Received: 25 August 2020 | Revised: 31 May 2021 | Accepted: 30 June 2021

DOI: 10.1002/mp.15094

RESEARCH ARTICLE

MEDICAL PHYSICS

Specific Borrmann classification in advanced gastric cancer by an ensemble multilayer perceptron network: a multicenter research

Siwen Wang[1,2] | **Di Dong**[1,2] | **Wenjuan Zhang**[3] | **Hui Hu**[4] | **Hailin Li**[1,5] | **Yongbei Zhu**[1,5] | **Junlin Zhou**[3] | **Xiuhong Shan**[4] | **Jie Tian**[1,5,6,7]

[1]CAS Key Laboratory of Molecular Imaging, Beijing Key Laboratory of Molecular Imaging, The State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China

[2]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China

[3]Department of Radiology, Lanzhou University Second Hospital, Lanzhou, Gansu, China

[4]Department of Radiology, Affiliated Renmin Hospital of Jiangsu University, Zhenjiang, Jiangsu, China

[5]Beijing Advanced Innovation Center for Big Data-Based Precision Medicine, School of Medicine and Engineering, Beihang University, Beijing, China

[6]Engineering Research Center of Molecular and Neuro Imaging of Ministry of Education, School of Life Science and Technology, Xidian University, Xi'an, Shaanxi, China

[7]Key Laboratory of Big Data-Based Precision Medicine (Beihang University), Ministry of Industry and Information Technology, Beijing, China

Correspondence
Jie Tian, CAS Key Laboratory of Molecular Imaging, Institute of Automation, Chinese Academy of Sciences, Beijing, 100190, China.
Email: jie.tian@ia.ac.cn

Xiuhong Shan, Department of Radiology, Affiliated Renmin Hospital of Jiangsu University, Zhenjiang, Jiangsu, 212002, China.
Email: 13913433095@163.com

Junlin Zhou, Department of Radiology, Lanzhou University Second Hospital, Lanzhou, Gansu, 730030, China.
Email: ery_zhoujl@lzu.edu.cn

Funding information
National Key R&D Program of China, Grant/Award Number: 2017YFA0205200; National Natural Science Foundation of China, Grant/Award Number: 82022036, 91959130, 81971776, 81771924, 62027901, 81227901, 81930053 and 81527805; Beijing Natural Science Foundation, Grant/Award Number: L182061; Strategic Priority Research Program of Chinese Academy of Sciences, Grant/Award Number: XDB 38040200; Chinese Academy of Sciences, Grant/Award Number: GJJSTD20170004 and QYZDJ-SSW-JSC005; Youth Innovation Promotion Association CAS, Grant/Award Number: 2017175; Project of High-Level Talents

Abstract

Purpose: Borrmann classification in advanced gastric cancer (AGC) is necessarily associated with personalized surgical strategy and prognosis. But few radiomics research studies have focused on specific Borrmann classification, and there is yet no consensus regarding what machine learning methods should be the most effective.

Methods: A combined size of 889 AGC patients was retrospectively enrolled from two centers. Radiomic features were extracted from tumors manually delineated on preoperative computed tomography images. Two classification experiments (Borrmann I/II/III vs. IV and Borrmann II vs. III) were conducted. In each task, we combined three common feature selection methods and five typical machine learning classifiers to construct 15 basic classification models, and then fed the 15 predictions to a designed multilayer perceptron (MLP) network.

Results: In internal and external validation cohorts, the proposed ensemble MLP yielded good performance with area under curves of 0.767 and 0.702 for Borrmann I/II/III vs. IV, as well as 0.768 and 0.731 for Borrmann II vs. III. Considering the imbalanced distribution of four Borrmann types (I, 2.9%; II, 12.8%; III, 69.5%; IV, 14.7%), the ensemble MLP surpassed the overfitting barrier and attained fine specificity (0.667 and 0.750 for Borrmann I/II/III vs. IV; 0.714 and 0.620 for Borrmann II vs. III) and sensitivity (0.795 and 0.610 for Borrmann I/II/III vs. IV; 0.652 and 0.703 for Borrmann II vs. III). Also, survival analysis showed

Siwen Wang, Di Dong, Wenjuan Zhang, and Hui Hu are joint first authors.

Team Introduction in Zhuhai City, Grant/Award Number: HLHPTP201703

that patients could be significantly risk stratified by MLP predicted types in both experiments ($p < 0.0001$, log-rank test).

Conclusions: This study proposed an MLP-based ensemble learning architecture, which could identify Borrmann type IV automatically and improve the differentiation of Borrmann type II from III. The study provided a new view for specific Borrmann classification in clinical practice.

KEYWORDS
advanced gastric cancer, Borrmann classification, ensemble learning, multilayer perceptron networks, radiomics

1 | INTRODUCTION

Gastric cancer (GC) remains common around the world and takes the third leading role in causes of cancer mortality as estimated in 2018.[1,2] Accurate classification of advanced gastric cancer (AGC) is one of key factors of personalized treatment strategies. The Borrmann classification system, developed in 1926 according to the gross appearance, is still widely used by surgeons, pathologists, endoscopists, and radiologists worldwide.[3]

Research has revealed that Borrmann type IV is regarded as an independent prognostic factor for AGC and it usually has a worse prognosis than other Borrmann types.[4] Thus, it naturally makes sense to distinguish Borrmann type IV from the other types. Borrmann type I tumors have special morphological characteristics as nodular polypoid and the lesions generally invade the mucosa, submucosa, and muscularis, but rarely invade the serosa.[5] However, Borrmann type II and III tumors are both ulcerative, and it may be difficult to distinguish the tumor invasion in surrounding tissues from inflammatory edema or adipose deposition of surrounding tissues, as the spread of tumor microvascular vessels is larger than the actual invasion of tumors. But Borrmann type III AGC is recommended a larger surgical resection margin (at least 5 cm) than Borrmann type II (at least 3 cm)[6] and usually displays worse prognosis than the latter.[7] And the accuracies for assessing Borrmann type II and III through computed tomography (CT) are relatively low (79.7% and 80.0%),[5] wherein assessments by junior radiologists even fall behind senior radiologists. Thus, further differentiating Borrmann type II from III accurately is necessary for developing a reasonable surgical plan and evaluating the prognosis.

Many modalities including double contrast barium meal, endoscopy, endoscopic ultrasonography (EUS), double contrast-enhanced ultrasonography (DCEUS), and CT have been applied for evaluating Borrmann classification preoperatively.[5] Among these modalities, the accuracy of double contrast barium meal technique is low. Endoscopy and EUS are invasive to some extent, and patients with upper gastrointestinal stenosis or obstruction are not suitable for such examinations. According to a previous research that compared the accuracy of multidetector CT (MDCT) with DCEUS in Borrmann classification determination,[5] DCEUS may act as a complementary tool in preoperatively assessing the gross appearance of GC; however, it more depends on the operator's experience. Thereupon, CT with high resolution and multiplanar reformatted views is now the most commonly used imaging method in preoperative examination of GC, which plays a very important part in determining the location, size, and infiltration depth of tumors.

These years, the advent of radiomics has improved the understanding of medical images. The primary concept of radiomics lies in mining high-throughput image features quantitatively and making connections between these features with tumor heterogeneity related to clinical issues.[8–11] Machine learning models, the core of radiomics, show great potential in cancer diagnosis wherein various feature selection and classification methods have been explored. Parmar et al.[12] explored 14 feature selection methods and 12 classification methods, and found random forest (RF) surpassed all the others. Wu et al.[13] conducted a comparative study on 24 feature selection methods and 3 classification methods, and the results showed that naïve Bayes (NB) yielded highest classification performance. Similar comparative study design was also derived in several research studies.[14–16] However, a new problem emerges: which feature selection method and which classifier on earth are most suitable for certain classification tasks? Among most radiomics research studies applied so far, there is no received coherent conclusion.

To this end, it is natural to expect ensemble learning that takes full advantage of multiple classification models. Boosting algorithm,[17] is a popular ensemble strategy that discovers and utilizes complex information from different models. On the other hand, multilayer perceptron (MLP) networks have shown excellent ability in coping with highly variable predictions by fully connecting compositional layers of neurons.[18,19] These

may inspire a new train of thought to adopt the MLP architecture in boosting-based ensemble learning in this study.

As far as we know, only one CT-based radiomics study was conducted to differentiate Borrmann type IV GC from primary gastric lymphoma.[20] However, this study may not provide a comprehensive view of specific Borrmann classification in AGC. In the current study, we thus advocated an MLP-based ensemble learning architecture to identify Borrmann type IV automatically and improve the differentiation of Borrmann type II from III.

2 | MATERIALS AND METHODS

2.1 | Research dataset

The Ethical Committees of Affiliated Renmin Hospital of Jiangsu University (center 1) and Lanzhou University Second Hospital (center 2) both granted this retrospective research and informed consent was waived. Patient recruitment criteria are given in Text S1 (see Supplementary Materials). Borrmann types are defined by Borrmann system and CT image interpretation (Text S2).[21,22] Totally, a combined size of 889 AGC patients was enrolled from the two centers. A summary of 597 consecutive AGC patients in center 1 (December 2011 to December 2016) fulfilled the recruitment criteria. The four Borrmann types accounted for 2.7%, 7.2%, 74.2%, and 15.9%, respectively. A total of 292 AGC patients treated at center 2 (January 2013 to December 2015) were analyzed as an independent external cohort. The four Borrmann types occupied 3.4%, 24.3%, 59.9%, and 12.3%, respectively.

The contrast-enhanced portal venous phase CT images were all available. Image acquisition procedure is described in Text S3. Detailed CT protocols are given in Table S1. By contouring along the margin of tumor on the slice with largest tumor area, all the tumor regions of interest (ROIs) were manually delineated using ITK-SNAP (version 3.6, http://www.itksnap.org) and thereafter validated by senior radiologists blind to other information of corresponding patients. Baseline clinical factors included age, sex, clinical T stage, and clinical N stage.

2.2 | Overall experimental design

Two binary classification experiments were conducted. For experiment A, we investigated if the proposed method could distinguish Borrmann type IV (denoted as '0') from the other three types (denoted as '1'). For experiment B, the differentiation of Borrmann type II (denoted as '0') from III (denoted as '1') was carried out only concerning Borrmann type II and III AGC patients. In each experiment, a random sample of 70% patients in center 1 was used for training, 30% were left out for internal validation, and patients in center 2 were used for external validation. Figure 1 illustrates the overall design in detail.

Two-dimensional (2D) radiomic feature extraction was initially conducted based on algorithms in Pyradiomics (version 2.1.1) and implemented by Python (version 3.7, https://www.python.org/), which was compliant with the Image Biomarker Standardization Initiative (IBSI) benchmarks.[23] The image types included original images, Laplacian of Gaussian (LoG)-filtered images, and wavelet-filtered images. Herein, the LoG filter was an edge enhancement filter with the width of the Gaussian kernel set to 1.0, 3.0, and 5.0 mm. The wavelet filter used Coiflet1 to yield four decompositions by applying either a High or a Low pass filter in each of the two dimensions, including LH, HL, HH, and LL. All the CT image slices and corresponding ROI segmentations were interpolated with B-spline interpolation algorithm to have a uniform pixel spacing of 1.0×1.0 mm^2, which helped ensure a common spatial resolution for the reproducibility of radiomic features. The gray values were discretized into equally spaced bins using a fixed bin size of 10 Hounsfield Units to allow for different ranges of intensities in ROIs, while still keeping the texture and intensity-based features informative and comparable. A total of 758 radiomic features were extracted from each CT image, including 14 morphology features, 144 intensity features, and 600 texture features (details are summarized in Table S2 and Text S4). All the extracted radiomic features were standardized by z-score method using the mean and standard deviation parameters calculated based on the training cohort. The processed radiomic features should have mean values of 0 and standard deviation values of 1.

2.3 | Training basic classification models

We trained basic classification models by integrating conventional feature selection methods and machine learning classifiers in the training cohorts based on R (version 3.6.0; https://www.r-project.org/; R packages are summarized in Text S5) and Python.

Three common feature selection methods (Figure 1C): the least absolute shrinkage and selection operator (LASSO),[24] the minimum redundancy maximum relevance (mRMR),[25] and recursive feature elimination (RFE)[26] were adopted. The LASSO method uses L1 regularization to obtain sparse features and finds a potential feature representation. The mRMR algorithm calculates feature importance or ranking to generate a set of top ranked features. The RFE method adopts a backwards feature selection strategy to find the optimal feature subset.

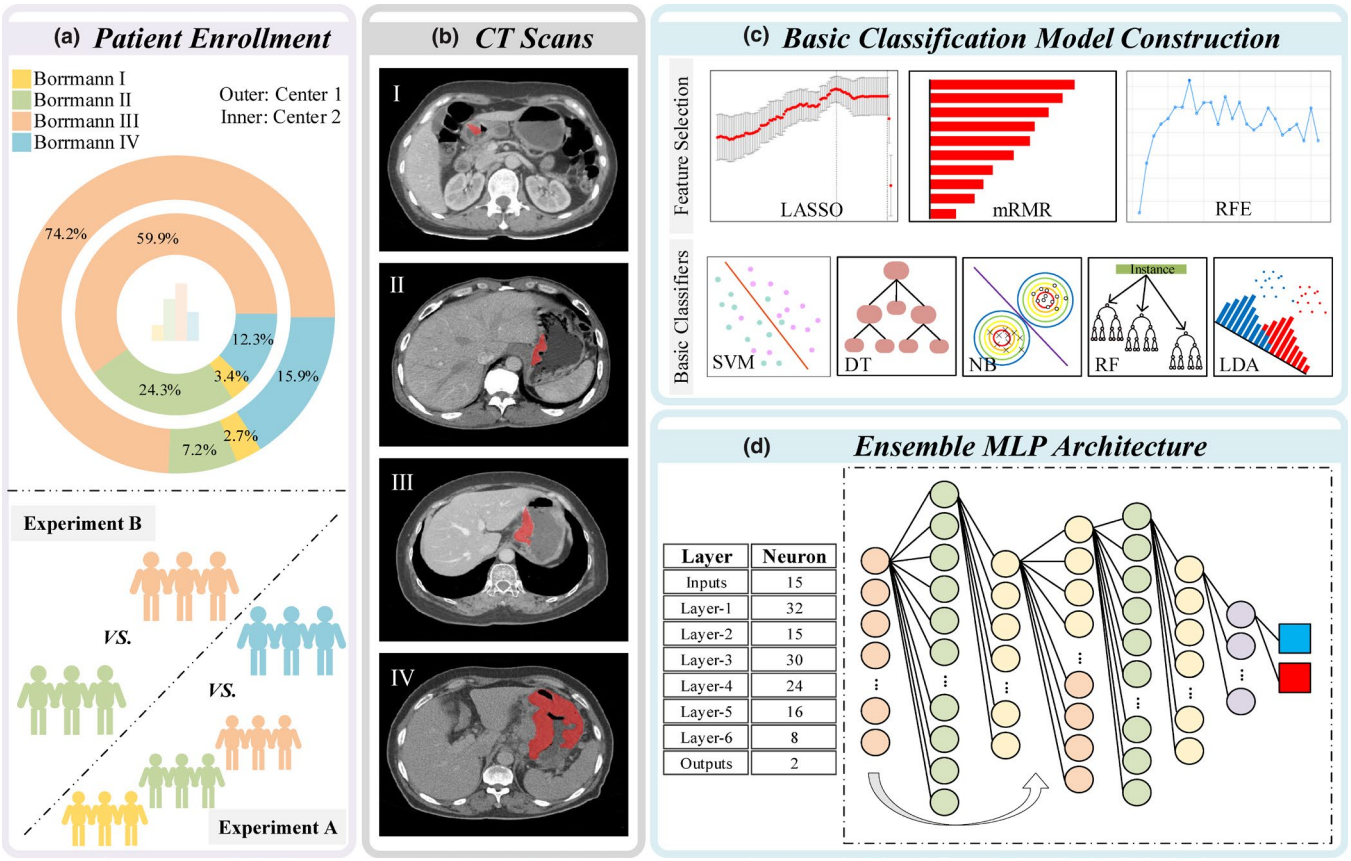

FIGURE 1 The overall study workflow, including (a) patient enrollment, (b) CT scan preprocessing, (c) basic classification model construction, and (d) ensemble MLP network development. LASSO, least absolute shrinkage selection operator; mRMR, minimum redundancy maximum relevance; RFE, recursive feature elimination; SVM, support vector machine; DT, decision tree; NB, naïve Bayes; RF, random forest; LDA, linear discriminant analysis; MLP, multilayer perceptron

Five typical machine learning classifiers in cancer prediction (Figure 1C): support vector machine (SVM),[27] decision tree (DT),[27] NB,[13] RF,[28] and linear discriminant analysis (LDA)[29] were used to construct classification models subsequently. The SVM maps the input vectors into a higher dimensional feature space and identifies a hyperplane to separate the data into two classes by maximizing the marginal distance between the hyperplane and the closest data points to boundary. DT acts as a tree-structured scheme that forms the inputs as nodes and creates the decision outcomes as leaves to specifically conjecture about the class of a new sample. NB is a probabilistic classifier based on Bayes' rule and strong conditional independence assumption among features. RF combines randomly sampled tree vectors and gives the final predictive outcome that gets the majority of votes. LDA tries to find a linear combination of features of different categories and in turn characterizes or distinguishes them.

Detailed hyperparameter and optimization settings of feature selection methods and machine learning classifiers are summarized in Text S6. In this study, we combined the feature selection methods and machine learning classifiers to give 15 basic classification models in experiment A and B, respectively.

2.4 | MLP-based multimodel ensemble learning

Developments in MLP structures have enabled the ensemble learning of multiple model predictions.[28] The pioneering work of MLP focused on fully connecting neurons and training in a layer-by-layer fashion. In this study, we distilled this insight into radiomic feature-based multimodel ensemble learning for Borrmann classification. The predictions of the 15 basic classification models were used as the inputs and fed to our designed six-layer perceptron for ensemble learning. Figure 1D illustrates this layout schematically.

The main computational units of our ensemble MLP are six fully connected layers. The leftmost is the input layer with predictions of 15 basic classification models as neurons. The rightmost is the output layer with two neurons. The middle are fully connected layers in which we experiment with different numbers of hidden neurons. Specifically, we concatenate the outputs after the second fully connected layer and the original inputs to encourage the information flow.

The ensemble MLP was run for 100 epochs using a stochastic gradient descent optimizer, cross entropy

loss function, and a learning rate of 0.001 in a full batch learning, which meant that all the training samples were fed to the MLP in each iteration. The ensemble MLP was trained in Python using PyTorch (version 1.3.0) and performed on a machine with an Intel Core i9-9900K CPU and 16 GB memory. For model availability, we have uploaded online the basic classification model parameters and weights as well as codes for the MLP network (please see http://www.radiomics.net.cn/post/137).

2.5 | Model performance evaluation and statistical analysis

The classification ability of the ensemble MLP in experiment A and B was compared with the 15 basic classification models, with regard to area under receiver operating characteristic (ROC) curves (AUC), specificity, sensitivity, and Youden index (= specificity +sensitivity - 1).[30] Also, we empirically assessed the clinical effectiveness of the ensemble MLP by decision curve analysis (DCA) and Kaplan–Meier survival analysis along with log-rank tests. In both experiments, we collected the overall survival (OS) information of AGC patients in external validation cohort, aiming to validate whether the predictions by ensemble MLP was also able to risk stratify the survival outcomes. The OS was measured from the date of surgery to the date of tumor-related death or the date of the last follow-up, which was patient-specific. Further in experiment B, subgroup analysis according to clinical T/N stage was carried out to investigate whether the ensemble MLP could differentiate Borrmann type II from III in certain subgroups.

In univariate analysis, two-sample *t*-test was used for numerical variables. A Fisher's exact test or chi-square test was applied for categorical characteristics. A two-tailed $P < 0.05$ represents a statistical significance level. All the statistical analysis was conducted in R software.

3 | RESULTS

3.1 | Experiment A: specific identification of Borrmann type IV

3.1.1 | Baseline clinical factors and radiomic feature discovery

As shown in Table 1 and Table S3, Borrmann type IV only occupied 18.4%, 10.1%, and 12.3% of the training, internal validation, and external validation cohorts, respectively. No significant differences between the training and internal validation cohorts were caught for the four clinical factors ($P = 0.055–0.762$). The clinical N stage was significantly associated with Borrmann type I/II/III vs. IV ($P = 0.028$, 0.005, 0.021) in all the three cohorts. Radiomic feature discovery is shown in Figure 2A and explained in Table S4.

3.1.2 | Model performance assessment

In internal validation cohort, the ensemble MLP (AUC: 0.767, 95% confidence interval [CI]: 0.634–0.901) surpassed all the 15 basic classification models (AUCs: 0.558–0.764) with a specificity of 0.667 and a sensitivity of 0.795 (Figure 2B). Performance of a junior radiologist with six-year experience was also compared. The radiologist was shown CT images of all the Borrmann types and asked to judge patients as either Borrmann type I/II/III or IV. The radiologist achieved excellent sensitivity but very low specificity. In external validation cohort, the ensemble MLP achieved an AUC of 0.702 (95% CI: 0.627–0.777), a specificity of 0.750, and a sensitivity of 0.610.

Detailed performance comparison with basic classification models is shown in Figure S1 and Table S5. Delong tests between each two models are illustrated in a heatmap in Figure 2C and Figure S2a. Considering an over 4:1 ratio of Borrmann type I/II/III to IV, basic classification models seemed more likely to obtain very high sensitivity (median [range]: 0.870 [0.652–0.950] in internal validation cohort, 0.645 [0.504–0.879] in external validation cohort) and low specificity (0.500 [0.167–0.778] in internal validation cohort, 0.611 [0.250–0.861] in external validation cohort) (Figure 2D and Figure S3a). The ensemble MLP, in this case, achieved a highest Youden index of 0.462 over all the basic classification models (Youden index, 0.117–0.437) in internal validation cohort and a second highest Youden index of 0.360 in external validation cohort. We further presented the normalized confusion matrices for the ensemble MLP. As shown in Figure S4, the false positive rates and false negative rates were low, indicating that only a small part of misclassifications took place. The false positive rates and false negative rates were similar in values, which indicated that the ensemble MLP misclassified Borrmann IV as I/II/III and Borrmann I/II/III as IV to a similar extent.

3.1.3 | Clinical usefulness

Decision curves in internal validation cohort (Figure 2E) demonstrated that the ensemble MLP could provide more guidance than all Borrmann type IV scheme and no Borrmann type IV scheme. The median OS for all the AGC patients in center 2 was 28 months (observed: 133/292, 45.5%). Patients could be significantly risk stratified by actual Borrmann types (Borrmann I/II/III vs. IV) and the ensemble MLP predicted Borrmann types (log-rank test, $P < 0.0001$, Figure 2F) where Borrmann

TABLE 1 Baseline clinical factors of AGC patients in experiment A

	Center 1 (n = 597)						Center 2 (n = 292)		
	Training cohort (n = 418)			Internal validation cohort (n = 179)			External validation cohort (n = 292)		
Clinical factors	Borrmann I/II/III (n = 341)	Borrmann IV (n = 77)	*P*	Borrmann I/II/III (n = 161)	Borrmann IV (n = 18)	*P*	Borrmann I/II/III (n = 256)	Borrmann IV (n = 36)	*P*
Age, years			0.075			0.792			0.801
Mean ±SD	64.0 ± 9.0	61.8 ± 9.9		63.9±10.3	63.3 ± 9.4		55.3 ± 9.4	55.7 ± 9.7	
Sex, No. (%)			0.859			1.000			0.532
Male	247 (72.4)	55 (71.4)		119 (73.9)	13 (72.2)		197 (77.0)	26 (72.2)	
Female	94 (27.6)	22 (28.6)		42 (26.1)	5 (27.8)		59 (23.0)	10 (27.8)	
Clinical T stage, No. (%)			0.733			0.751			<0.001
T1	4 (1.2)	1 (1.3)		7 (4.3)	0 (0.0)		0 (0.0)	0 (0.0)	
T2	51 (15.0)	8 (10.4)		27 (16.8)	3 (16.7)		45 (17.6)	1 (2.8)	
T3	183 (53.6)	44 (57.1)		74 (46.0)	7 (38.9)		144 (56.2)	13 (36.1)	
T4	103 (30.2)	24 (31.2)		53 (32.9)	8 (44.4)		67 (26.2)	22 (61.1)	
Clinical N stage, No. (%)			0.028			0.005			0.021
N0	52 (15.2)	5 (6.5)		33 (20.5)	1 (5.6)		63 (24.6)	2 (5.5)	
N1	153 (44.9)	30 (39.0)		68 (42.2)	9 (50.0)		54 (21.1)	6 (16.7)	
N2	104 (3.5)	28 (36.4)		54 (33.5)	4 (22.2)		52 (20.3)	9 (25.0)	
N3	32 (9.4)	14 (18.2)		6 (3.7)	4 (22.2)		87 (34.0)	19 (52.8)	

Note: In univariate analysis, two-sample t-test was used for numerical variables. Fisher's exact test or Chi-square test was applied for categorical characteristics. AGC, advanced gastric cancer; SD, standard deviation.

type IV patients were significantly associated with shorter OS and worse prognosis.

3.2 | Experiment B: differentiation of Borrmann type II from III

3.2.1 | Baseline clinical factors and radiomic feature discovery

Borrmann type II occupied 8.5%, 9.6%, and 28.9% of the three cohorts, respectively (Table S6). No significant difference was captured between training and internal validation cohort in Borrmann classification (Chi-square test, $P = 0.706$). As analyzed in Table 2, clinical T stage and clinical N stage were both univariately significant in all the three cohorts. Radiomic feature discovery is illustrated in Figure 3A and Table S7.

3.2.2 | Classification performance evaluation of the ensemble MLP

The ensemble MLP still worked in experiment B. In internal validation cohort, the distinguishing ability of the ensemble MLP was indicated with an AUC of 0.768 (95% CI, 0.626–0.911), a specificity of 0.714, and a sensitivity of 0.652. Individual radiologist performances from one junior radiologist with 6-year experience and one intermediate radiologist with 10-year experience are also plotted, below the ROCs of the ensemble MLP (Figure 3B). Here, the radiologists were shown only CT images of Borrmann type II and III patients, and they were informed that the dataset represented either Borrmann type II or III, and that they should decide between the two. In external validation cohort, the ensemble MLP reached an AUC of 0.731 (95% CI, 0.664–0.799), a specificity of 0.620, and a sensitivity of 0.703.

Detailed performance comparison with basic classification models is given in Figure S5 and Table S8. Delong-test results between each two models in both validation cohorts are shown in Figure 3C and Figure S2b. Most basic classification models resulted in very high sensitivity (median [range]: 0.909 [0.523–1.000] in internal validation cohort, 0.937 [0.360–0.989] in external validation cohort) and corresponding low specificity (0.429 [0.000–0.786] in internal validation cohort, 0.254 [0.056–0.873] in external validation cohort). The ensemble MLP, however, could still compensate for the balance of specificity and sensitivity, achieving Youden indices of 0.366 and 0.323 in both

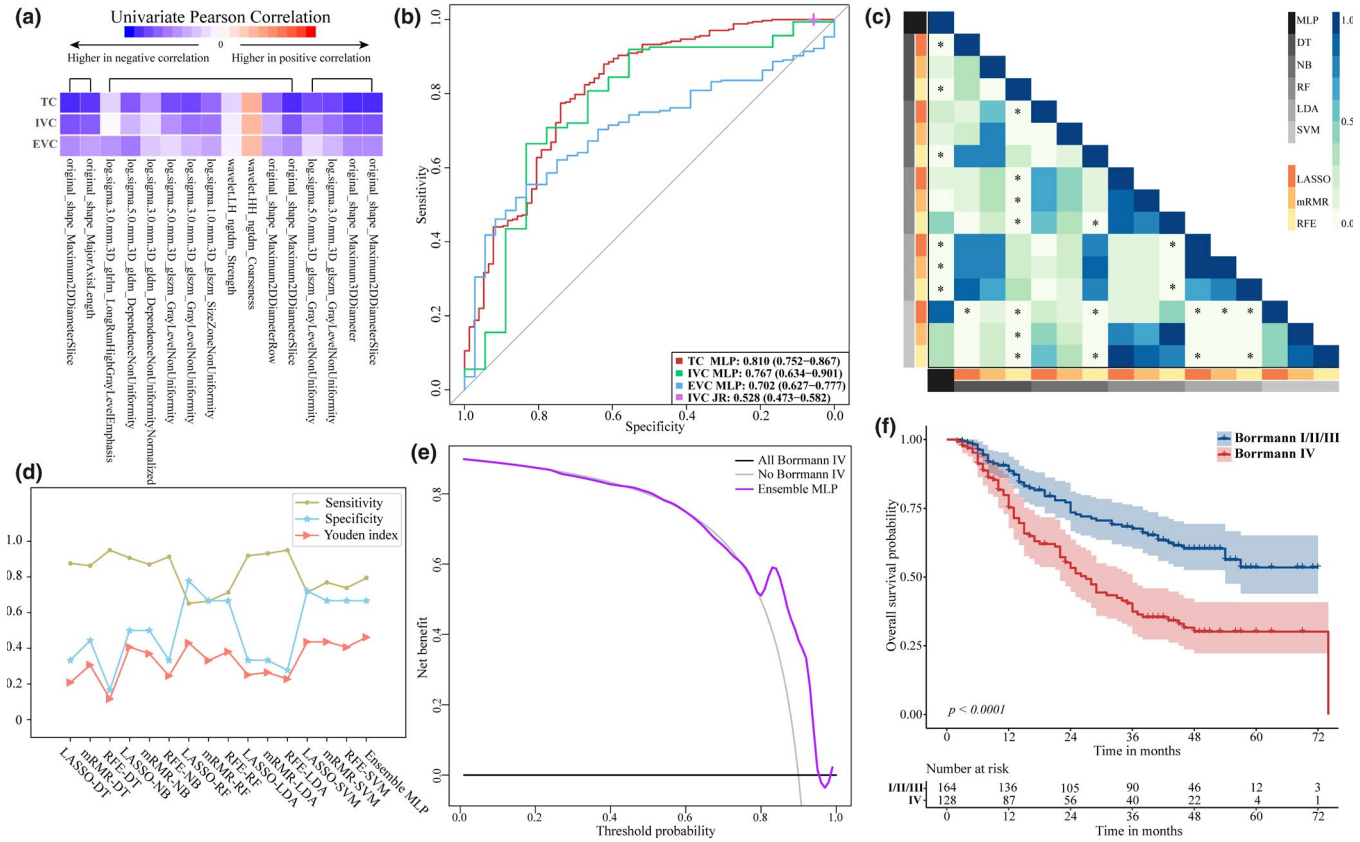

FIGURE 2 Results of experiment A. (a) Heatmap of univariate Pearson correlation coefficients between radiomic features and Borrmann types (I/II/III vs. IV). The upmost black lines summarized features selected by LASSO, mRMR, and RFE from left to right. (b) ROCs for ensemble MLP and junior radiologist performance. (c) Heatmap for Delong-tests between each two models in internal validation cohort. * represents a significant difference. (d) A line chart illustrating specificity, sensitivity, and Youden index in internal validation cohort. (e) Decision curves. (f) Kaplan-Meier curves by ensemble MLP predicted Borrmann I/II/III vs. IV types. LASSO, least absolute shrinkage and selection operator; mRMR, minimum redundancy maximum relevance; RFE, recursive feature elimination; TC, training cohort; IVC, internal validation cohort; EVC, external validation cohort; ROC, receiver operating characteristics; MLP, multilayer perceptron; JR, junior radiologist; DT, decision tree; NB, naïve Bayes; RF, random forest; LDA, linear discriminant analysis; SVM, support vector machine

validation cohorts (Figure 3D and Figure S3b). We also presented the normalized confusion matrices for our ensemble MLP in internal and external validation cohorts in Figure S4.

3.2.3 | Clinical use analysis

Decision curves in internal validation cohort (Figure 4A) demonstrated that the ensemble MLP could provide more net benefit than all Borrmann type II scheme or all Borrmann type III scheme in clinical practice. For survival analysis concerning Borrmann type II and III AGC patients in center 2, the median OS was 32.5 months (observed: 107/246, 43.5%). Patients could be significantly separated into the high-risk and low-risk groups by actual Borrmann types (Borrmann II vs. III) with a log-rank test $P = 0.0025$, indicating a worse prognosis of Borrmann type III than II. Meanwhile, Borrmann types predicted by ensemble MLP could also risk stratify AGC patients with a log-rank test $P < 0.0001$ (Figure 4B).

For clinical T stage, AGC patients in center 1 were divided into T1, T2, T3, and T4 subgroups. For clinical N stage, we simply separated patients into N0 and N+ (N1, N2, and N3) subgroups. As illustrated in Figure 4C and 4D, the ensemble MLP showed better differentiation ability in patients with more advanced clinical T/N stage.

4 | DISCUSSION

Specific Borrmann classification can help determine more appropriate surgical margins and improve the prognosis for AGC patients. However, to the best of our knowledge, there is yet no radiomics research predicting specific Borrmann types. In this study, we advocated an MLP-based multimodel ensemble learning architecture, which could identify Borrmann type IV automatically and improve the differentiation of Borrmann type II from III.

Borrmann type IV AGC is characterized by diffuse infiltration to the gastric wall without ulceration or distinct

TABLE 2 Baseline clinical factors of AGC patients in experiment B

	Center 1 (n = 486)						Center 2 (n = 246)		
	Training cohort (n = 340)			Internal validation cohort (n = 146)			External validation cohort (n = 246)		
Clinical factors	Borrmann II (n = 29)	Borrmann III (n = 311)	*P*	Borrmann II (n = 14)	Borrmann III (n = 132)	*P*	Borrmann II (n = 71)	Borrmann III (n = 175)	*P*
Age, years			0.744			0.387			0.370
Mean±SD	62.9 ± 9.9	63.6 ± 9.3		61.8 ± 13.0	65.0 ± 9.3		55.9 ± 8.6	54.8 ± 9.4	
Sex, No. (%)			0.065			0.906			0.332
Male	18 (62.1)	243 (78.1)		10 (71.4)	87 (65.9)		58 (81.7)	133 (76.0)	
Female	11 (37.9)	68 (21.9)		4 (28.6)	45 (34.1)		13 (18.3)	42 (24.0)	
Clinical T stage, No. (%)			<0.001			<0.001			<0.001
T1	3 (10.3)	5 (1.6)		1 (7.1)	1 (0.8)		0 (0.0)	0 (0.0)	
T2	14 (48.3)	38 (12.2)		7 (50.0)	16 (12.1)		26 (36.6)	12 (6.8)	
T3	12 (41.4)	165 (53.1)		5 (35.7)	69 (52.3)		35 (49.3)	106 (60.6)	
T4	0 (0.0)	103 (33.1)		1 (7.1)	46 (34.8)		10 (14.1)	57 (32.6)	
Clinical N stage, No. (%)			<0.001			0.002			0.029
N0	13 (44.8)	44 (14.1)		7 (50.0)	18 (13.6)		24 (33.8)	32 (18.3)	
N1	14 (48.3)	139 (44.7)		6 (42.9)	53 (40.2)		16 (22.5)	36 (20.6)	
N2	1 (3.4)	104 (33.4)		1 (7.1)	49 (37.1)		14 (19.7)	38 (21.7)	
N3	1 (3.4)	24 (7.7)		0 (0.0)	12 (9.1)		17 (23.9)	69 (39.4)	

Note: In univariate analysis, two-sample t-test was used for numerical variables. Fisher's exact test or Chi-square test was applied for categorical characteristics. AGC, advanced gastric cancer; SD, standard deviation.

elevation.[31] Because of the typical growth and morphologic characteristics, it is sometimes difficult to recognize the lesions in endoscopy examination. In contrast to endoscopy, high-contrast CT can perform better in lesion detection and characterization of Borrmann type IV.[22] Although not reaching an equivalent level to CT diagnosis, the ensemble MLP enabled the automatic identification of Borrmann type IV, which may help radiologists validate certain judgments. Furthermore, the incidence of Borrmann type IV is approximately 10–20% of all GC,[32] but it is usually regarded as an important independent prognostic factor.[3,4] In accordance with this, the significant risk stratification of AGC patients indicated by Kaplan-Meier survival curves in experiment A demonstrated similar prognostic power of actual Borrmann types and the ensemble MLP predicted types, which showed the acceptable classification ability of the ensemble MLP to some extent.

On the other hand, Borrmann type II AGC differs from III in biological characteristics and treatment decisions.[33] Survival analysis proved that Borrmann type III usually had worse prognosis than II, which was consistent with previous studies and the clinical common knowledge.[7] When differentiating the two types, Yan's study showed that the accuracies of defining Borrmann type II and III were 79.7% and 80.0% on MDCT.[5] However, CT might overclassify AGC because the extent of a microscopic tumor was greater than that of gross invasion.[5] Under such circumstance, the ensemble MLP derived from CT-based radiomics enabled the differentiation of Borrmann type II from III with good AUCs and Youden indices, acting as an auxiliary tool to help determine more reliable Borrmann types. Radiologists with less experience in CT diagnosis could be benefitted, and their work intensity may be reduced. Furthermore, subgroup analysis in experiment B showed that the differences between Borrmann type II and III were more easily captured in AGC patients with more advanced clinical T/N stage. More precise Borrmann classification may be defined for such patients preoperatively.

Analyses in both experiments were based on 2D CT image slices; however, our study for gastric cancer may not be limited by only using the largest image slices. As shown in Zhang et al.'s study,[34] 2D radiomic features (rather than 3D) were used to quantify gastric tumor characteristics. They compared the radiomic features extracted from 2D CT image slices with those from 3D CT volumes, and found that most of the two types of features had high correlations. This implied that 2D CT image analysis might be able to reflect enough information of the entire tumor to some extent.

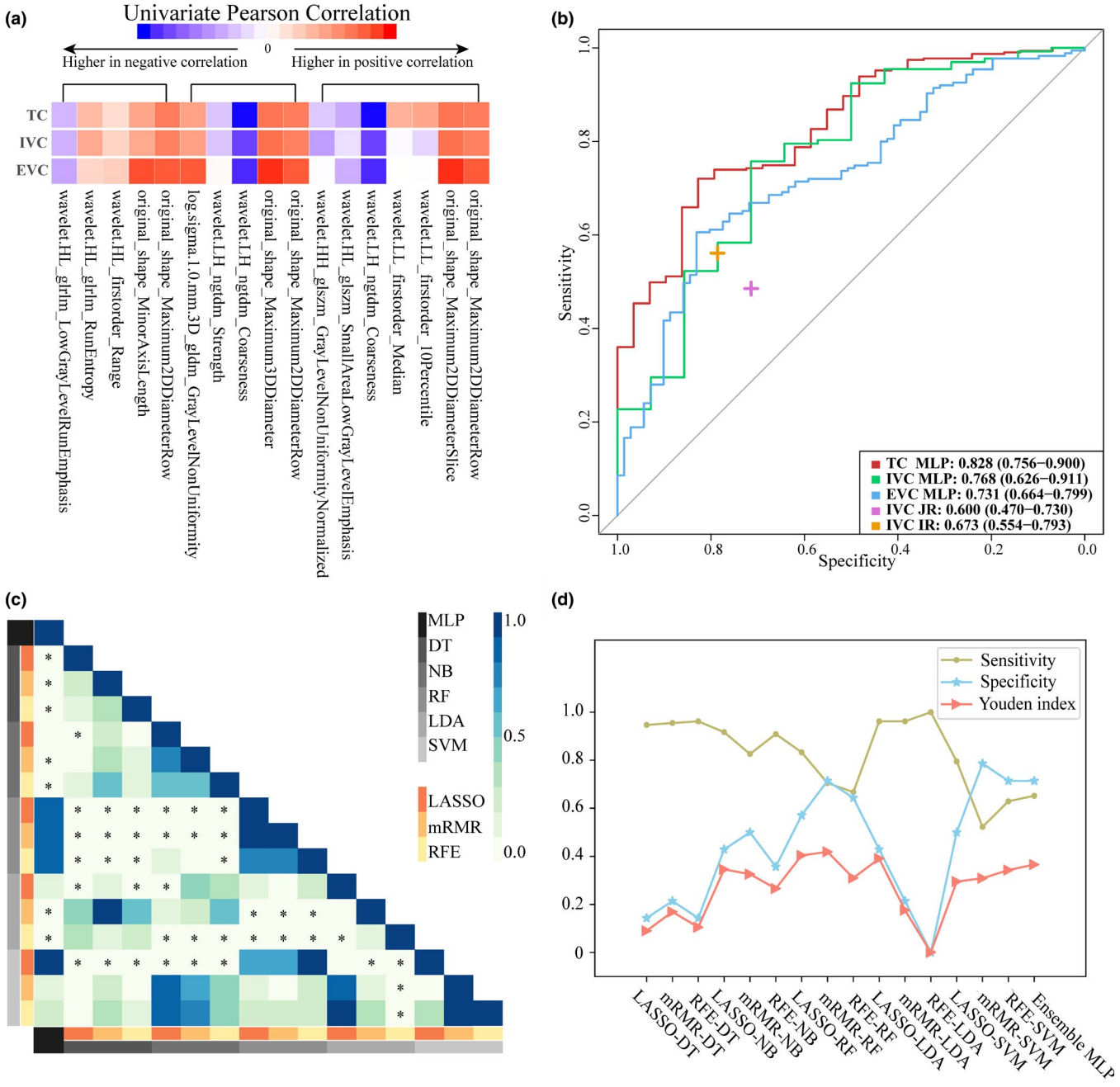

FIGURE 3 Results of experiment B. (a) Heatmap of univariate Pearson correlation coefficients between radiomic features and Borrmann types (II vs. III). The upmost black lines summarized features selected by LASSO, mRMR, and RFE from left to right. (b) ROC curves for ensemble MLP and individual radiologist performances. (c) Heatmap for Delong-tests between each two models in internal validation cohort. * represents a significant difference. (d) A line chart illustrating specificity, sensitivity, and Youden index in internal validation cohort. LASSO, least absolute shrinkage and selection operator; mRMR, minimum redundancy maximum relevance; RFE, recursive feature elimination; TC, training cohort; IVC, internal validation cohort; EVC, external validation cohort; ROC, receiver operating characteristics; MLP, multilayer perceptron; JR, junior radiologist; IR, intermediate radiologist; DT, decision tree; NB, naïve Bayes; RF, random forest; LDA, linear discriminant analysis; SVM, support vector machine

Moreover, Meng et al.'s study[35] has revealed that 2D CT annotations might be better for GC-based radiomic studies because of a better performance than 3D CT annotations. Also, they pointed out that 3D annotations sometimes brought more noise, which may interfere with effective information. Thus, analyzing the largest CT image slices can be a good choice.

In our study, the main Borrmann type of AGC was III (69.5%), far more than the incidence of I, II, and IV (2.9%, 12.8%, and 14.7%), which was similar to the distribution of Borrmann types in previous studies.[5,7] The imbalanced nature of Borrmann classification made conventional machine learning radiomics encounter the overfitting barrier more easily. From a large body of

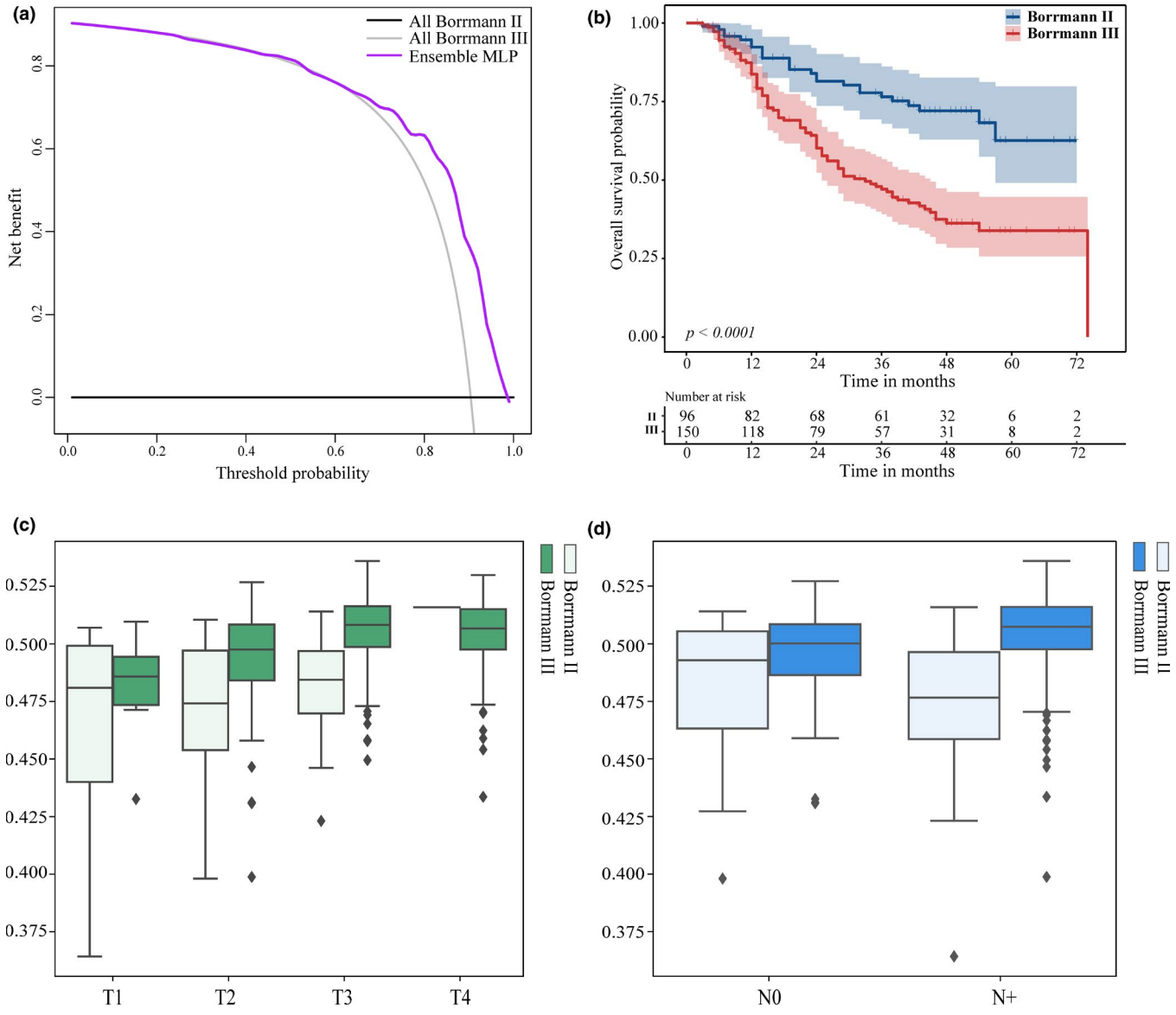

FIGURE 4 Results of experiment B. (a) Decision curves. (b) Kaplan-Meier curves by ensemble MLP predicted Borrmann types. (c) and (d) Subgroup analysis for ensemble MLP predicted Borrmann types based on clinical T/N stage. MLP, multilayer perceptron

literature in cancer prediction, each machine learning method may outperform others but still has defects in different facets.[18] Ensemble learning, however, is a viable methodology. First, the basic classification models were based on three different feature selection methods that measured the relationship between radiomic features and Borrmann types through different algorithms. Herein, LASSO generated the optimal feature subset via L1 regularization, RFE found the optimal feature subset via backwards feature selection, and mRMR calculated the feature importance/ranking. This may help provide more representative features. Second, the basic classification models were developed by using multiple machine learning classifiers that realized the classification tasks from complementary aspects of both linear regression and non-linear regression. The linear SVM and LDA, for example, focused on simple fitting, whereas RF and DT may figure out more appropriate kernel functions and build more complex connections between radiomic features and Borrmann types. Third, the MLP integrating the predictions of basic classification models may take full advantage of different feature selection results and diversiform machine learning classifiers. And the ensemble MLP produced multilevel features from hidden layers and encouraged feature reuse and information flow by feature concatenation. Specifically, optimizing the MLP with cross-entropy loss also helped alleviate the class imbalance problem quite well. The proposed networks are thus compatible. Nevertheless, other strategies to deal with the class imbalance problem (e.g., undersampling) may also be effective. That can be further explored in future studies.

There are also some limitations. First, tumor ROIs were manually segmented, thus making it a labor-intensive task. Automatic or semi-automatic segmentation may be better. Second, we followed the recommendation of filtering after image resampling by the IBSI benchmarks, which may lead to the failure of estimating how strongly this would affect the wavelet feature values to some extent. Thirdly, pathological Borrmann classification may not be acquired in some centers. Fourthly, there remains domain shift between the datasets from different centers, causing the model performance decrease in external validation. Further improvements in relieving domain shift may help achieve better results.

5 | CONCLUSIONS

In short, this study presented a CT radiomics-based ensemble MLP network and improved the specific Borrmann classification. The idea of this study not only posed focus on preoperative specific image-guided Borrmann classification in clinical practice, but also provided a non-invasive auxiliary tool for personalized strategy in AGC patients.

ACKNOWLEDGMENTS

This work was supported by National Key R&D Program of China (2017YFA0205200), National Natural Science Foundation of China (82022036, 91959130, 81971776, 81771924, 62027901, 81227901, 81930053, 81527805), the Beijing Natural Science Foundation (L182061), Strategic Priority Research Program of Chinese Academy of Sciences (XDB 38040200), Chinese Academy of Sciences under Grant No. GJJSTD20170004 and QYZDJ-SSW-JSC005, the Youth Innovation Promotion Association CAS (2017175), and the Project of High-Level Talents Team Introduction in Zhuhai City (Zhuhai HLHPTP201703). The authors would like to acknowledge the instrumental and technical support of Multimodal Biomedical Imaging Experimental Platform, Institute of Automation, Chinese Academy of Sciences. We also appreciate Yi Ding for his help in gathering data for this study.

CONFLICT OF INTEREST

The authors have no conflicts to disclose.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

REFERENCES

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2018;68(6):394-424.
2. Ferlay J, Colombet M, Soerjomataram I, et al. Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods. *Int J Cancer*. 2019;144(8):1941-1953.
3. Luo Y, Gao P, Song Y, et al. Clinicopathologic characteristics and prognosis of Borrmann type IV gastric cancer: a meta-analysis. *World J Surg Oncol*. 2016;14(1):49.
4. An JY, Kang TH, Choi MG, Noh JH, Sohn TS, Kim S. Borrmann type IV: an independent prognostic factor for survival in gastric cancer. *J Gastrointestinal Surgery*. 2008;12(8):1364-1369.
5. Yan C, Bao X, Shentu W, et al. Preoperative gross classification of gastric adenocarcinoma: comparison of double contrast-enhanced ultrasound and multi-detector row CT. *Ultrasound Med Biol*. 2016;42(7):1431-1440.
6. National Health Commission Of The People's Republic Of C. National Health Commission Of The People's Republic Of C. Chinese guidelines for diagnosis and treatment of gastric cancer 2018 (English version). *Chinese J Cancer Res*. 2019;31(5):707-737.
7. Li C, Oh SJ, Kim S, et al. Macroscopic Borrmann type as a simple prognostic indicator in patients with advanced gastric cancer. *Oncology*. 2009;77(3–4):197-204.
8. Lambin P, Rios-Velazquez E, Leijenaar R, et al. Radiomics: extracting more information from medical images using advanced feature analysis. *Eur J Cancer*. 2012;48(4):441-446.
9. Dong D, Tang L, Li Z-Y, et al. Development and validation of an individualized nomogram to identify occult peritoneal metastasis in patients with advanced gastric cancer. *Ann Oncol*. 2019;30(3):431-438.
10. Wang S, Feng C, Dong DI, et al. Preoperative computed tomography-guided disease-free survival prediction in gastric cancer: a multicenter radiomics study. *Med Phys*. 2020;47(10):4862-4871.
11. Dong D, Fang M-J, Tang L, et al. Deep learning radiomic nomogram can predict the number of lymph node metastasis in locally advanced gastric cancer: an international multicenter study. *Ann Oncol*. 2020;31(7):912-920.
12. Parmar C, Grossmann P, Bussink J, Lambin P, Aerts HJ. Machine learning methods for quantitative radiomic biomarkers. *Sci Rep*. 2015;5:13087.
13. Wu W, Parmar C, Grossmann P, et al. Exploratory study to identify radiomics classifiers for lung cancer histology. *Front Oncol*. 2016;6:71.
14. Parmar C, Grossmann P, Rietveld D, Rietbergen MM, Lambin P, Aerts HJ. Radiomic machine-learning classifiers for prognostic biomarkers of head and neck cancer. *Front Oncol*. 2015;5:272.
15. Zhang B, He X, Ouyang F, et al. Radiomic machine-learning classifiers for prognostic biomarkers of advanced nasopharyngeal carcinoma. *Cancer Lett*. 2017;403:21-27.
16. Leger S, Zwanenburg A, Pilz K, et al. A comparative study of machine learning methods for time-to-event survival data for radiomics risk modelling. *Sci Rep*. 2017;7(1):13206.
17. Freund Y, Schapire RE. Experiments with a new boosting algorithm. *ICML*. 1996;96:148-156.
18. Mamoshina P, Vieira A, Putin E, Zhavoronkov A. Applications of deep learning in biomedicine. *Mol Pharm*. 2016;13(5):1445-1454.
19. Bourlard H, Kamp Y. Auto-association by multilayer perceptrons and singular value decomposition. *Biol Cybern*. 1988;59(4–5):291-294.
20. Ma Z, Fang M, Huang Y, et al. CT-based radiomics signature for differentiating Borrmann type IV gastric cancer from primary gastric lymphoma. *Eur J Radiol*. 2017;91:142-147.
21. Association JGC. Japanese classification of gastric carcinoma: 3rd English edition. *Gastric Cancer*. 2011;14(2):101-112.
22. Kim JI, Kim YH, Lee KH, et al. Type-specific diagnosis and evaluation of longitudinal tumor extent of borrmann type IV

gastric cancer: CT versus gastroscopy. *Korean J Radiol*. 2013;14(4):597-606.
23. Zwanenburg A, Vallières M, Abdalah MA, et al. The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology*. 2020;295(2):328-338.
24. Tibshirani R. Regression shrinkage and selection via the lasso. *J Roy Stat Soc: Ser B (Methodol)*. 1996;58(1):267-288.
25. Peng H, Long F, Ding C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans Pattern Anal Mach Intell*. 2005;8:1226-1238.
26. Granitto PM, Furlanello C, Biasioli F, Gasperi F. Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products. *Chemometrics and Intelligent Laboratory Systems*. 2006;83(2):83-90.
27. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. *Comput Struct Biotechnol J*. 2015;13:8-17.
28. Xiao Y, Wu J, Lin Z, Zhao X. A deep learning-based multi-model ensemble method for cancer prediction. *Comput Methods Programs Biomed*. 2018;153:1-9.
29. Fernández-Delgado M, Cernadas E, Barro S, Amorim D. Do we need hundreds of classifiers to solve real world classification problems? *J Mach Learning Res*. 2014;15(1):3133-3181.
30. Hawass N. Comparing the sensitivities and specificities of two diagnostic procedures performed on the same group of patients. *Br J Radiol*. 1997;70(832):360-366.
31. Kim DY, Kim HR, Kim YJ, Kim S. Clinicopathological features of patients with Borrmann type IV gastric carcinoma. *ANZ J Surg*. 2002;72(10):739-742.
32. Bollschweiler E, Boettcher K, Hoelscher AH, et al. Is the prognosis for Japanese and German patients with gastric cancer really different? *Cancer*. 1993;71(10):2918-2925.
33. Association JGC. Japanese gastric cancer treatment guidelines 2014 (ver. 4). *Gastric Cancer*. 2017;20(1):1-19.
34. Zhang W, Fang M, Dong DI, et al. Development and validation of a CT-based radiomic nomogram for preoperative prediction of early recurrence in advanced gastric cancer. *Radiother Oncol*. 2020;145:13-20.
35. Meng L, Dong DI, Chen X, et al. 2D and 3D CT Radiomic features performance comparison in characterization of gastric Cancer: a multi-center study. *IEEE J Biomed Health Informat*. 2021;25(3):755-763.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Wang S, Dong D, Zhang W, et al. Specific Borrmann classification in advanced gastric cancer by an ensemble multilayer perceptron network: a multicenter research. *Med Phys*. 2021;00:1–12. https://doi.org/10.1002/mp.15094