Machine learning for predicting the risk stratification of 1-5 cm gastric gastrointestinal stromal tumors based on CT

Cui Zhang
Tongde Hospital of Zhejiang Province

Jian Wang
Tongde Hospital of Zhejiang Province

Yang Yang
First Affiliated Hospital of Bengbu Medical College

Baling Dai
Tongde Hospital of Zhejiang Province

Zhihua Xu
Tongde Hospital of Zhejiang Province

Fangmei Zhu
Tongde Hospital of Zhejiang Province

Huajun Yu (✉ yuhuajunde@163.com)
Zhejiang Hospital

Research Article

Keywords: Gastrointestinal stromal tumors, Machine learning, Risk assessment, X-ray computed tomography

Posted Date: January 13th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2461211/v1

License: ☺️  📧 This work is licensed under a Creative Commons Attribution 4.0 International License.  Read Full License
Abstract

**Background:** To predict the malignancy of 1-5 cm gastric gastrointestinal stromal tumors (GISTs) in a CT risk assessment by machine learning (ML) using three models - Logistic Regression (LR), Decision Tree (DT) and Gradient Boosting Decision Tree (GBDT).

**Methods:** 309 patients with gastric GISTs enrolled were divided into three cohorts for training (n=161), as well as internal validation (n=70) and external validation (n=78). Scikit-learn software was used to build three classifiers. Sensitivity, specificity, accuracy and area under the curve (AUC) were calculated to evaluate the performance of three models. The diagnostic difference between ML models and radiologists were compared in internal validation cohort. Important features were analyzed and compared in LR and GBDT.

**Results:** GBDT achieved the largest AUC values (0.981 and 0.815) among three classifiers in training and internal validation cohorts and greatest accuracy (0.923, 0.833 and 0.844) in three cohorts. LR was found to have the largest AUC value (0.910) in external validation cohort. DT yielded the worst accuracy (0.790 and 0.727) and AUC (0.803 and 0.700) both in two validation cohorts. GBDT and LR showed more favorable performances than two radiologists. Long diameter was demonstrated to be the same and most important CT feature for GBDT and LR.

**Conclusions:** ML classifiers were considered to be promising in prediction of risk classification of gastric GISTs less than 5 cm based on CT, especially GBDT and LR due to the high accuracy and strong robustness. Long diameter was found as the most important feature for risk stratification.

**Background**

Gastrointestinal stromal tumors (GISTs) are neoplasms that arise from Cajal cells of gastrointestinal tract mesenchyme [1]. Nonetheless they were reported growing throughout the whole digestive tract, stomach was the site with highest incidence (50–60%) [2]. As malignant potential tumor, although less than 2cm, about 10~30% GISTs will develop into malignancy, and the risk of recurrence and metastasis are significantly increased [3].

National Institutes of Health (NIH) classification system has been proposed to stratify the risk assessment of GISTs. At present, modified NIH risk stratification criteria and the latest Chinese consensus guidelines (2017 Edition) from Chinese Society of Clinical Oncology (CSCO) Expert Committee on GIST divided GISTs into very low, low, intermediate and high risk groups according to tumor size, location, the mitosis index, and whether the tumor ruptures [4, 5]. Very low and low risk GISTs generally grew slowly and had a low incidence of recurrence and metastasis, while GISTs in intermediate and high risk stages had more invasive behaviors [6]. Risk classification was also conducive to the formulation of clinical treatment plans [7]. 2~5 cm GISTs with low risk can be completely resected with endoscopic technology, however, surgical operation were necessary in advanced ones in China [8]. Another report suggested periodic follow-up by endoscopic ultrasound (EUS) is recommended for GISTs less than 1 cm [9]. However, localized GISTs (greater than 1 cm) warrant resection followed by adjuvant therapy with imatinib for lesions with intermediate to high risk [9–11]. It was reported that only 2.2% of gastric GISTs with diameters less than 1 cm were to be high-risk, whereas 1~2 cm GISTs had a malignant risk rate of 10.1% [12]. Most of grading system indicated that GISTs larger than 5cm had a great tendency to be at high risk. Therefore, it is was clinically meaningful to identify the high-risk 1~5 cm gastric GISTs preoperatively.

Contrast-enhanced CT (CE-CT) scan can clearly show the anatomical structure of gastric mesenchymal tumor, and can be more conducive to showing the internal and peripheral conditions of lesions, including density, necrosis, ulcers, bleeding, blood vessels in tumor, as well as invasion of surrounding tissue, lymph node metastasis and distant metastasis [13, 14].

A variety of studies using univariate or multivariate logistic regression analysis about predicting the risk stratification of GISTs based on CT imaging have been reported [15–21]. Tumor size was found to be the independent risk factor, even the only one, for high-risk malignant GISTs [15, 20]. Besides, other features such as percentage of tumor necrosis, growth pattern, intratumoral angiogenesis, margin and enhancement pattern were also demonstrated to contribute to the discovery of high-grade GISTs on CT images [16, 18, 21].

Machine learning (ML) algorithm provides the possibility to dig valuable data about the significant and intricate connections between enormous data items. ML has been applied to disease identification, differential diagnosis and prognosis with outstanding performance and promising prospect [22–25]. To our knowledge, no research on the use of ML for GISTs risk classification based on CT scan imaging has been attempted. In this study, 309 patients’ CT images of gastric GISTs less than 5cm were collected to assess the malignancy risk using three models - Logistic Regression (LR), Decision Tree (DT) and Gradient Boosting Decision Tree (GBDT).

**Methods**
Patient selection

This retrospective study was approved by Institutional Ethics Committee of our hospital and the need for informed consent was waived. Patients from two centers with gastric GISTs were enrolled in this research from January, 2012 to September, 2022. The criteria of patient inclusion were as follows: (a) patients who had complete CT images (including unenhanced, arterial and portal venous phases images) within 15 days before surgery; (b) solitary and primary lesion; (c) lesions without neoadjuvant treatment; (d) the lesion larger than 1 cm and smaller than 5 cm in the long diameter. (e) patients who had detailed clinical data (including age, gender, clinical symptoms and tumor markers). The inclusion and exclusion of patients were shown in Fig. 1. Finally, 231 patients (109 men and 122 women; mean age, 59.47 ± 10.13 years) from our hospital and 78 patients (41 men and 37 women; mean age, 62.69 ± 10.78 years) from other hospital were included in our series. Another 78 patients served as external validation cohort. Clinical characteristics of every patient were collected including patient age, gender, symptom and tumor marker. All of GISTs from surgical resection consisted of low-grade malignancy group and high-grade malignancy group. Low-grade malignancy category consisted of GISTs with very low or low risk and high-grade malignancy group included GISTs with intermediate or high risk. NCCN Guidelines in 2022 [5] were applied to stratify risk assessment.

Ct Examination

All patients underwent abdominal CE-CT examination using 64-slice spiral CT (Siemens, Forchheim, Germany or Philips Medical Systems, Cleveland, OH, USA). The parameters of CT imaging were set as follows: 120 kV tube voltage, 150–250 mA tube current, 0.5 s tube rotation time, 64 × 0.625 mm detector collimation, 350 × 350 mm field of view, 5 mm section thickness and 1-1.5mm reconstruction interval. Subsequently, arterial phase (delay 30–40 s) and the portal venous phase (delay 60–70 s) images were obtained with 2 mL/kg of iodinated contrast medium injected intravenously at a rate of 3 ml/s.

Image Analysis

Two radiologists (reviewer 1 with 6 and reviewer 2 with 13 years' experience in abdominal imaging) reviewed CT scan images independently, and final findings were reached by consensus without knowledge of the surgical and pathological information of every patient. The determined CT imaging features included (a) the CT attenuation value (Hounsfield units, HU) in unenhancement phase (CTU), (b) arterial phase (CTA) and (c) venous phase (CTV) of the tumor, (d) degree of enhancement in arterial phase (DEAP) and (e) in portal venous phase (DEPP), (f) enhanced potentiality in arterial phase (EPA) and (g) in portal venous phase (EPV), (h) long diameter (LD), (i) short diameter (SD), (j) the ratio of long diameter to short diameter (LD/SD), (k) contour (round; oval; irregular), (l) necrosis (yes or no), (m) calcification (yes or no), (n) surface ulceration (yes or no), (o) intratumoral angiogenesis (yes or no) and (p) peripheral enlarged lymph node (LN) (yes or no). The CT attenuation value was measured by drawing the region of interest (ROI) on the tumor in the same axial image avoiding vessels, calcification, and the necrotic regions. DEAP or DEPP was obtained by subtracting CTU from CTA or CTV respectively. EPA or EPV was equal to DEAP or DEPP divided by CTU. Enlarged lymph node was considered present if the shortest axis diameter of lymph node was more than 10 mm. A part of CT features referred to our previous report [26].

Machine Learning

Scikit-learn software was used to build three classifiers -DT, GBDT and LR for our data. The detailed methods were described in the website of official documentation (https://scikit-learn.org/), which also be applied to our previous research [25]. Three datasets (training, internal validation and external validation cohort) do not have any intersection in our study. Training dataset was aimed to train model, internal validation cohort to adjust parameters and external validation cohort to evaluate the model performance. For each model, sensitivity, specificity, accuracy, and area under the curve (AUC) together with 95% confidence intervals (95% CI) were calculated to evaluate the performance of each classifier.

Grid Search Strategy For Selecting Optimal Parameters

In order to find the optimal parameters of three models, the grid search strategy in scikit-learn software was used. The detail of grid search method was described in the model selection module in the website of official documentation (https://scikit-learn.org/stable/model_selection.html#model-selection).
Logistic Regression (Lr)
LR is the most conventional approach to measure the relationship between discrete response variable and several covariates by estimating probabilities. It can be written as: \( p = 1/(1 + e^{-z}) \). \( z \) refers to logistic regression model. The response variable can take two values (0 as no and 1 as yes) according to \( p \) smaller than 0.5 or not.

The final optimal parameters of LR were set as following: \( C = 100, \) random_state = 12, penalty = ‘l1’, solver = ‘liblinear’. Other parameter factors were set as default in sklearn software module.

Decision Tree (Dt)
DT as a binary method that can classify data by calculating their characteristics. Decision nodes, branches and leaves are the three main components of DT. DT starts with a node and extends to many branches and child nodes, finally to leaves. The criterion used in our model were Gini’s Diversity Index, a measure of node impurity. The standard CART algorithm implemented using sciki-learn library in Python was applied to build decision tree.

The parameters set in the DT were: random_state = 0, max_features = 6, max_depth = 6, criterion = ‘gini’. Other parameters were set as default in sklearn software module.

Gradient Boosting Decision Tree (Gbdt)
GBDT is an ensemble classifier based on bootstrap sampling, and its purpose is to improve the generalization ability and robustness by combining the predicted results of multiple base learners (i.e. weak decision trees). The weight is adjusted with iteration, so that the higher weight will be assigned to the data poorly classified. Total 15 weak decision trees were created in GBDT model in this study (e.g. a tree was showed in Fig. S1).

The following showed the parameter factors in the GBDT: learning_rate = 0.1, max_depth = 8, random_state = 0, min_samples_leaf = 2. Other parameters were also set as default in sklearn software.

Performance Comparison Between Radiologists And Models
The diagnostic performance differences between three ML models and two radiologists were compared in internal validation cohort.

Feature Variable Analysis
GBDT and LR showed excellent diagnostic effency in the prediction of risk classification of gastric GISTs on account of the high accuracy and strong robustness. LR is famous for determining the beneficial features to support decision by linear analysis, since the result is easy to explain. Firstly, significant CT features were determined by univariate analysis. Secondly, variable with \( P \) less than 0.05 were as the input data to calculate the independent risk factors for high-risk malignant GISTs. In order to find out the top five important features for high-grade malignant GISTs, the function of Feature_Importance was performed. The description of feature importance was in the website: https://scikit-learn.org/stable/modules/ensemble.html#gradient-tree-boosting). According to the official documentation description, individual decision trees in the GBDT model intrinsically perform feature selection by selecting appropriate split points. This information can be used to measure the importance of each feature. The basic idea is: the more often a feature is used in the split points of a tree, the more important that feature is. Subsequently, the feature variables of LR and GBDT were compared.

Statistical Analysis
Continuous distributed data were showed as mean ± SD, and categorical variables were expressed as n (%). Univariate analysis using t test or Mann-Whitney U test for continuous variables and Fisher’s exact test for categorical variables were performed to compare CT features between the low-grade malignancy and high-grade malignancy groups. Variables with \( P < 0.05 \) were considered as significant features and included in the LR multivariate analysis. The final features with \( P < 0.05 \) from multivariate logistic regression model indicated the significant predictors of high risk GISTs. Statistical analyses were performed using SPSS version 22.0 (SPSS Inc., Chicago, IL, USA). A statistically significant difference was defined as two - sided \( P \) value < 0.05.
### Results

#### Clinical characteristics of patients

Table S1 showed the clinical characteristics of three cohorts. Results of the univariate analysis indicated that patients in three databases had no significant difference in variables of age, gender distribution, clinical symptom and tumor marker between the low-grade malignancy and high-grade malignancy groups (all $P > 0.05$). The results in Table 1 indicated that clinical characteristics of patients had no contribution to distinguish between low-grade and high-grade GISTs.

#### Table 1

Univariate analysis of CT features of GISTs in training cohort, internal validation cohort and external validation cohort.

<table>
<thead>
<tr>
<th>CT features</th>
<th>Training cohort</th>
<th>Internal validation cohort</th>
<th>External validation cohort</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(n = 161)</td>
<td>(n = 70)</td>
<td>(n = 78)</td>
</tr>
<tr>
<td></td>
<td>Low-grade malignancy</td>
<td>High-grade malignancy</td>
<td>$P$ value</td>
</tr>
<tr>
<td>CTU(HU)</td>
<td>33.90 ± 8.58</td>
<td>33.86 ± 6.38</td>
<td>0.976</td>
</tr>
<tr>
<td>CTA(HU)</td>
<td>53.24 ± 12.51</td>
<td>55.39 ± 15.60</td>
<td>0.359</td>
</tr>
<tr>
<td>CTV(HU)</td>
<td>67.38 ± 15.84</td>
<td>68.84 ± 15.11</td>
<td>0.592</td>
</tr>
<tr>
<td>DEAP(HU)</td>
<td>19.33 ± 11.15</td>
<td>21.53 ± 13.55</td>
<td>0.289</td>
</tr>
<tr>
<td>DEPP(HU)</td>
<td>34.47 ± 16.42</td>
<td>34.97 ± 13.55</td>
<td>0.582</td>
</tr>
<tr>
<td>EPA</td>
<td>0.65 ± 0.67</td>
<td>0.65 ± 0.39</td>
<td>0.971</td>
</tr>
<tr>
<td>Epv</td>
<td>1.14 ± 1.22</td>
<td>1.07 ± 0.44</td>
<td>0.685</td>
</tr>
<tr>
<td>LD(mm)</td>
<td>24.78 ± 10.39</td>
<td>33.81 ± 12.94</td>
<td>0.000</td>
</tr>
<tr>
<td>SD(mm)</td>
<td>20.43 ± 9.62</td>
<td>28.19 ± 10.65</td>
<td>0.000</td>
</tr>
<tr>
<td>LD/SD</td>
<td>1.24 ± 0.26</td>
<td>1.20 ± 0.16</td>
<td>0.286</td>
</tr>
<tr>
<td>Contour</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Round</td>
<td>60(52.63%)</td>
<td>14(29.79%)</td>
<td></td>
</tr>
<tr>
<td>Oval</td>
<td>38(33.33%)</td>
<td>15(31.91%)</td>
<td></td>
</tr>
<tr>
<td>Irregular</td>
<td>16(14.04%)</td>
<td>18(38.30%)</td>
<td></td>
</tr>
<tr>
<td>Necrosis</td>
<td>24(21.05%)</td>
<td>20(42.55%)</td>
<td>0.007</td>
</tr>
<tr>
<td>Calcification</td>
<td>12(10.53%)</td>
<td>4(8.51%)</td>
<td></td>
</tr>
<tr>
<td>Surface ulceration</td>
<td>10(8.77%)</td>
<td>10(21.28%)</td>
<td>0.037</td>
</tr>
<tr>
<td>BVT</td>
<td>8(7.02%)</td>
<td>11(23.40%)</td>
<td>0.006</td>
</tr>
<tr>
<td>LN</td>
<td>0</td>
<td>1(2.13%)</td>
<td>0.295</td>
</tr>
</tbody>
</table>

CTU/CTA/CTV, the CT attenuation value in unenhancement phase/arterial phase/portal venous phase; DEAP/ DEPP, degree of enhancement in arterial phase/portal venous phase; EPA/EPv, enhanced potentiality in arterial phase/portal venous phase; LD, long diameter; SD, short diameter; BVT, blood vessels in tumor; LN, peripheral enlarged lymph nodes.

$P$ values written in bold indicate significant difference between low-grade malignancy and high-grade malignancy groups.
Univariate Analysis Of CT Data

Univariate analysis of CT imaging features was exhibited in Table 1. Results showed that LD, SD, contour, presence of necrosis and surface ulceration were significant different features in distinguishing two groups in three cohorts. Size in high-grade malignancy GISTs was found larger than low-grade ones. Lesions with oval and irregular contour were seen more commonly in the high-grade malignancy group than in the low-grade group. Necrosis and surface ulceration were more likely to be found in the high-grade group. Intratumoral angiogenesis was significant different feature when compared the two groups in training and external validation cohort, but not in internal validation cohort showed in Fig. 2. The other CT imaging variables showed no significant difference in all three cohorts.

Model Evaluation

Results of diagnostic performance of LR, DT and GBDT were described in Table 2 and Fig. 3. Due to the compromise between sensitivity and specificity, accuracy and AUC were considered as better diagnostic indicators. GBDT achieved the largest AUC (0.981 and 0.815) among all three classifiers in training and internal validation cohorts. Nevertheless, we found that LR had the largest AUC (0.910), followed by GBDT (0.819) and DT (0.700) in external validation cohort. The worst accuracy (0.790 and 0.727) and AUC (0.803 and 0.700) were gained in DT model both in internal and external validation cohort. GBDT and LR performed well among three models and in three cohorts with high accuracy and strong robustness.

Table 2
Diagnostic performance analysis of LR, DT and GBDT models.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Group</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Accuracy</th>
<th>AUC (95%CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR</td>
<td>Training cohort</td>
<td>0.918</td>
<td>0.492</td>
<td>0.792</td>
<td>0.815 (0.744–0.885)</td>
</tr>
<tr>
<td></td>
<td>Internal validation cohort</td>
<td>0.941</td>
<td>0.437</td>
<td>0.792</td>
<td>0.815 (0.602–0.904)</td>
</tr>
<tr>
<td></td>
<td>External validation cohort</td>
<td>0.852</td>
<td>0.688</td>
<td>0.818</td>
<td>0.910 (0.810–0.978)</td>
</tr>
<tr>
<td>DT</td>
<td>Training cohort</td>
<td>0.966</td>
<td>0.639</td>
<td>0.870</td>
<td>0.883 (0.826–0.941)</td>
</tr>
<tr>
<td></td>
<td>Internal validation cohort</td>
<td>0.941</td>
<td>0.429</td>
<td>0.790</td>
<td>0.803 (0.587–0.845)</td>
</tr>
<tr>
<td></td>
<td>External validation cohort</td>
<td>0.787</td>
<td>0.500</td>
<td>0.727</td>
<td>0.700 (0.545–0.856)</td>
</tr>
<tr>
<td>GBDT</td>
<td>Training cohort</td>
<td>0.986</td>
<td>0.770</td>
<td>0.923</td>
<td>0.981 (0.957–1.000)</td>
</tr>
<tr>
<td></td>
<td>Internal validation cohort</td>
<td>0.882</td>
<td>0.714</td>
<td>0.833</td>
<td>0.815 (0.704–0.920)</td>
</tr>
<tr>
<td></td>
<td>External validation cohort</td>
<td>0.918</td>
<td>0.563</td>
<td>0.844</td>
<td>0.819 (0.686–0.952)</td>
</tr>
</tbody>
</table>

LR, Logistic regression; DT, Decision tree; GBDT, Gradient boosting decision tree; AUC, area under the curve; CI, confidence interval.

Performance Comparison Between Radiologists And Models

Table 3 displayed the two reviewers’ diagnostic performance in internal validation cohort. GBDT and LR showed more favorable performances than two radiologists.

Table 3
Results of radiologists’ diagnostic performance in internal validation cohort.

<table>
<thead>
<tr>
<th></th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Accuracy</th>
<th>AUC (95%CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR</td>
<td>0.653</td>
<td>0.524</td>
<td>0.614</td>
<td>0.602 (0.574–0.666)</td>
</tr>
<tr>
<td>DT</td>
<td>0.735</td>
<td>0.619</td>
<td>0.700</td>
<td>0.717 (0.682–0.832)</td>
</tr>
</tbody>
</table>

AUC, area under the curve; CI, confidence interval.

Feature Variable Analysis

LD, SD, contour, necrosis, surface ulceration and intratumoral angiogenesis were selected as significant features by univariate analysis to input into multivariate analysis. Table 4 indicated that only LD was an independent risk factor for high-grade malignant GISTs (P < 0.001,
The top five important features from GBDT were ranked in Fig. 4. LD ranked the most important feature among all CT features with importance score of 0.202, followed by SD (0.175), DEPP (0.115), CTU (0.088) and DEAP (0.064). LD was demonstrated as the only same and most important feature for LR and GBDT in terms of feature variable analysis.

Table 4
Results of feature variable analysis in LR model.

<table>
<thead>
<tr>
<th>CT features</th>
<th>β</th>
<th>OR</th>
<th>OR (95%CI)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>LD</td>
<td>0.064</td>
<td>1.066</td>
<td>1.034–1.099</td>
<td>0.000</td>
</tr>
<tr>
<td>SD</td>
<td>0.550</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contour</td>
<td>0.288</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Necrosis</td>
<td>0.658</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Surface ulceration</td>
<td>0.404</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BVT</td>
<td>0.248</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

LD, long diameter; SD, short diameter; BVT, blood vessels in tumor; OR, odds ratio; CI, confidence interval.

P values written in bold indicate significant difference.

Discussion

To our best knowledge, this is the first research on prediction of malignancy of gastric GISTs on CT images by machine learning. In addition, this study has the largest sample size among relevant studies, so that the reliability of the results can be guaranteed. Various qualitative and quantitative variables extracted from CT signs were inputted into LR, DT and GBDT models. The results of model evaluation analysis were different but not inconsistent among three cohorts. For training cohort, GBDT had the largest sensitivity, specificity, accuracy and AUC among three models. What's more, GBDT gained the largest AUC in training and internal validation set and performed well in all three cohorts in terms of accuracy, although the AUC was not as good as LR in the external validation cohort. Furthermore, GBDT and LR showed more outstanding performance than two radiologists. However, DT was not as outstanding as GBDT and LR. Therefore, GBDT and LR were suggested to be promising ML models in prediction of risk classification of gastric GISTs on CT giving consideration to the high accuracy and strong robustness.

GBDT, an ensemble method based on bootstrap sampling, was demonstrated to be the favorable algorithm with highly predictive efficiency, which has been reported in a variety of previous researches [27–30]. The excellent performance of GBDT classifier may not be surprising due to the ability of increasing the weight of weak decision trees for optimizing the algorithm. As a classic machine learning algorithm to solve two classification problems, LR assumes that the data conform to the Bernoulli distribution, and calculates the parameters through the maximum likelihood function method to achieve two classification. Additional, LR model obtains probability prediction, that can be utilized preferably to assist decision-making better. However, LR has some limitations. First, logistic regression analysis, difficult to fit the true distribution of data, has only linear decision boundary. It may weak the ability of classification task resulting from the limitation of over-fitting and multicollinearity. In terms of DT, a tree composed of decision nodes, branches, and leaves is generated using the training dataset and classify or predict the test dataset. In this study, the ability of DT in predicting the risk grading of gastric GISTs may be weaker than LR and GBDT.

When it came to feature variable selection, long diameter was found to be the only common CT feature for LR and GBDT in this study. Several studies using multivariate logistic regression analysis revealed that the size of GISTs was the only independent risk factor for differentiation the high-grade malignant GISTs [15, 20], so did the gastric GISTs with size shorter than 5 cm in our series. The study by Kim et al. [20] reported that, for the GISTs ≤ 5 cm, identifying malignant from benign by tumor size was not possible based on CT scan images, which was different from our results. It may be related to the different grouping definition and difference in composition ratio of tumors. Mazzei et al. [31] found that the maximum diameter of GISTs with high mitotic index (> 5 mitoses) was larger than those with low mitotic index (≤ 5 mitoses), which revealed that the larger the tumor was, the faster it grew and the higher the degree of malignancy it had. However, oval and irregular contour, presence of necrosis, surface ulceration and intratumoral angiogenesis were seen more frequently in the high-grade malignancy group, which were similar with other studies [16, 18, 20, 21], these features were excluded from predictive factors selection in LR. GBDT algorithm could find complicated and impalpable feature relationship between data to support decision-making that may not detected in logistic regression analysis [32]. In this study, DEPP, CTU and DEAP were selected as important features despite that they were not significant different factors in univariate analysis. The important features that GBDT concentrated on were seemingly
unimportant features in traditional statistical method, and highly predictive performance can be obtained by the contributions from these variables.

Artificially determined CT imaging features were used as the input variables of three ML models to predict the risk grading of gastric GISTs in this study, and great results were obtained, especially in GBDT and LR. Compared with the diagnostic ability of radiologists, machine learning has achieved more welcome results, which may have a guiding prospect for doctors in daily diagnostic work, especially for the junior ones. It may promisingly provide theoretical and practical support for texture analysis or deep learning since ML may play an important role in feature selection.

There are some limitations in our study. First, our sample size was small for ML. ML classifiers can highlight their advantages in the context of large data, amount of predictor variables or complex relationship. Second, four risk grades were further divided into two, so the results were unable to meet the requirement of each risk classification. Simple ML model cannot meet the needs of four risk levels prediction, but the convolutional neural network can. It puts the next step of research on the agenda. Third, only three simple ML models were implemented in our research, including the classic LR. We will attempt next other and more complex ML models to assess the risk stratification, such as random forest, support vector machine, k-nearest neighbors, etc.

Conclusions

In summary, GBDT and LR showed outstanding performance with high accuracy and strong robustness in the risk assessment of gastric GISTs less than 5 cm on CT imaging. The long diameter of lesion was found to be the most important feature for risk stratification.

Abbreviations


Declarations

Acknowledgements

Not applicable.

Author contributions

(I) Conception and design: C. Zhang, J. Wang, H.J. Yu; (II) Administrative support: J. Wang, H.J. Yu; (III) Provision of study materials or patients: All authors; (IV) Collection and assembly of data: C. Zhang, Y. Yang, B.L. Dai, Z.H. Xu, F.M. Zhu; (V) Data analysis and interpretation: C. Zhang, Y. Yang, B.L. Dai, Z.H. Xu, F.M. Zhu; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Funding

This research was supported by Zhejiang Provincial Natural Science Foundation of China (No. LGF21H030004).

Availability of data and materials

All datasets presented and analyzed in this study were interpreted and provided by the corresponding author.

Ethics approval and consent to participate

This retrospective study was approved by our institutional review board (Tongde Hospital of Zhejiang Province No. 2021-040).

Consent for publication

Not applicable.

Competing interests
References


**Figures**

601 patients with pathologically confirmed gastric gastrointestinal stromal tumors were retrospectively collected from January 2012 to September 2022

292 Excluded:
(a) The lesion smaller than 1 cm or larger than 5 cm in the long diameter (n = 263);
(b) Multiple lesions (n = 3);
(c) Integrated CT data or fuzzy CT imagings (n = 13);
(d) Preoperative CT interval more than 2 months (n = 5);
(e) Incomplete clinical data (n = 6);

Finally enrolled patients n = 309
(Very low and low risk group = 225; intermediate and high risk group = 84.)

Figure 1
Flowchart shows inclusion and exclusion criteria for the study.

**Figure 2**

CT image of low-grade and high-grade malignant GISTs. (A-B). A low-grade malignant GIST in cardia in a 62-year-old man. (A-B). Axial and sagittal CT scans in portal venous phase show a round mass with 2.5 cm of the long diameter (white arrows). The lesion has a homogeneous enhancement pattern without necrosis, calcification, and intratumoral angiogenesis in tumor. (C-D). A high-grade malignant GIST in gastric body in a 57-year-old woman. Axial and coronal CT images in portal venous phase show an irregular neoplasm with 4.7 cm of the long diameter (white arrows). Intratumoral angiogenesis (black arrows) are seen in the lesion and the mass shows heterogeneous enhancement with necrotic portion (*) within the tumor.
Receiver operating characteristic (ROC) curves of three models to predict the risk stratification of gastric GISTs. (A). ROC curve of three models in training cohort. The largest area under the curve (AUC) was GBDT (0.981), followed by DT (0.883) and LR (0.815). (B). ROC curve of three models in internal validation cohort. GBDT and LR had the equal AUC (0.815), and DT gained the smallest AUC (0.803). (C). ROC curve of three models in external validation cohort. LR gained the best AUC (0.910), followed by GBDT (0.819) and DT (0.700).
Figure 4

Five top important features in GBDT model. Five top important features were as follows: LD (importance score 0.202), SD (0.175), DEPP (0.115), CTU (0.088) and DEAP (0.064).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- FigS1.docx
- TableS1.docx