IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 18, 2021, accepted March 9, 2021, date of publication March 29, 2021, date of current version April 8, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3069191

Multimodal AI System for the Rapid Diagnosis and Surgical Prediction of Necrotizing Enterocolitis

WENJING GAO[1,*], YUANYUAN PEI[1,*], HUIYING LIANG[1], JUNJIAN LV[2], JIALE CHEN[2], AND WEI ZHONG[2]

[1]Institute of Pediatrics, Guangzhou Women and Children's Medical Center, Guangzhou Medical University, Guangzhou 511436, China
[2]Department of Neonatal Surgery, Guangzhou Women and Children's Medical Center, Guangzhou Medical University, Guangzhou 511436, China

Corresponding author: Wei Zhong (zhongwei123455@gmail.com)

This work was supported by the National Key Research and Development Program under Grant 2018YFC1315402.

*Wenjing Gao and Yuanyuan Pei are co-first authors.

ABSTRACT The rapid diagnosis and surgical prediction of necrotizing enterocolitis (NEC) remain a challenge because its complex pathogenesis has not been completely elucidated, and no single medical examination is specific for diagnosing NEC. Artificial intelligence (AI) has proven the robustness of multivariate analysis and been widely used in the diagnosis of complex diseases in the past decade. In this paper, a new multimodal AI system including feature engineering, machine learning (ML), and deep learning (DL) was constructed based on abdominal radiographs (ARs) and clinical data. A total of 4,535 ARs from 1,823 suspected NEC patients were analyzed by transfer learning, and then medical images and clinical parameters from 827 suspected NEC patients were used to train, validate, and test the AI system. Our results demonstrated that the system was effective in diagnosing NEC. In addition, the clinical datasets obtained one week before surgery from 379 NEC patients were studied by the multimodal AI system, and the results showed that it was capable of predicting which NEC patients required surgery. We compared the results in external test sets with those made by clinicians and found that the diagnostic and surgical predictive ability of the AI system was equivalent to that of experienced clinicians. This multimodal AI system can help clinicians improve diagnostic efficiency, reduce the number of missed diagnoses, and facilitate early diagnosis and treatment to prevent disease progression or even death.

INDEX TERMS Abdominal x-ray, AI, diagnosis, multimodal, necrotizing enterocolitis, surgery.

I. INTRODUCTION

NEC is one of the most devastating gastrointestinal emergencies in the neonatal care unit [1]. Usually, there are no clinical warning signs for acute NEC. It is estimated that up to 50% of patients need surgical intervention, 46.5% of patients do not survive after surgery, and 20% to 50% of the survivors develop long-term sequelae, such as recurrence, intestinal stenosis, short bowel syndrome, slowed growth, and neurodevelopmental disorders [2]. NEC consists of a group of complex multivariable diseases that are difficult to describe, detect, and diagnose [3], [4]. Numerous international groups have recently highlighted NEC as a research priority and have made efforts to move the field forward [1], [5], [6]. Despite the intense research performed in this field, an effective method for the rapid diagnosis and the prediction of surgical indications of NEC has not been found. In 2020, Hooven *et al.* [7] predicted NEC using microbiome data, achieving a precision-recall AUC value of only 0.7. Shi *et al.* [4] showed that the prediction and diagnosis of NEC were satisfactory in the Era of Metabolomics and Proteomics, despite the small quantity of training data [4]. Additionally, van Druten *et al.* [2], [3], [8] diagnosed NEC based on abdominal X-rays with artificial intelligence, but the results were not expected through a single medical examination. Multimodal models represent attempts to effectively simplify the complexity of multiple-factor diagnosis [9]. This study aims to establish a new multimodal AI system for NEC patients. Combined with feature engineering, a multimodal AI system was constructed via machine learning (ML) and deep learning (DL) models in series. The system was evaluated with a dataset derived from 2,245 NEC patients from

The associate editor coordinating the review of this manuscript and approving it for publication was Shadi Alawneh.

Guangzhou Women and Children's Medical Center, China, collected from 2011 to 2020. Then, we carried out a series of experimental studies. The diagnosis of NEC is heavily dependent on abdominal radiography (AR) [10]. Therefore, in the first stage, 4,535 ARs of 1,823 suspected NEC patients were collected and divided into a training dataset, a validation dataset, and an internal test dataset. Then, three DL models were made computationally effective with the training and validating ARs. We selected the best model, SENet-154, by comparing the AUC value, sensitivity, specificity, precision and accuracy of the three models in the internal test dataset. In the second stage, the SENet-154 model was trained and validated with the clinical data of 827 suspected NEC patients and 379 confirmed NEC patients obtained one week before surgery, respectively. The radiomics signatures of the ARs were obtained by transferring the learning and fine-tuning the model parameters, and then the top-performing significant features were selected by mRMR. In the third stage, the light gradient boosting machine (LightGBM) classifier was used to predict the diagnosis and surgical eligibility of NEC in combination with the radiomics signature and clinical parameters (Fig. 2). The model captured valuable information from ARs that cannot be detected by the human eye. Afterward, thermographic images were generated, which improved the value of the diagnosis and prediction of surgical eligibility for NEC. Diagnostic value: AUC 0.9337 (95% CI: 0.9028, 0.9646), sensitivity 0.9427 (95% CI: (0.9138, 0.9716)), specificity 0.8246 (95% CI: (0.7774, 0.8718)), precision 0.9476 (95% CI: (0.9199, 0.9753)), accuracy 0.9157 (95% CI: (0.8812, 0.9502)). Surgery-predictive value: AUC 0.9413 (95% CI: 0.8998, 0.9828), sensitivity 0.8500 (95% CI: 0.7869, 0.9131), specificity 0.9535 (95% CI: 0.9163, 0.9907), precision 0.9714 (95% CI: 0.9419, 1.0000), accuracy 0.8861 (95% CI: 0.8300, 0.9422). Our multimodal AI system was comparable to experienced clinicians in diagnosing and predicting the surgery eligibility for NEC. This study can be used as an auxiliary means for the clinical diagnosis and surgical eligibility prediction of NEC.

II. MATERIALS AND METHODS

A. DATASETS

1) SETTINGS AND PARTICIPANTS

This was a retrospective study on data obtained from hospitalized infants in Guangzhou Women and Children's Medical Center, China, between December 2011 and March 2020. The study was approved by the Ethics Committee of Guangzhou Women and Children's Medical Center (No. GO-2016–017) and conducted in accordance with the ethical guidelines of the Declaration of Helsinki of the World Medical Association. Informed written consent was obtained from all participants at the initial hospital visit.

We collected 2,245 consecutive infants with abdominal distention, bradycardia, lethargy, temperature instability, emesis, bloody stool, feeding intolerance, and gastric retention. The gold-standard set was a diagnosis of medical NEC and surgical NEC as defined by senior pediatricians and pediatric surgeons according to Bell's staging criteria [11] modified by Walsh and Kliegman [12]. The criteria for screening eligible infants are summarized in Table SI [13], [14]. Radiological signs were the primary criteria for the diagnosis and surgical eligibility prediction of NEC, and clinical parameters (abdominal signs and clinical and laboratory findings) were the secondary criteria. One primary and one secondary sign were used to define NEC/surgical NEC. The non-NEC group presented with intestinal dysmotility, gastroesophageal reflux, megacolon, intestinal malrotation, intestinal atresia, lactose intolerance, meconium ileus, and intestinal stenosis, which were diagnosed based on standard clinical and radiological results. Patients were excluded if they presented with spontaneous intestinal perforation ($n = 7$) and congenital malformations in other organs ($n = 4$). Accordingly, 2,234 of the 2,245 patients remained, with one group including 1,201 non-NEC and 622 NEC patients and an independent group of 411 NEC patients that included surgical NEC and medical NEC. For the first group, patients were excluded due to a lack of complete clinical parameters or poor diagnostic image quality. Finally, there were 827 patients in the first group, with 342 NEC (1,234 records) and 485 non-NEC (758 records) patients with a complete set of ARs and clinical parameters. Likewise, in the independent group of 411 NEC patients, in which 379 patients had complete case data and were thus included, 268 patients (with 268 records) had received conservative treatment without surgical intervention, while the remaining 111 patients (226 records) had undergone invasive open surgery (Fig. 1).

2) DATA DEFINITIONS

The collected NEC and non-NEC datasets included clinical patient information obtained between diagnosis and discharge from the NICU. Preoperative data were used to predict surgical probability. All data were reviewed by two board-certified pediatricians (L.J.J. and C.J.L. each with 7-14 years of experience). ARs were obtained from the infants every 6 hours. All ARs were generated using a CanoScan LiDE 700F (Canon, Beijing, China). The ARs were extracted from the picture archiving and communication system. For each infant in the NEC and non-NEC groups, three to six representative radiomics signatures (Table SII) were selected for image analysis. For surgical NEC patients, one to three AR images were collected from each patient. For medical NEC patients with multiple samples, a single AR image was randomly selected. In addition to ARs, clinical parameters previously reported in the literature [3], [7], [13], [15]–[17] were also extracted. A total of 23 clinical parameters that included demographic data (age, birth weight, gestational age, etc.), vital signs and clinical symptoms (heart rate, vomiting, bloody stool, abdominal distention, etc.), and laboratory parameters (hemoglobin, CRP, WBC, etc.) obtained during hospitalization were included. All clinical parameters were collected within 24 hours before each AR examination. Individual patient data were labeled NEC/surgical NEC when

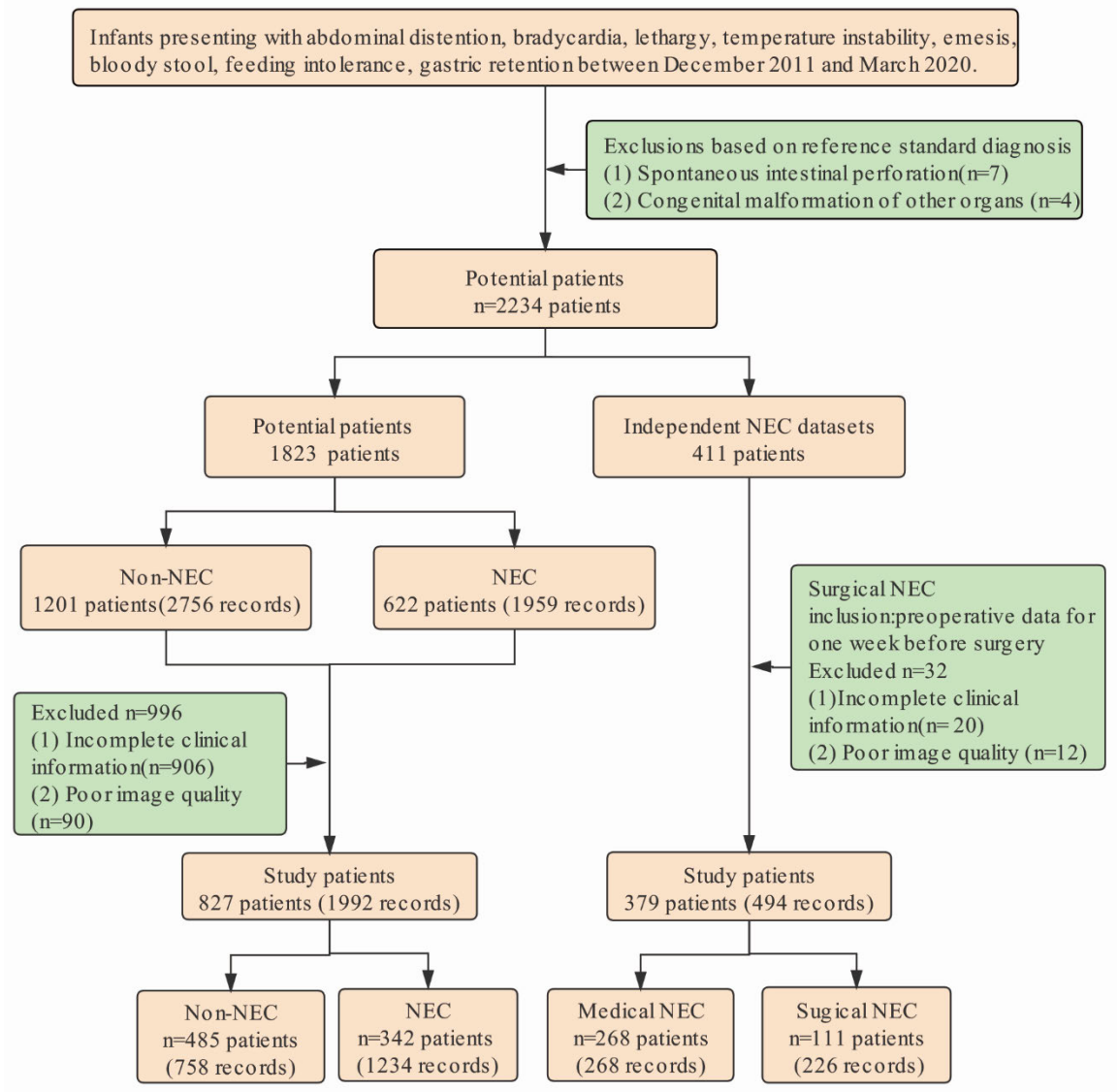

FIGURE 1. The flowchart shows study case selection according to exclusion criteria to diagnose NEC and predict NEC.

the ARs and their corresponding clinical parameters met the diagnostic criteria.

B. METHODS

1) EXPERIMENTAL PROCEDURE

In this paper, we proposed a multimodal AI system to address NEC diagnosis and surgical indication prediction. Fig. 2 shows a flowchart of our experiments, which can be divided into three stages:

1) Data preprocessing, transfer learning and best AR-DL model selection: Three DL models, ResNesT-50, SENet-154, and SE-ResNet were pretrained on ImageNet datasets, the 4,535 ARs of 1,823 suspected NEC patients were standardized and augmented, and then transfer learning was performed with the 4,535 ARs and the three models. The best model was SENet-154 as evaluated by five key indicators (AUC, sensitivity, specificity, precision, accuracy).
2) Selection of the radiomics characteristics of the ARs: SENet-154 was trained and tested with the data from the diagnosis group (1,992 ARs of 827 patients) and the surgical indication prediction group (494 ARs of 379 patients), and then the radiomics characteristics of the penultimate layer were obtained to determine the top variables through mRMR (minimum redundancy maximum correlation).
3) Establishment of a multimodal AI system and external validation (comparison between multimodal AI system and clinicians): A multimodal AI system based on ARs and other clinical parameters was constructed by LightGBM with the data from the diagnosis and surgical

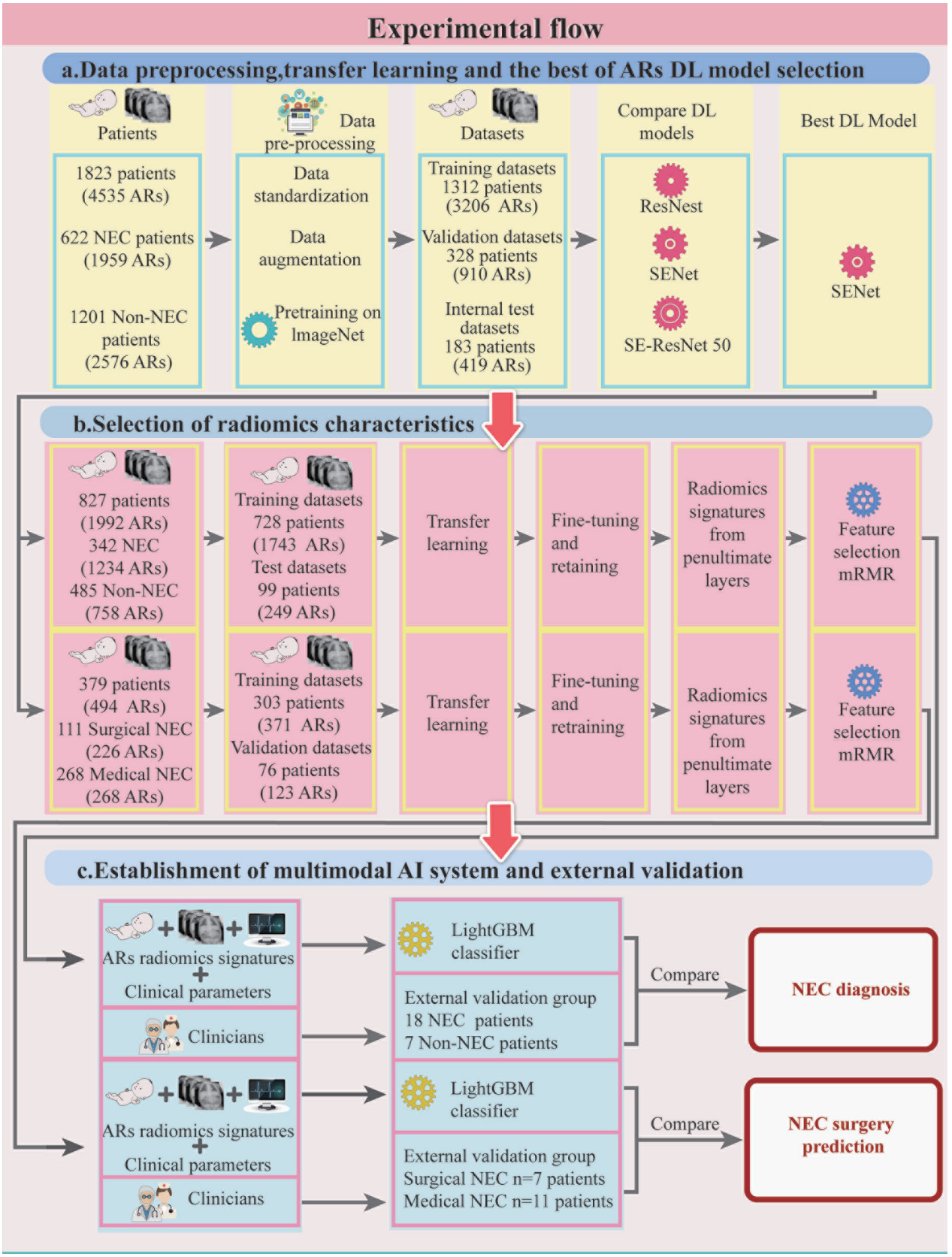

FIGURE 2. Experimental flow. a. Data preprocessing, transfer learning and best AR-DL model selection. b. Selection of radiomics characteristics of ARs. c. Establishment of the multimodal AI system and external validation (comparison between the multimodal AI system and clinicians).

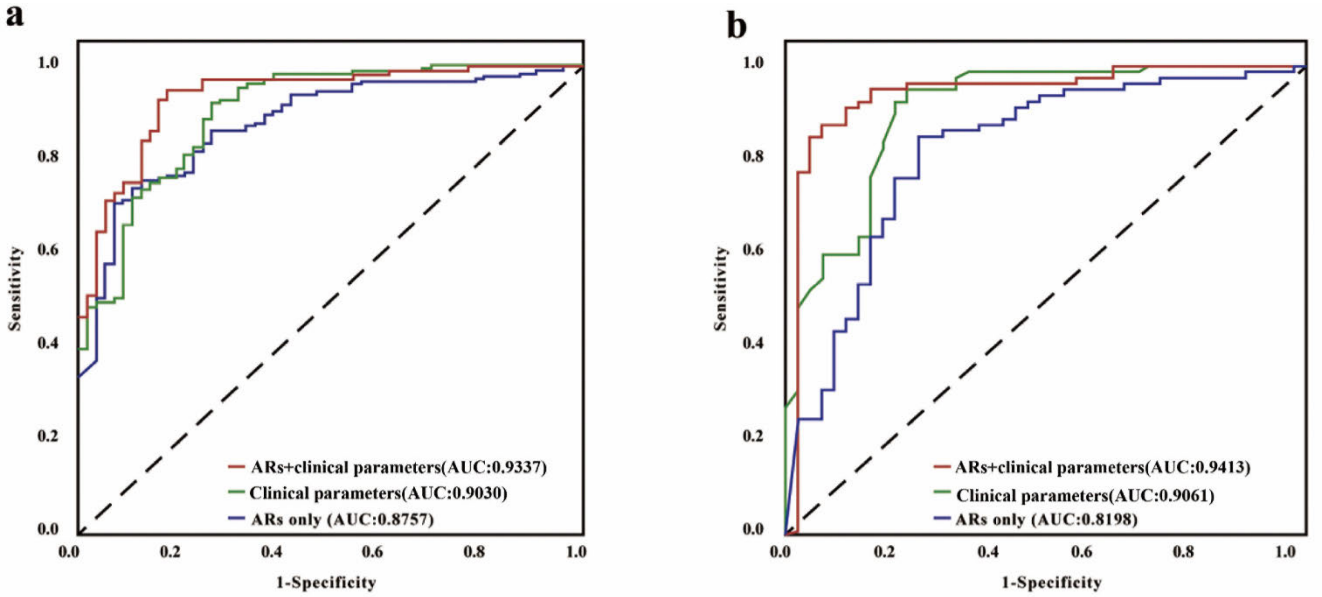

FIGURE 3. Comparison of the classification performance in identifying NEC and predicting surgical NEC. Performance is reported for three sets of features (clinical parameters, ARs, and both ARs and clinical parameters). a. Results for NEC diagnosis in test datasets with suspicious patients presenting with similar clinical symptoms of NEC. b. Results for differentiating medical NEC and surgical NEC in validation datasets.

indication prediction groups. As shown in Fig. 3, the AR dataset, the clinical parameter dataset, and the combined AR and clinical parameter dataset were run on the integrated AI system, and the results demonstrated that the multimodal AI system was superior to the model based on clinical data alone. Finally, four indicators (sensitivity, specificity, precision, and error rate) were used as evaluation criteria for external verification.

The results showed that the diagnostic and surgical prediction ability of the multimodal AI system was equivalent to that of experienced clinicians.

When NEC is diagnosed based on the modified Bell stage criteria, the label of NEC is set to 1, whereas the label of non-NEC is set to 0. For NEC patients, NEC is diagnosed as surgical NEC based on the gold standard, in which case the label is 1; otherwise, it is 0.

2) DATA PREPROCESSING

AR preprocessing: The ARs were converted into portable network graphics format and resized to 224 × 224 pixels to fit a deep convolutional neural network. The ARs were standardized by z-score normalization (subtraction by the mean intensity value and division by the standard deviation of intensity values) to reduce the effect of different reconstruction parameters. During the training portion, augmentation, including width/height shift, horizontal/vertical flip, rotation, brightness and contrast changes, and zoom, was used to expand the training dataset and improve the generalizability of the model.

This diagnostic study followed the 2015 Standards for Reporting of Diagnostic Accuracy (STARD). All ARs of the 1823 suspected patients were randomly assigned into 1 of the following three datasets: (1) the training dataset, comprising 1,312 patients (3,206 ARs), used to optimize the network weights; (2) the validation dataset, comprising 328 patients (910 ARs), used to optimize the hyperparameters; and (3) the internal test dataset, comprising 183 patients (419 ARs). The best DL model was chosen by transfer learning. Then, the DL model was trained and tested in the dataset comprising 827 patients with complete information, and the best performance DL model was fine-tuned using the pretrained weights. The 827-patient dataset were divided into two nonoverlapping datasets. The first, the training dataset, comprised 88% of the data (1,743 records of 728 patients with AR and clinical parameter data). This training set was used to update the DL system parameter. The second, the test dataset, including the remaining 12% (249 records of 99 patients with AR and clinical parameter data). The test dataset was used for independent testing. Because no hyperparameter optimization was performed, there was no need for a separate validation set. Finally, 80% of the dataset (371 records of 303 NEC patients with full AR and clinical parameter data) was used to train the surgery predicting ability of the model, and 20% of the dataset (123 records of 76 NEC patients encompassing ARs and clinical parameters) were applied as a validation dataset to select the surgery predictive hyperparameters.

If clinical parameters were missed by over 50%, the clinical parameters were excluded from the study. The remaining missed values were handled by the LightGBM classifier, which assigns them to the branch of a split that minimizes the loss [18].

3) AR DEEP LEARNING MODEL SELECTION

The ARs were loaded onto a computer with a Linux operating system (Ubuntu 16.04.4; Canonical, London, England) and a Torch deep learning framework (Version: 1.4.0;

https://pytorch.org/), a graphics processing unit acceleration operated on CUDA 10.2/cuDNN 7.6.0 (Nvidia Corporation, Santa Clara, Calif), an NVIDIA Quadro M6000 24 GB GPU for training and testing, and 256 GB RAM.

Three different effective attention DL models were evaluated in this study. Split-attention networks [19], squeeze-and-excitation networks [20] and SE-ResNet (a squeeze-and-excitation network used directly with the residual network) [20]. The three different DL models for feed-forward convolutional neural networks improve the learned feature representations to boost performance across image classification. Given an input feature map, the attention module sequentially infers attention maps. ResNeSt is a modular split-attention block, and we stacked these blocks in the style of ResNet-50 (ResNeSt-50). The squeeze-and-excitation (SE) module is an architectural unit designed to improve the representational power of a network by integrating SE blocks with a modified ResNeXt to form a SENet-154 network (SENet-154). The SE module is a lightweight and general module that can be seamlessly integrated into residual networks with 50 layers and negligible overhead and is end-to-end trainable along with base CNNs (SE-ResNet-50).

First, the DL models were pretrained on 1.2 million everyday color images from ImageNet (http://www.image-net.org/) that consisted of 1,000 categories (referred to as pretrained). For medical image analysis, the transfer learning method was implemented to identify discriminate radiomics signatures. All layers in each DL model except the last layer were inherited from the pretrained DL model for fine-tuning. Utilizing the 4,535 ARs from the 1,823 suspected patients, the best DL model was selected by evaluating the performance in the internal test datasets.

The three DL models consist of their intrinsic structures and one fully connected layer. A batch normalization layer was applied prior to the fully connected layer, and a softmax activation layer was connected to the fully connected layer, which was used to yield the prediction probabilities of the nodule candidates. To reduce the possibility of overfitting during the training process, it was necessary to take several measures. (1) Regularization: L2 regularization was used, which added a cost to the loss function of the network for large weights. As a result, a simpler model that was forced to learn only the relevant patterns in the training datasets was obtained. (2) Dropout: A dropout layer was included after the batch normalization layer with dropout probability. (3) Early stop: Training was stopped if no improvement was seen in the validation loss following a decrease in the learning rate.

4) DL FOR CLASSIFICATION AND RADIOMICS SIGNATURE SELECTION

After transfer learning, of the 4,535 ARs of the 1,823 suspected NEC patients used to select senet-154, 1,992 ARs (827 patients) were used for NEC diagnosis and 494 ARs (379 patients) were used for surgery prediction, after which the model was fine-tuned. Finally, the features of the penultimate layer of the network structure were extracted as radiomics features. mRMR is a feature selection method. For continuous features, the F-statistic of mRMR was used to calculate the correlation with class (correlation), and the Pearson correlation coefficient was used to calculate the correlation between features (redundancy). Previous studies have shown that ranking variables may help to improve the performance of the model. In this study, the mRMR algorithm was applied to select strong discriminative radiomics features. The minimum redundancy and maximum correlation were used for feature selection to identify the top variables [21].

5) ML ON RADIOMICS SIGNATURES AND CLINICAL PARAMETERS

LightGBM is an ML classifier and an open-source Python implementation of a gradient boosting framework (https://lightgbm.readthedocs.io/en/latest/). LightGBM mainly consists of two algorithms: the gradient-based one-side sampling (GOSS) algorithm, which excludes most samples with small gradients from the perspective of sample reduction and uses the remaining samples to calculate information gain; the exclusive feature bundling (EFB) algorithm, which bundles mutually exclusive features from the perspective of feature reduction; and a greedy algorithm with a constant approximation ratio, which is used to solve the problem. The LightGBM is a tree model, as each node in the tree can be converted to IF-THEN rules that are easily understandable, and can significantly outperform the other tree learning classifiers, such as extreme gradient boosting (XGBoost) and random forest [22]. Many data scientists have applied LightGBM to solve classification problems and have achieved excellent results. LightGBM has also been successfully used in medical studies [18]. Radiomics signatures and clinical parameters were used for diagnosing NEC and predicting surgical NEC with LightGBM as the classifier.

6) FEATURE IMPORTANCE

We implemented LightGBM to interpret the effects and relative contributions of the radiomics signatures and clinical parameters on the diagnosis and prediction models. The LightGBM classifier has the ability to evaluate the importance of features. Feature importance refers to a class of techniques for assigning scores to be input to a predictive model that indicates the relative importance of each feature. In LightGBM, feature importance is calculated as the number of times the feature is used in a model. A total of 58 feature importance values for diagnosing NEC and 49 feature importance values for predicting surgical NEC were obtained and ranked.

7) STATISTICAL ANALYSIS

The associations between clinical parameters and NEC/surgical NEC were evaluated with the independent Mann-Whitney U test for continuous variables and the

chi-square test for categorical variables. Two-sided $p < 0.05$ was considered to indicate a statistically significant difference. The predictive power and diagnostic performance indicators were measured using the area under the receiver operating characteristic curve (AUC), sensitivity, specificity, and precision, and accuracy was calculated from the confusion matrix to quantify the performance of the proposed classification algorithms. Youden's index was used to determine the optimal threshold. Since the number of reliably labeled NECs was very limited, we can estimate the 95% confidence interval (95% CI) for the performance indicators using the standard approach for computing the confidence interval for proportions [21] by using the following formula in Python:

$$CI = \hat{P} \pm Z * \sqrt{\frac{\hat{P}\left(1-\hat{P}\right)}{N}}$$

where z denotes the significance level of the confidence interval (the number of standard deviations of the Gaussian distribution). Here, we used a 95% confidence interval, for which the corresponding value of z is 1.96. $\hat{P}$ is the proportion.

8) ACTIVATION MAPPING

Convolutional layers naturally retain spatial information that is lost in fully connected layers, and we expected the last convolutional layers to have the best compromise between high-level semantics and detailed spatial information. To generate activation maps, we focused on explaining the output layer decisions only. Grad-CAM uses the gradient information flowing into the last convolutional layer of the CNN to assign importance values to each neuron for a particular decision of interest and generate activation heatmaps (gradients) during backpropagation [23]. The heat maps were then normalized and enlarged to match the input image size, thus indicating where the DL model was focused. The blue and red colors on the heat maps indicate lower and higher activation values, respectively. PyTorch tool kits and Python 3.7 were required to implement this model.

9) EXTERNAL VALIDATION

The external test set included 25 patients (11 with medical NEC, 7 with surgical NEC and 7 with non-NEC) and was used to compare the performance of the AI system with that of clinicians with different levels of experience in both making a binary decision between NEC and non-NEC and in predicting the NEC patients who required surgery. Each patient in this set had only one record. The patients were anonymized and independently presented to 2 senior clinicians (10-20 years of work experience), 2 junior clinicians (3-10 years of work experience), and 1 resident (2 years of work experience). These clinicians were invited to read the same ARs and clinical parameters as those used for the AI system. The analysis was performed under double-blind conditions.

III. RESULTS

A. STUDY ON THE NEC COHORT

A total of 827 patients (median age 2.00 days (0.00, 12.00) [IQR]) were included; 342 patients (41.35%) were positive for NEC (median age 0.00 days (0.00, 6.00) [IQR] days), and 485 (58.65%) were categorized as non-NEC (median age 7.00 days (2.00, 20.00) [IQR]). The clinical parameters of the patients with NEC and non-NEC are shown in Table 1. Eighteen of the 23 clinical parameters were significantly different between the NEC group and the non-NEC group ($P < 0.05$), whereas sex, gastric residual, vomiting, abdominal distention, and lethargy were similar. A total of 111 patients (median age 0.00 days (0.00, 10.75) [IQR]) underwent surgery, and 268 patients (median age 0.00 days (0.00,8.00) [IQR]) underwent conservative treatment (Table 2). There were no significant differences in 14 of the variables between the surgical and conservative groups. However, heart rate, hemoglobin, percentage of neutrophils, CRP, gestational age, decreased bowel sounds, mechanical ventilation, tenderness, and lethargy were significantly different.

B. DL MODEL PERFORMANCE AND PARAMETER OPTIMIZATION

Manipulation of some hyperparameters produced significant changes in DL model performance, and other parameters produced reasonable models over a variety of settings. The three DL models' performances and architectures are shown in Supplemental Table SIII. As mentioned earlier, we focused on three popular attentional DL models: ResNeSt-50, SENet-154 and SE-ResNet-50. These models predicted a probability score for each image and the likelihood of an AR being detected as NEC. By comparing this probability with the cut-off threshold, we can derive a binary label showing whether the image represents NEC. An ideal model should predict all NEC samples with a probability close to 1 and all non-NEC samples with a probability close to 0. All the models achieved very promising results on the internal test dataset, with SENet-154 achieving optimal diagnostic power, with an AUC value of 0.8531 (95% CI: 0.8192-0.8870), a sensitivity of 0.6881 (95% CI: 0.6437-0.7325), a specificity of 0.7742 (95% CI: 0.7342-0.8142), a precision of 0.7394 (95% CI: 0.6974-0.7814) and an accuracy of 0.7327 (95% CI: 0.6893, 0.7760) in the internal test dataset. Therefore, SENet-154 was finally chosen as the best transferred network model for the subsequent analyses.

During the training, the Adam optimizer with initial learning rate = 0.008, beta_1 = 0.9, and beta_2 = 0.999 was used; the learning rate was stepped down in plateau fashion by a factor of 0.9 every 1 epoch. We applied weight decay with a value of 0.01 and dropout rate of 0.55.

C. RESULTS WITH ON DIFFERENT DATASETS

In the diagnosis of NEC, 18 clinical parameters were found to be significantly correlated with NEC in the univariable analysis, and the top 40 radiomics signatures with

TABLE 1. Clinical characteristics of NEC and non-NEC patients.

Characteristics	All Patients (n=827)	Non-NEC (n=485)	NEC (n=342)	P
Age, median (days)	2.00 (0, 12)	7.00 (2, 20)	0.00 (0, 6)	<0.001**
Male, no. (%)	494 (59.73)	281 (57.94)	213 (62.28)	0.24
Birth weight (kg)	2.28 (1.71, 2.80)	2.76 (2.25, 3.28)	1.98 (1.50, 2.46)	<0.001**
Gestational age (weeks)	34.5 (30.6, 38)	37.8 (34.5, 39.0)	32.30 (30, 36.1)	<0.001**
Vital signs and clinical symptoms				
Gastric residual, no. (%)	10 (12.17)	6 (12.50)	4 (1.12)	0.82
Vomiting, no. (%)	819 (99.64)	479 (99.79)	340 (99.42)	0.76
Bloody stool, no. (%)	115 (13.99)	45 (9.38)	70 (20.47)	<0.001**
Decreased bowel sounds, no. (%)	511 (62.17)	233 (48.54)	278 (81.29)	<0.001**
Abdominal distention, no. (%)	818 (99.51)	477 (99.38)	341 (99.71)	0.87
PDA, no. (%)	144 (17.52)	45 (9.38)	99 (28.95)	<0.001**
Mechanical ventilation, no. (%)	354 (43.07)	132 (27.50)	222 (64.91)	<0.001**
Tenderness, no. (%)	27 (3.28)	0 (0.00)	27 (7.89)	<0.001**
Lethargy, no. (%)	143 (17.39)	83 (17.29)	60 (17.54)	0.99
Heart rate (beats/min)	144 (133, 153)	142 (132, 151)	145 (134., 154.75)	0.0006**
SBP (mmHg)	73 (65, 80)	76 (69, 82)	71.00 (63, 79)	<0.001**
DBP (mmHg)	41 (35, 49)	43 (38, 51)	40.00 (33, 47)	<0.001**
Breath (beats/min)	44 (38, 549)	45 (40, 49)	43.00 (36, 49)	<0.001**
Temperature (°C)	36.8 (36.7, 37)	36.9 (36.8, 37)	36.80 (36.7, 37)	<0.001**
Laboratory parameters				
Hemoglobin (g/L)	125.1 (107, 149)	136 (112, 159)	121 (105, 144)	<0.001**
PLT (10^9/L)	275 (188, 388)	324 (240, 439)	249 (160, 352)	<0.001**
Percentage of neutrophil (%)	49 (36,63)	46 (33,61)	51 (38,63)	<0.001**
CRP (mg/L)	3.16 (0.5,27.8)	1.34 (0.5,9.6)	4.54 (0.5,33.23)	<0.001**
WBC (10^9/L)	10.5 (7.55,14)	10.95 (8.5,14.3)	10.00 (7,13.6)	<0.001**

Data are presented as medians (IQRs) or n (%).
P values are derived from the Mann-Whitney U test or the χ^2 test.
**Indicates a significant difference at $P < 0.01$; *Indicates a significant difference at $P < 0.05$.
Abbreviations: IQR=interquartile range; PDA = patent ductus arteriosus, SBP = systolic pressure, DBP = diastolic pressure, PLT = platelet count, CRP = C-reactive protein, WBC = white blood cell.

the highest mRMR ranks were selected. The multimodal model (cutoff: 0.4) achieved an AUC of 0.9337 (95% CI: 0.9028, 0.9646). The AR-alone model (cutoff: 0.8) achieved an AUC of 0.8757 (95% CI: 0.8347, 0.9166), and the clinical parameters-alone model (cutoff: 0.3) achieved an AUC of 0.9030 (95% CI: 0.8662, 0.9398). By adding clinical parameters to the AR-based radiomics signatures in the multimodal model, the AUC improved by 5.80%, sensitivity improved by 8.85%, specificity improved by 1.76%, precision improved by 1.05%, and accuracy improved by 7.23%.

In predicting surgical necessity, 9 clinical parameters with $p < 0.05$ in the univariable analysis were retained (Table 2), and 40 radiomics signatures were selected using the mRMR method. The multimodal model (cutoff: 0.8) achieved an AUC of 0.9413 (95% CI: 0.8998, 0.9828) by using both ARs and clinical parameters, 0.8198 (95% CI: 0.7519, 0.8877) by using ARs alone (cutoff: 0.1), and 0.9061 (95% CI: 0.8546, 0.9576) by using clinical parameters alone (cutoff: 0.5). The addition of clinical parameters to the AR-based radiomics signatures improved the AUC obtained by the multimodal model by 12.15%. The specificity, precision, and accuracy

TABLE 2. Clinical characteristics of patients with definite medical NEC or surgical NEC.

Characteristics	All Patients (n=379)	Medical NEC (n=268)	Surgical NEC (n=111)	P
Age, median (day)	0 (0, 9)	0 (0, 8)	0 (0, 10.75)	0.17
Male, no. (%)	231 (60.94)	160 (59.70)	71 (63.96)	0.51
Birth weight (kg)	2.09 (1.69, 2.51)	2.14 (1.55, 2.64)	2.07 (1.75, 2.41)	0.34
Gestational age (weeks)	34.2 (31.1, 37)	35 (31.37, 37.1)	33.75 (31, 36.1)	0.003*
Vital signs and clinical symptoms				
Gastric residual, no. (%)	4 (1.05)	3 (1.19)	1 (0.90)	0.71
Vomiting, no. (%)	377 (99.47)	266 (99.25)	111 (100.00)	0.89
Bloody stool, no. (%)	81 (21.37)	64 (23.88)	17 (15.32)	0.08
Decreased bowel sounds, no. (%)	297 (78.36)	189 (70.52)	108 (97.30)	<0.001**
Abdominal distention, no. (%)	377 (99.47)	266 (99.25)	111 (100.00)	0.89
PDA, no. (%)	108 (28.49)	81 (30.22)	27 (24.32)	0.30
Mechanical ventilation, no. (%)	221 (58.31)	116 (43.28)	105 (94.59)	<0.001**
Tenderness, no. (%)	18 (4.75)	6 (22.24)	12 (10.81)	0.0009**
Lethargy, no. (%)	72 (18.99)	59 (22.01)	13 (11.71)	0.02*
Heart rate (beats/min)	142 (132, 153)	142 (132, 150)	143 (133, 155)	0.005*
SBP (mmHg)	69.00 (63, 78)	69.00 (62, 76)	69 (64, 78)	0.11
DBP (mmHg)	39.00 (33, 46)	39.00 (32, 45)	39 (34, 48)	0.06
Breath (beats/min)	45.00 (40, 51.25)	46.00 (40, 50.75)	44.5(39, 52)	0.46
Temperature (°C)	36.80 (36.5, 36.9)	36.70 (36.5, 36.9)	36.8 (36.7, 37)	0.44
Laboratory parameters				
Hemoglobin (g/L)	134 (109.25, 163)	149 (119.5, 174.5)	120 (102, 143)	<0.001**
PLT (10^9/L)	253 (178, 342)	252 (187.5, 316.5)	253 (158, 393.5)	0.27
Percentage of neutrophil (%)	56 (42, 66)	52 (4.5, 63)	59 (45.5, 72)	0.03*
CRP (mg/L)	5.1(0.5, 31)	0.5 (0.5, 3.4)	14.15 (2.67, 53.6)	0.008*
WBC (10^9/L)	10 (6.30, 13.5)	10.2 (7.3, 13.05)	9.95 (5.8, 14.98)	0.09

Data are presented as medians (IQRs) or n (%).
P values were derived from the Mann-Whitney U test or the χ^2 test.
**Indicates a significant difference at P < 0.01; *Indicates a significant difference at P < 0.05.
Abbreviations: IQR=interquartile range; PDA = patent ductus arteriosus, SBP = systolic pressure, DBP = diastolic pressure, PLT = platelet count, CRP = C-reactive protein, WBC = white blood cell.

increased by 16.28%, 8.53%, and 4.06%, respectively (Fig. 3 and Table 3).

D. ACTIVATION MAPPING OF DL NETWORKS

To determine regions within the ARs responsible for the network predictions, we mapped the network's activation maps over the final convolutional layer (Fig. 4). This analysis allowed us to highlight the relevant regions (both lesion core and perilesional area) with the greatest impact on predictions. We observed that the network tended to fixate on the ileum, colon and jejunum. Most contributions to the predictions came in the form of large uninterrupted areas of relatively lower AR density—spanning regions within and beyond the lesion. Areas with higher AR density, however, contributed the least to the predictions. We also observed that high-density bone tissue was disregarded, as it was likely to be found in most images and was thus noninformative. This visual mapping demonstrates that attention regions within and without the lesion were both crucial for characterization and eventual prediction.

E. FEATURE IMPORTANCE

We established a multimodal model integrating 18 clinical parameters and 40 radiomics signatures. The 40 radiomics

TABLE 3. Performance comparison of multimodal data in diagnosing NEC and predicting surgical necessity.

	AUC	SE	SP	PR	ACC
Diagnose of NEC					
All features	0.9337 (0.9028, 0.9646)	0.9427 (0.9138, 0.9716)	0.8246 (0.7774, 0.8718)	0.9476 (0.9199, 0.9753)	0.9157 (0.8812, 0.9502)
ARs only	0.8757 (0.8347, 0.9166)	0.8542 (0.8104, 0.8980)	0.807 (0.7580, 0.8560)	0.9371 (0.9069, 0.9673)	0.8434 (0.7983, 0.8885)
Clinical parameters only	0.903 (0.8662, 0.9398)	0.9271 (0.8948, 0.9594)	0.7193 (0.6635, 0.7751)	0.9175 (0.8833, 0.9517)	0.8795 (0.8391, 0.9199)
Surgery prediction of NEC					
All features	0.9413 (0.8998, 0.9828)	0.85 (0.7869, 0.9131)	0.9535 (0.9163, 0.9907)	0.9714 (0.9419,1.0000)	0.8861 (0.8300, 0.9422)
ARs only	0.8198 (0.7519, 0.8877)	0.875 (0.8166, 0.9334)	0.7907 (0.7188, 0.8626)	0.8861 (0.8300, 0.9422)	0.8455 (0.7816, 0.9093)
Clinical parameters only	0.9061 (0.8546, 0.9576)	0.9 (0.8470, 0.9530)	0.7907 (0.7188, 0.8626)	0.8889 (0.8334, 0.9444)	0.8618 (0.8008.0.9228)

Note. AUC=Area under the curve; SE=Sensitivity; SP=Specificity; PR=Precision; ACC=Accuracy; 95% confidence intervals in brackets.

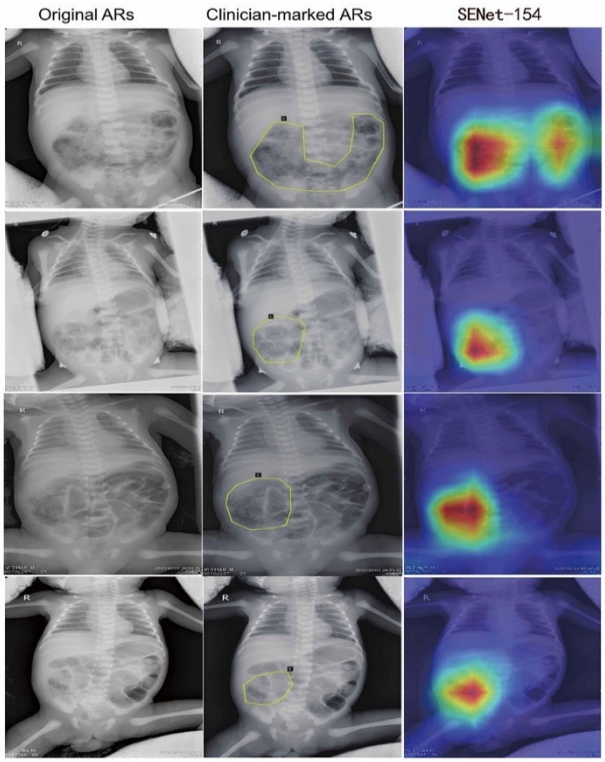

FIGURE 4. Activation mapping by the DL network. Visual highlights of the most "important" regions within the input image with the greatest contribution to maximizing the outputs of the SENet-154 final diagnosis or prediction layer are shown. Rows represent 4 randomly selected samples. The first column shows the original ARs. The second column shows annotations with different abdominal regions, labeled with serial numbers. Column 3 represents the activation heatmaps, which provide a better visual reference.

signatures were identified as the most significant diagnostic contributor, with a contribution of 28.84%. In addition, demographic parameters, such as gestational age, age, and birth weight, also contributed to the diagnosis of NEC, with a contribution of 17.30%. As expected, inflammatory markers (CRP, percentage of neutrophils, WBC, PLT, hemoglobin) contributed 25.03%. Interestingly, respiratory function (breath and mechanical ventilation) parameters were also found to be related to NEC. The reason may be that patients lacking oxygen often have poor overall health (Table SIV and Fig. 5(a)).

We also computed importance for the features involved in predicting NEC surgery based on 9 clinical parameters and 40 radiomics signatures. Radiomics signatures remained the most important contributor, but their contribution increased to 51.29%. Inflammatory markers comprising CRP, hemoglobin, percentage of neutrophils but excluding WBC and PLT parameters were also important, but their contribution decreased to 23.53%, and intestinal symptoms bloody stools and decreased bowel sounds) made no contribution. The patients diagnosed with medical NEC or surgical NEC shared similarities in the intestinal symptoms they presented (Table SV and Fig. 5(b)).

F. COMPARISON AND VALIDATION OF THE DIAGNOSTIC ABILITY OF THE AI SYSTEM

We further evaluated the strengths and weaknesses of the multimodal model with respect to clinicians utilizing confusion matrices (Fig. 6). The receiver operating characteristic curve and diagnostic and predicted performance of the multimodal model and clinicians in an external validation set are illustrated in Fig. 6 and Tables SVI and SVII. These results demonstrated that the multimodal model performed significantly better than 3 (2 years and 3–10 years of experience) clinicians and senior clinicians with 10–20 years of experience.

In diagnosing NEC, the multimodal model demonstrated a senior clinician-level performance, with an AUC of 0.98, a sensitivity of 1.00, a specificity of 0.86, and precision of 0.95. The overall accuracy of the multimodal model was 0.96, while that of the best-performing clinician, who had

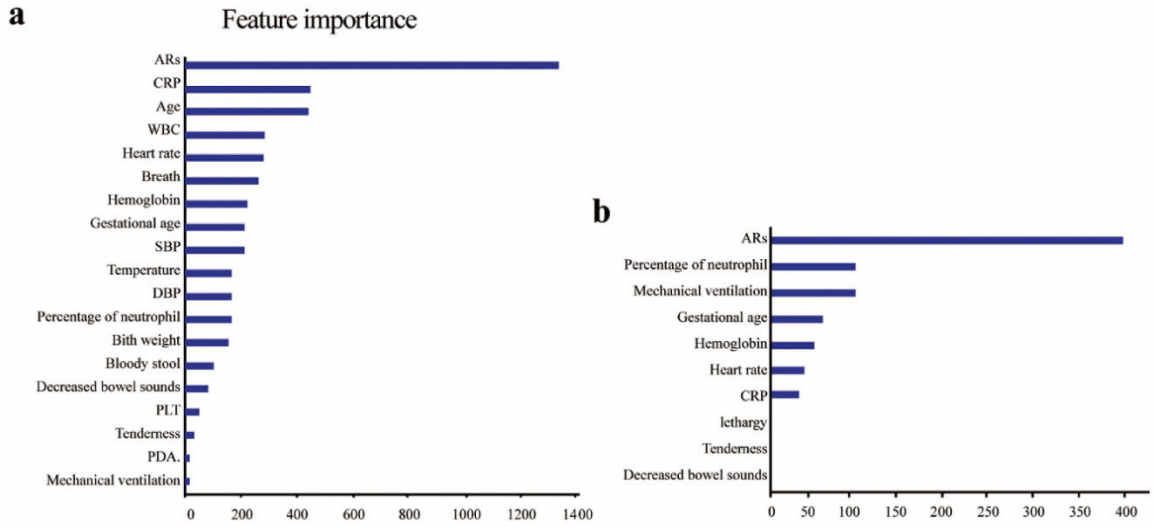

FIGURE 5. Illustration of features contributing to diagnosing NEC and predicting surgical eligibility. (a) The relative contributions of ARs and clinical parameters in identifying NEC. (b) The relative contributions of ARs and clinical parameters in predicting surgically eligible NEC. The larger the feature importance value is, the greater the importance of the variable to the model.

10–20 years of experience, was 0.86. In predicting surgical NEC, the multimodal model was superior to the two senior clinicians with 10-20 years of work experience, the two junior clinicians with 3-10 years of work experience, and the residents with 2 years. The multimodal model achieved an AUC of 0.90, a sensitivity of 0.86%, a specificity of 0.91%, a precision of 0.86%, and an accuracy of 0.89%.

IV. DISCUSSION

A. PREDICTIVE FEATURES

Our multimodal AI system proved to be accurate in the diagnosis and surgical eligibility prediction of NEC. In a recent multispecialist survey, Ahle *et al.* [24] reported 90% agreement in the value of ARs for confirming an NEC diagnosis, providing guidance in decisions on surgery and abdominal radiography as the first-choice treatment modality and the most important radiographic signs, respectively. CRP, gestational age, hemoglobin, percentage of neutrophils, and mechanical ventilation appeared in both multimodal AI systems for detecting NEC and predicting the surgical form of the disease. These five clinical parameters delineate a distinct subset of NEC in terms of surgical management. Pourcyrous *et al.* [25] concluded that CRP becomes abnormal in both stage II and stage III NEC. In infants with NEC, persistently elevated CRP after initiation of appropriate medical management suggests associated complications that may require surgical intervention. D'Angelo *et al.* [10] reported that the NEC group had a lower gestational age than the non-NEC group. Preterm newborns have many intestinal vulnerabilities that permit microbial pathogens to invade tissue and may cause bowel perforation and even death. Cai *et al.* [26] analysis revealed that a decrease in hemoglobin was a risk factor for NEC. Here, we found that an increased percentage of neutrophils was correlated with NEC and surgical NEC, and increased mechanical ventilation was related to NEC, while decreased mechanical ventilation was associated with surgical NEC. D'Angelo *et al.* [10] also showed that several biochemical alterations, such as raised or depressed WBCs and thrombocytopenia, can be observed in infants affected by NEC.

B. COMPARISONS WITH PREVIOUS MODELS

Many studies have described several potentially useful biomarkers, isolated largely from serum, stool, and urine samples, for discriminating NEC and surgical NEC. Cakir *et al.* [27] estimated the predictive value of endocan and interleukin (IL)-33 biomarkers using statistical analyses. They found that the serum levels of these biomarkers were significantly higher in the NEC group than in the control group on the 1st, 3rd, and 7th days. Serum endocan and IL-33 levels gradually increased in patients who underwent surgery. Heath *et al.* [14] also studied biomarkers in NEC and found that high amounts of intestinal alkaline phosphatase (IAP) in stool and low IAP enzyme activity were associated with a diagnosis of NEC and may serve as useful biomarkers for the disease. Ng *et al.* [28] discovered that specific circulating miR-1290 biomarkers provide the greatest diagnostic usefulness for identifying both mild medical and severe surgical NEC. Combined with C-reactive protein, these biomarkers achieved a sensitivity of 0.83, a specificity of 0.96, a positive predictive value of 0.75, and a negative predictive value of 0.98. Ng *et al.* [29] combined gut barrier proteins, liver fatty acid-binding protein (L-FABP), intestinal fatty acid-binding protein (I-FABP), and trefoil factor 3 (TFF3) biomarkers and the LIT score to differentiate NEC and identify the most severely affected surgical NEC. Median values of the biomarkers and the LIT score in the NEC group served as cutoff values for identifying NEC and achieved a specificity

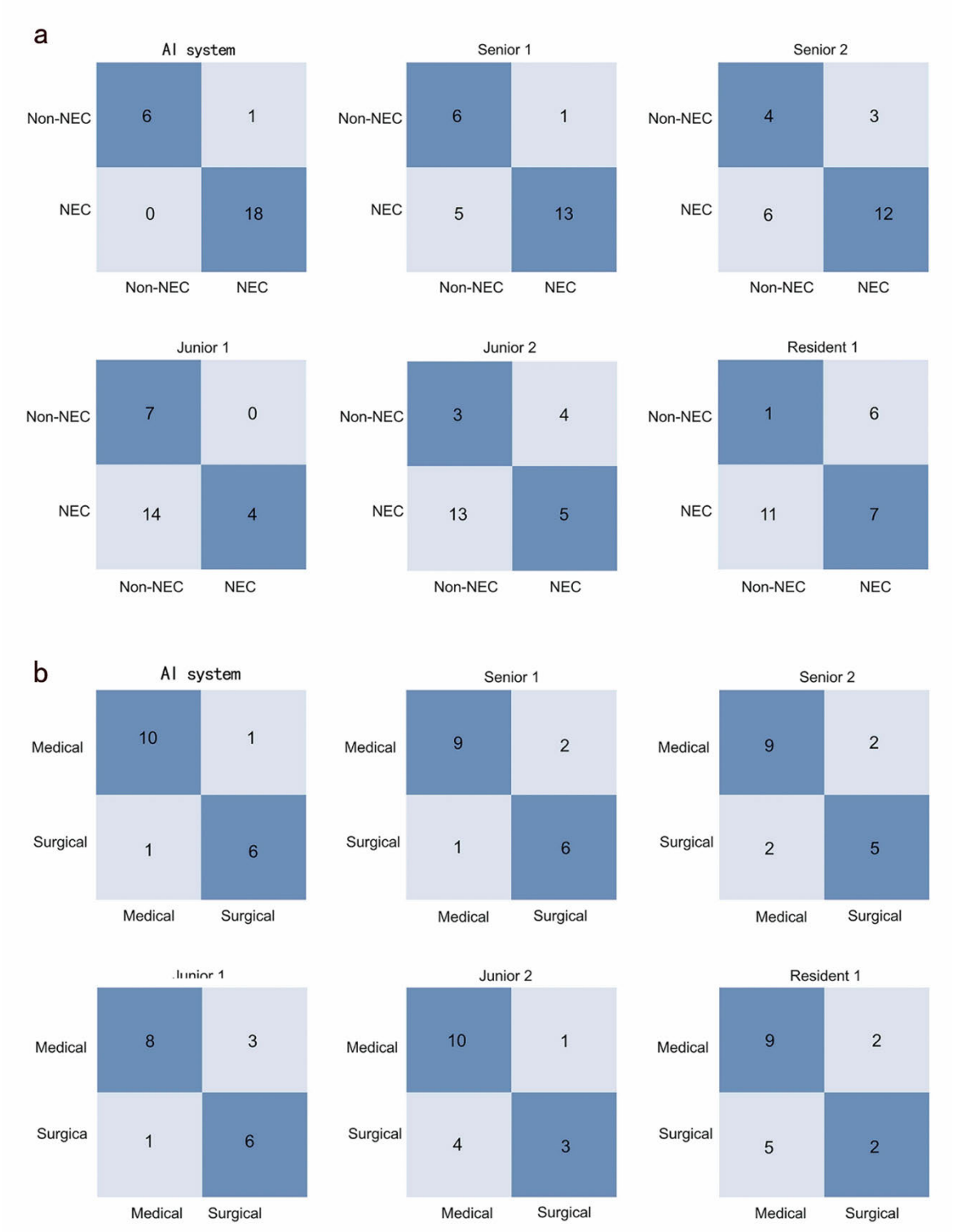

FIGURE 6. Confusion matrices for diagnosing NEC and predicting surgical NEC for the AI system and individual clinicians.

of 95% or more and a sensitivity of 50%. The median LIT score of 4.5 achieved a sensitivity and specificity of 83% and 100%, respectively, in predicting NEC. Although biomarkers are powerful for detecting NEC and predicting surgical NEC, their analysis is expensive and time consuming. No biomarkers have been identified to prospectively diagnose NEC and predict surgical NEC. Therefore, it is challenging to establish the biological significance of biomarkers.

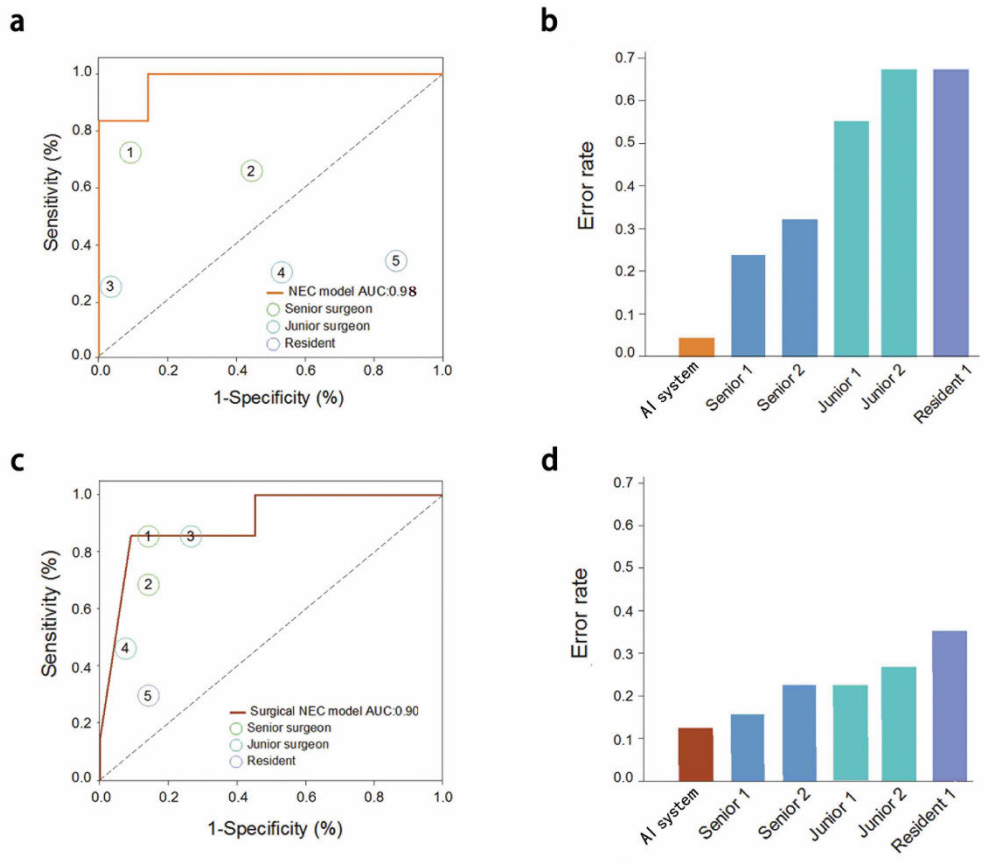

FIGURE 7. Comparison between the AI system and five clinicians of different experience levels in binary classifying NEC versus non-NEC (a, b) and surgical NEC versus medical NEC (c, d).

Previous studies have applied ML models to aid in the diagnosis of NEC. Mueller *et al.* [30] developed an algorithm using artificial neural networks (ANNs) to predict prematurely born infants at the highest risk of NEC. Small gestational age (SGA) and artificial ventilation were the first and second most useful among all 57 variables included for the ANN. As predictive tools, the ANNs provided an indication for the relative importance of the 57 variables in the final decision-making. Ntonfo *et al.* [31] presented a novel approach to the early diagnosis of NEC through thermal image analysis. Preliminary results showed that IQR and kurtosis measures were good discriminants in the detection of NEC.

Despite ARs or clinical parameters being the primary evidence for the disease and NEC diagnoses requiring all-sided information, there are no existing studies that have combined ARs and clinical parameters as two modalities. Unlike traditional methods, the proposed multimodal AI system combines different data modalities with the LightGBM classifier. The LightGBM classifier is based on decision tree algorithms and is used for variable ranking and classification. From the combined information, the LightGBM classifier not only quantified the feature importance of the clinical parameters part and the 40 selected radiomics signatures but also classified and determined the surgical eligibility of NEC.

C. MODEL INTERPRETATION

One disadvantage of DL is that the model usually runs as a black box. However, it is necessary for clinicians to understand the reasons why a model makes such a prediction in the clinic, especially when timely detection is necessary. Grad-CAM was used to produce the attention map highlighting the important regions in the ARs for diagnosing the target object (NEC or non-NEC) and histopathologic features.

D. LIMITATION

The multimodal AI system should be further improved by considering genetic information, microbiome data, and the numerous altered biochemical parameters missing in this study. Indeed, Hooven *et al.* [7] used clinical and microbiome data and achieved an AUC above 0.90, with 75% of dominant predictive samples for NEC-affected infants identified at least 24 hours prior to disease onset. Due to the low number of eligible patients and the small number of clinical parameters, the AI system might not have yet achieved the best performance, especially in distinguishing medical NEC and surgical NEC. For the AI system to perform the real-time

diagnosis of NEC in a clinic, it needs to be evaluated in the form of prospective trials. Should this happen, it would likely lead to improved outcomes. Due to the limitation of the ARs, the selected radiomics signatures may not reflect all cases completely. Further studies could involve the use of images from different manufacturers.

V. CONCLUSION

In this paper, we identified the significant features of ARs and clinical data that were closely related to the diagnosis and surgical prediction of NEC with feature engineering using artificial intelligence. Then, a multimodal AI system was established with ML and DL models in series. The AI system was tested on a dataset derived from patients from Guangzhou Women and Children's Medical Center and ultimately demonstrated favorable accuracy in diagnosing NEC and predicting surgical NEC. After validation, the multimodal AI system proved to be a useful auxiliary diagnostic tool for helping clinicians improve their efficiency and accuracy. Future work should entail the determination of characteristic factors to improve the accuracy of the AI system and supplement a prospective randomized case-control study on the treatment of NEC.

ACKNOWLEDGMENT

The authors wish to acknowledge Dr. Yongke Cao for his linguistic assistance during the preparation of this manuscript.

DECLARATIONS

A. CONFLICT OF INTEREST STATEMENT

The authors have no conflicts of interest relevant to this article to disclose.

B. AVAILABILITY OF DATA AND MATERIAL (DATA TRANSPARENCY)

The data are available through the corresponding author.

C. CODE AVAILABILITY (SOFTWARE APPLICATION OR CUSTOM CODE)

The custom code can be obtained through the corresponding author.

REFERENCES

[1] D. F. Niño, C. P. Sodhi, and D. J. Hackam, "Necrotizing enterocolitis: New insights into pathogenesis and mechanisms," *Nature Rev. Gastroenterol. Hepatol.*, vol. 13, no. 10, pp. 590–600, Oct. 2016, doi: 10.1038/nrgastro.2016.119.

[2] J. van Druten, M. S. Sharif, S. S. Chan, C. Chong, and H. Abdalla, "A deep learning based suggested model to detect necrotising enterocolitis in abdominal radiography images," in *Proc. Int. Conf. Comput., Electron. Commun. Eng. (iCCECE)*, London, U.K., Aug. 2019, pp. 118–123.

[3] J. van Druten, M. S. Sharif, M. Khashu, and H. Abdalla, "A proposed machine learning based collective disease model to enable predictive diagnostics in necrotising enterocolitis," in *Proc. Int. Conf. Comput., Electron. Commun. Eng. (iCCECE)*, Southend, U.K., Aug. 2018, pp. 101–106.

[4] Y. Shi, P. Payeur, M. Frize, and E. Bariciak, "Thermal and RGB-D imaging for necrotizing enterocolitis detection," in *Proc. IEEE Int. Symp. Med. Meas. Appl. (MeMeA)*, Bari, Italy, Jun. 2020, pp. 1–6.

[5] B. L. Frost, B. P. Modi, T. Jaksic, and M. S. Caplan, "New medical and surgical insights into neonatal necrotizing enterocolitis: A review," *JAMA Pediatrics*, vol. 171, no. 1, pp. 83–88, Jan. 2017, doi: 10.1001/jamapediatrics.2016.2708.

[6] R. M. Patel, J. Ferguson, S. J. McElroy, M. Khashu, and M. S. Caplan, "Defining necrotizing enterocolitis: Current difficulties and future opportunities," *Pediatric Res.*, vol. 88, no. S1, pp. 10–15, Aug. 2020, doi: 10.1038/s41390-020-1074-4.

[7] T. Hooven, Y. C. Lin, and A. Salleb-Aouissi, "Multiple instance learning for predicting necrotizing enterocolitis in premature infants using microbiome data," in *Proc. ACM Conf. Health, Inference, Learn.*, Toronto, ON, Canada, Apr. 2020, pp. 99–109.

[8] J. van Druten, M. Khashu, S. S. Chan, S. Sharif, and H. Abdalla, "Abdominal ultrasound should become part of standard care for early diagnosis and management of necrotising enterocolitis: A narrative review," *Arch. Disease Childhood-Fetal Neonatal Ed.*, vol. 104, no. 5, pp. F551–F559, Sep. 2019, doi: 10.1136/archdischild-2018-316263.

[9] K. Zhang *et al.*, "Clinically applicable AI system for accurate diagnosis, quantitative measurements, and prognosis of COVID-19 pneumonia using computed tomography," *Cell*, vol. 181, no. 6, pp. 1423.e11–1433.e11, Jun. 2020, doi: 10.1016/j.cell.2020.04.045.

[10] G. D'Angelo, P. Impellizzeri, L. Marseglia, A. S. Montalto, T. Russo, I. Salamone, R. Falsaperla, G. Corsello, C. Romeo, and E. Gitto, "Current status of laboratory and imaging diagnosis of neonatal necrotizing enterocolitis," *Italian J. Pediatrics*, vol. 44, no. 1, p. 84, Jul. 2018, doi: 10.1186/s13052-018-0528-3.

[11] M. J. Bell *et al.*, "Neonatal necrotizing enterocolitis. Therapeutic decisions based upon clinical staging," *Ann. Surg.*, vol. 187, no. 1, pp. 1–7, Jan. 1978, doi: 10.1097/00000658-197801000-00001.

[12] M. C. Walsh and R. M. Kliegman, "Necrotizing enterocolitis: Treatment based on staging criteria," *Pediatric Clinics North Amer.*, vol. 33, no. 1, pp. 179–201, Feb. 1986, doi: 10.1016/s0031-3955(16)34975-6.

[13] J. Neu and W. A. Walker, "Necrotizing enterocolitis," *New England J. Med.*, vol. 364, no. 3, pp. 255–264, Jan. 2011, doi: 10.1056/NEJMra1005408.

[14] M. Heath, R. Buckley, Z. Gerber, P. Davis, L. Linneman, Q. Gong, B. Barkemeyer, Z. Fang, M. Good, D. Penn, and S. Kim, "Association of intestinal alkaline phosphatase with necrotizing enterocolitis among premature infants," *JAMA Netw. Open*, vol. 2, no. 11, Nov. 2019, Art. no. e1914996, doi: 10.1001/jamanetworkopen.2019.14996.

[15] C. Irles, G. González-Pérez, S. C. Muiños, C. M. Macias, C. S. Gómez, A. Martínez-Zepeda, G. C. González, and E. L. Servitje, "Estimation of neonatal intestinal perforation associated with necrotizing enterocolitis by machine learning reveals new key factors," *Int. J. Environ. Res. Public Health*, vol. 15, no. 11, p. 2509, Nov. 2018, doi: 10.3390/ijerph15112509.

[16] C. Battersby, N. Longford, K. Costeloe, and N. Modi, "Development of a gestational age–specific case definition for neonatal necrotizing enterocolitis," *JAMA Pediatrics*, vol. 171, no. 3, pp. 256–263, Mar. 2017, doi: 10.1001/jamapediatrics.2016.3633.

[17] B. Wingfield, S. Coleman, T. M. McGinnity, and A. J. Bjourson, "A metagenomic hybrid classifier for paediatric inflammatory bowel disease," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Vancouver, BC, Canada, Jul. 2016, pp. 1083–1089.

[18] N. Barda, D. Riesel, A. Akriv, J. Levy, U. Finkel, G. Yona, D. Greenfeld, S. Sheiba, J. Somer, E. Bachmat, G. N. Rothblum, U. Shalit, D. Netzer, R. Balicer, and N. Dagan, "Developing a COVID-19 mortality risk prediction model when individual-level data are not available," *Nature Commun.*, vol. 11, no. 1, pp. 1–9, Sep. 2020, doi: 10.1038/s41467-020-18297-9.

[19] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, H. Lin, Z. Zhang, Y. Sun, T. He, J. Mueller, R. Manmatha, M. Li, and A. Smola, "ResNeSt: Split-attention networks," 2020, *arXiv:2004.08955*. [Online]. Available: http://arxiv.org/abs/2004.08955

[20] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 7132–7141.

[21] N. De Jay, S. Papillon-Cavanagh, C. Olsen, N. El-Hachem, G. Bontempi, and B. Haibe-Kains, "MRMRe: An R package for parallelized mRMR ensemble feature selection," *Bioinformatics*, vol. 29, no. 18, pp. 2365–2368, Sep. 2013, doi: 10.1093/bioinformatics/btt383.

[22] G. Ke *et al.*, "LightGBM: A highly efficient gradient boosting decision tree," in *Proc. Adv. Neural Inf. Process. Syst. (NIP)*, Long Beach, CA, USA, 2017, pp. 3146–3154.

[23] G. Montavon, W. Samek, and K.-R. Müller, "Methods for interpreting and understanding deep neural networks," *Digit. Signal Process.*, vol. 73, pp. 1–15, Feb. 2018, doi: 10.1016/j.dsp.2017.10.011.
[24] M. Ahle, H. G. Ringertz, and E. Rubesova, "The role of imaging in the management of necrotising enterocolitis: A multispecialist survey and a review of the literature," *Eur. Radiol.*, vol. 28, no. 9, pp. 3621–3631, Sep. 2018, doi: 10.1007/s00330-018-5362-x.
[25] M. Pourcyrous, "C-reactive protein in the diagnosis, management, and prognosis of neonatal necrotizing enterocolitis," *Pediatrics*, vol. 116, no. 5, pp. 1064–1069, Nov. 2005, doi: 10.1542/peds.2004-1806.
[26] N. Cai, W. Fan, M. Tao, and W. Liao, "A significant decrease in hemoglobin concentrations may predict occurrence of necrotizing enterocolitis in preterm infants with late-onset sepsis," *J. Int. Med. Res.*, vol. 48, no. 9, Sep. 2020, Art. no. 0300060520952275, doi: 10.1177/0300060520952275.
[27] U. Cakir, C. Tayman, E. Yarci, H. Halil, M. Buyuktiryaki, H. O. Ulu, C. Yucel, and S. S. Oguz, "Novel useful markers for follow-up of necrotizing enterocolitis: Endocan and interleukin-33," *J. Maternal-Fetal Neonatal Med.*, vol. 33, no. 14, pp. 2333–2341, Jul. 2020, doi: 10.1080/14767058.2018.1548601.
[28] P. C. Ng *et al.*, "Plasma miR-1290 is a novel and specific biomarker for early diagnosis of necrotizing enterocolitis—Biomarker discovery with prospective cohort evaluation," *J. Pediatrics*, vol. 205, pp. 83.e10–90.e10, Feb. 2019, doi: 10.1016/j.jpeds.2018.09.031.
[29] E. W. Y. Ng, T. C. W. Poon, H. S. Lam, H. M. Cheung, T. P. Y. Ma, K. Y. Y. Chan, R. P. O. Wong, K. T. Leung, M. M. T. Lam, K. Li, and P. C. Ng, "Gut-associated biomarkers L-FABP, I-FABP, and TFF3 and LIT score for diagnosis of surgical necrotizing enterocolitis in preterm infants," *Ann. Surgery*, vol. 258, no. 6, pp. 1111–1118, Dec. 2013, doi: 10.1097/SLA.0b013e318288ea96.
[30] M. Mueller, S. N. Taylor, C. L. Wagner, and J. S. Almeida, "Using an artificial neural network to predict necrotizing enterocolitis in premature infants," in *Proc. Int. Joint Conf. Neural Netw.*, Atlanta, Georgia, Jun. 2009, pp. 2172–2175.
[31] G. M. K. Ntonfo, M. Frize, and E. Bariciak, "Detection of necrotizing enterocolitis in newborns using abdominal thermal signature analysis," in *Proc. IEEE Int. Symp. Med. Meas. Appl. (MeMeA)*, Turin, Italy, May 2015, pp. 36–39.

WENJING GAO received the M.Sc. degree from Northwest Normal University, in 2018. She is currently an Image Algorithm Engineer with the China Guangzhou Women and Children's Medical Center. Her research interests include pattern recognition, machine learning, and deep learning.

YUANYUAN PEI received the Ph.D. degree in medicine from Sun Yat-sen University, in 2009. She has been engaged in Research and Development in the biomedical engineering field for ten years. Since 2020, she has been a Clinical Researcher with the Data Center of Guangzhou Women and Children's Medical Center. Her current research interest includes the intelligent application of health care big data.

HUIYING LIANG received the Ph.D. degree from Southern Medical University. He is currently the Director with the Data Center of Guangzhou Women and Children's Medical Center. His main research interest includes the intelligent application of health care big data.

JUNJIAN LV graduated from Sun Yat-sen University, in 2005. He received the qualification of the Associate Chief Physician in 2019. For 15 years, he has been engaged in the diagnosis and treatment of neonatal diseases and he has been responsible for the perioperative management of neonatal surgical diseases, since 2015. His current research interests include NEC, CDH, and ECMO.

JIALE CHEN received the master's degree in pediatric surgery from West China Hospital, Sichuan University. He has eight years of experience in neonatal disease research.

WEI ZHONG received surgical training from the Boston Children's Hospital, America, from October to December 2016. She is currently the Chief Physician and a Master Supervisor of Pediatric Surgery with the Guangzhou Women and Children's Medical Center. She has been engaged in first-line clinical work in pediatric surgery for 25 years, specializing in neonatal diseases and perioperative management. From September 2003 to February 2005, she worked as a General Surgeon with The Royal Children's Hospital, Melbourne, Australia. From April to June 2008, she was with the Cincinnati Children's Hospital Fetal Medical Center for further study.