of individual undifferentiated early gastric cancer (UEGC) patients. Additionally, machine learning (ML) adopts more optimized algorithms and more clear feature extraction. Models built using random forest classifier (RFC) have the highest predictive accuracy in Entropy, Haralick full angle (Haralick_all), Haralick 30° (Haralick_30), Inverse gap full angle (IG_all), Inverse gap 45° (IG_45), Inverse gap 0°

(IG_0), and Inertia value 45° (IV_45). Further research is needed to develop these models for clinical practice.

Research motivation

The evaluation results indicate that the method of selecting radiological and textural features becomes more effective in the discrimination of LNM from UEGC patients. In addition, an ML-based prediction model developed using RFC can be used to derive treatment options and identify LNM that can improve clinical outcomes.

Research objectives

GLCM based feature extraction significantly correlated with LNM. The top 7 GLCM based factors included Inertia value 0°, IV_45, IG_0, IG_45, IG_all, Haralick_30, Haralick_all, and Entropy. The areas under the receiver operating characteristic (ROC) curve (AUCs) of the RFC model, support vector machine (SVM), eXtreme gradient boosting (XGBoost), artificial neural network (ANN), and decision tree (DT) ranged from 0.805 [95% confidence interval (CI): 0.258-1.352] to 0.925 (95%CI: 0.378-1.472) in the training set and from 0.794 (95%CI: 0.237-1.351) to 0.912 (95%CI: 0.355-1.469) in the testing set, respectively. The RFC (training set: AUC: 0.925, 95%CI: 0.378-1.472; testing set: AUC: 0.912, 95%CI: 0.355-1.469) model incorporating Entropy, Haralick_all, Haralick_30, IG_all, IG_45, IG_0, and IV_45 had the highest predictive accuracy.

Research methods

We retrospectively selected 526 cases of UEGC confirmed by pathological examination after radical gastrectomy without endoscopic treatment in four tertiary hospitals between January 2015 to December 2021. GLCM-based features were extracted from grayscale images and ML was applied to the classification of candidate predictive variables. In order to evaluate robustness and clinical utility of each model, the following were made: ROC, decision curve analysis, and clinical impact curve.

Research results

Identifying a potential biomarker that predicts LNM is proven to be very useful in determining treatment.

Research conclusions

To develop a ML-based integral procedure to construct the LNM gray level co-occurrence matrix (GLCM) prediction model.

Research perspectives

The risk of LNM is the most important consideration in determining treatment strategies for UEGC. Therefore, identifying a potential biomarker that predicts LNM is proven to be very useful in determining treatment.

5%

SIMILARITY INDEX

PRIMARY SOURCES

- | | | |
|---|--|-----------------|
| 1 | doaj.org
Internet | 40 words — 1% |
| 2 | Zhijun Gao, Xin Wang. "Chapter 5 Spectrum Sensing Algorithm Based on Random Forest in Dynamic Fading Channel", Springer Science and Business Media LLC, 2021
Crossref | 29 words — 1% |
| 3 | www.dovepress.com
Internet | 28 words — 1% |
| 4 | www.frontiersin.org
Internet | 25 words — 1% |
| 5 | bmcgastroenterol.biomedcentral.com
Internet | 18 words — 1% |
| 6 | Amrita Naik, Damodar Reddy Edla, Ramesh Dharavath. "Prediction of Malignancy in Lung Nodules Using Combination of Deep, Fractal, and Gray-Level Co-Occurrence Matrix Features", Big Data, 2021
Crossref | 16 words — < 1% |
| 7 | tessera.spandidos-publications.com
Internet | 14 words — < 1% |
| 8 | doctorpenguin.com
Internet | |

13 words — < 1%

- 9 Yasufumi Koterazawa, Manabu Ohashi, Masaru Hayami, Rie Makuuchi, Satoshi Ida, Koshi Kumagai, Takeshi Sano, Souya Nunobe. "Minimum resection length to ensure a pathologically negative distal margin and the preservation of a larger remnant stomach in proximal gastrectomy for early upper gastric cancer", Gastric Cancer, 2022

12 words — < 1%

Crossref

EXCLUDE QUOTES ON

EXCLUDE BIBLIOGRAPHY ON

EXCLUDE SOURCES

EXCLUDE MATCHES

< 12 WORDS

< 12 WORDS