Computer Methods and Programs in Biomedicine 200 (2021) 105937

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

Applying a random projection algorithm to optimize machine learning model for predicting peritoneal metastasis in gastric cancer patients using CT images

Seyedehnafiseh Mirniaharikandehei[a,*], Morteza Heidari[a], Gopichandh Danala[a], Sivaramakrishnan Lakshmivarahan[b], Bin Zheng[a]

[a] School of Electrical and Computer Engineering, University of Oklahoma, Norman, OK 73019, USA
[b] School of Computer Sciences, University of Oklahoma, Norman, OK 73019, USA

ARTICLE INFO

Article history:
Received 25 August 2020
Accepted 6 January 2021

Keywords:
Gastric cancer
Quantitative features
Computed tomography
Random projection
Feature dimensionality reduction

ABSTRACT

Background and Objective: Non-invasively predicting the risk of cancer metastasis before surgery can play an essential role in determining which patients can benefit from neoadjuvant chemotherapy. This study aims to investigate and test the advantages of applying a random projection algorithm to develop and optimize a radiomics-based machine learning model to predict peritoneal metastasis in gastric cancer patients using a small and imbalanced computed tomography (CT) image dataset.

Methods: A retrospective dataset involving CT images acquired from 159 patients is assembled, including 121 and 38 cases with and without peritoneal metastasis, respectively. A computer-aided detection scheme is first applied to segment primary gastric tumor volumes and initially compute 315 image features. Then, five gradients boosting machine (GBM) models embedded with five feature selection methods (including random projection algorithm, principal component analysis, least absolute shrinkage, and selection operator, maximum relevance and minimum redundancy, and recursive feature elimination) along with a synthetic minority oversampling technique, are built to predict the risk of peritoneal metastasis. All GBM models are trained and tested using a leave-one-case-out cross-validation method.

Results: Results show that the GBM model embedded with a random projection algorithm yields a significantly higher prediction accuracy (71.2%) than the other four GBM models ($p<0.05$). The precision, sensitivity, and specificity of this optimal GBM model are 65.78%, 43.10%, and 87.12%, respectively.

Conclusions: This study demonstrates that CT images of the primary gastric tumors contain discriminatory information to predict the risk of peritoneal metastasis, and a random projection algorithm is a promising method to generate optimal feature vector, improving the performance of machine learning based prediction models.

1. Introduction

Although the occurrence of gastric cancer has declined recently, it remains the third leading cause of cancer-related death worldwide [1]. While surgery remains the only curative treatment option, preoperative neoadjuvant chemotherapy (NAC) has demonstrated favorable results with increased therapeutic resection rates and improved survival [2]. Preventing the adverse effect of NAC, patients with different disease stages must be distinguished from each other [3] because, for each step of the disease, the treatment would be different [4]. Recent studies demonstrated that applying preoperative NAC for advanced gastric cancer patients with peritoneal metastasis (PM) yielded a much better clinical outcome and enhanced the overall survival rate [5,6]. Thus, an accurate assessment of the presence of the PM is essential for the selection of appropriate patients for NAC. Since the overall accuracies of subjectively reading endoscopic ultrasound and computed tomography (CT) images are not completely reliable [3,4], an alternative technique is needed to facilitate the assessment of tumor stages and the risk of PM.

Recently, the novel radiomics technique has been applied to extract quantitative information from medical images with a large pool of image features, and the data mining of image feature pool offers an exciting approach to build machine learning (ML) mod-

[*] Corresponding author.
E-mail address: snmirnia@ou.edu (S. Mirniaharikandehei).

https://doi.org/10.1016/j.cmpb.2021.105937
0169-2607/© 2021 Elsevier B.V. All rights reserved.

Table 1
Distribution of demographic information and several related clinical results of study cases in the dataset.

	Category	Cases with PM	Cases without PM
Total Cases		121	38
Age (years old)	< 45	11 (6.9%)	5 (3.1%)
	45 – 65	72 (45.2%)	23 (14.4%)
	> 65	38 (23.8%)	10 (6.2%)
	Mean ± SD	59.49 ± 11.97	59.11 ± 8.75
	Median	61	60
Gender	Men	94 (59.1%)	(18.8%)
	Women	27 (16.9%)	8 (5.0%)
Tumor Location	Upper	37 (23.2%)	19 (11.9%)
	Medium	20 (12.6%)	7 (4.4%)
	Lower	50 (31.4%)	12 (7.5%)
	Diffuse	14 (8.8%)	0
Pathological Staging after Surgery	I	0	38 (23.9%)
	II	26 (16.4%)	0
	III	32 (20.1%)	0
	IV	63 (39.6%)	0
Bormann Type	1	1 (0.6%)	0
	2	21 (13.2%)	11 (6.9%)
	3	94 (59.1%)	25 (15.7%)
	4	5 (3.1%)	2 (1.3%)

els and predict clinical outcomes [7,8]. Although several radiomics based ML models have been reported to differentiate and stage gastric cancer patients [9,10], these studies computed radiomics features from the tumor region that is manually segmented from one CT slice selected by the radiologist. Meanwhile, the correlation analysis based method was used to determine a small set of image features, which cannot eliminate the redundancy of the selected features. Thus, discriminatory power and prediction accuracy of these ML models were limited. To overcome such limitations, we in this study propose to develop and evaluate a new computer-aided detection (CAD) scheme aiming to predict the risk of PM among gastric cancer patients. First, our scheme segments primary gastric tumor volume in 3D CT image data, which can better compute image features related to the heterogeneity of the tumors. Second, to reduce the dimensionality of feature space and better identify orthogonal or non-redundant image features from a large pool of initially computed radiomics features, we investigate and apply a random projection algorithm (RPA). Third, to avoid bias in generating feature vector, RPA is embedded in a multi-feature fusion-based machine learning (ML) model to predict the risk of PM, which is trained and tested using (1) a synthetic minority oversampling technique (SMOTE) to balance numbers of cases in two classes and (2) a leave-one-case-out (LOCO) cross-validation method. The details of the study design, experimental procedures, data analysis results, and discussions are presented in the following sections of this article.

2. Materials and methods

2.1. Image dataset

In this study, we use a retrospective dataset of abdominal computed tomography (CT) images. To avoid potential case selection bias, the dataset initially contains 219 consecutive patients who were diagnosed and treated with gastric cancer. Then, by excluding the cases that were unresectable or undetectable based on CT examinations and poor image quality as determined by the radiologists in the retrospective review, 159 cases are included in this study dataset. Among these patients, 121 cases have PM, and 38 cases do not have PM. Table 1 summarizes the distribution of general demographic information and several related clinical results of these 159 patients in this dataset.

Each patient had an abdominal CT imaging examination during the original cancer diagnosis before surgery. All CT examinations were performed using a multidetector CT machine (GE Discovery CT750 HD, GE Healthcare). Each patient is requested to fast from food overnight and drank 600-1000ml water orally to distend the stomach prior to the CT examination. The contrast-enhanced CT images are obtained with a delay of 28 s (arterial phase), 55 s (portal phase), and 120 s (venous phase) after administration of infused 1.5 ml/kg body weight iodinated contrast agent (Optiray 320 mg I/mL, Bayer Schering Pharma) intravenously at a flow rate of 2.5 ml/s. The CT scanning parameters include (1) tube voltage switching between 120 kVp and 140 KVp in spectral imaging mode, (2) tube current automatically optimizing with the maximum limit of 200 mA, (3) tube rotation time of 0.76 – 0.80 s, (4) detector collimation of 64 × 0.625 mm, (5) field of view with 350 – 500 mm, and (6) the image matrix with 512 × 512 pixels and reconstruction thickness of 2.5 mm. The venous phase CT images were selected and used to segment tumors, compute image features, and build the machine learning prediction model in this study.

2.2. Tumor segmentation

By recognizing the heterogeneity of tumors in the clinical images, we modified and implemented a hybrid tumor segmentation scheme that used a dynamic programming method [11,12] to adaptively identify growing thresholds of a multi-layer topographic region growing algorithm and initial contour in active contour algorithm. Specifically, the tumor segmentation scheme involves the following steps. First, a Weiner filter is applied to reduce image noise. Second, an initial seed is placed at the center of the tumor region of one CT slice in which the tumor has its most significant area. To reduce inter-operator variability in choosing the initial seed and increase the robustness of segmentation results demonstrated in the previous study [13], a predefined window with the size of (5,5) around the initial seed is automatically created. A pixel with the minimum value inside the window is detected and selected as the first seed point. Third, to automatically determine the first threshold value for the region growing algorithm, a new predefined window with size of (5,5), which ensures to fully locate inside all tumor regions of our dataset and avoid potential risk of growing leakage at the first growth layer, is created around the new seed point. Then, the scheme computes the pixel value differ-

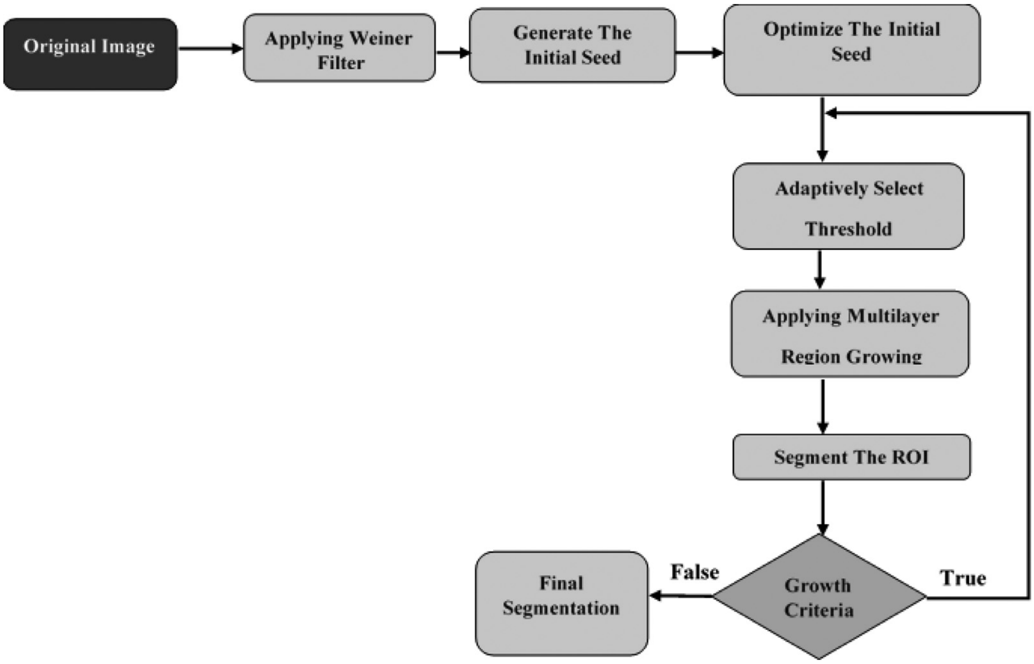

Original Image
Applying Weiner Filter
Generate The Initial Seed
Optimize The Initial Seed
Adaptively Select Threshold
Applying Multilayer Region Growing
Segment The ROI
Growth Criteria
True
False
Final Segmentation

Fig. 1. The block diagram of the 2D tumor region segmentation.

ences between the center pixel and boundary pixels and identifies the maximum difference. Subsequently, the region growing threshold is determined as $T_1 = V_c + 0.25 \times D_{max}$, where V_c is the pixel value of the center pixel and D_{max} is the computed maximum pixel value difference inside the bounding window. This threshold value is applied to define the first layer of region growing to segment tumor region depicting on one CT image slice.

Fourth, after determining the first layer of tumor region growth, the growing threshold of the second layer is $T_2 = T_1 + \beta C_1$ where C_1 is the computed contrast of the first layer, and β is a coefficient (i.e., 0.5). This multi-layer region growing continues until the growth ratio between two adjacent layers is two times bigger than the size of the last growing layer. Last, after the region growing algorithm stops, the scheme selects the boundary contour of the last region growing layer as the initial region contour. The active contour algorithm is followed to expand or shrink the contour curve for the best fitting tumor boundary. Fig. 1 and Fig. 2 illustrate the block diagram of this tumor segmentation scheme and an image example of applying the above steps to segment a tumor region depicting on one CT slice, respectively.

Subsequently, after segmenting the tumor region on one CT slice, the CAD scheme continues to perform tumor region segmentation by scanning in both up and down directions until no tumor region is detected in the next adjacent CT slice. In this process, the central point of the tumor region detected in the adjacent CT slice is mapped into the new CT slice as the initial region growing seed. Then, the tumor region segmentation in this targeted slice is automatically performed from the mapped growing seed. Additionally, a tumor growing boundary condition is limited by the adjacent slice to facilitate the multi-layer region-growing process and avoid growth leakage. Fig. 3 shows an example of the segmentation of tumor regions depicting several CT image slices of one case. In this way, 3D tumor volume can be segmented and computed.

2.3. Feature extraction

Once 3D tumor volume is segmented, the CAD scheme is applied to compute a large set of radiomics-based image features, which include 315 features extracted and computed from each segmented 2D tumor region (ROI) depicting on one CT image slice. These features were categorized into four main groups, including (a) the grayscale-run length (GLRLM) features in which 44 2-dimensional features are extracted. (b) The Gray Level Difference Methods (GLDM) probability density function features in which from each probability density function representing statistical texture features of ROI, four features of mean, median, standard deviation, and variance are computed. (c) Wavelet domain features in which the image is first decomposed into four components comprising low and high scale decomposition in either X or Y direction by wavelet transform [14]. Then, the GLCM features [15], as well as 21 tumor density [16] and GLDM features [17], are extracted from those components. (d) Laplacian of Gaussian (LoG) features in which a Gaussian smoothing filter is first applied to reduce the sensitivity to the noise, and then the Laplacian filter sharpens the image's edge and highlights rapid intensity changes inside the region [18]. Next, from the extracted points after applying the LoG filters, the mean, median, and standard deviation are computed. Fig. 4 shows the flow diagram of the feature extraction process.

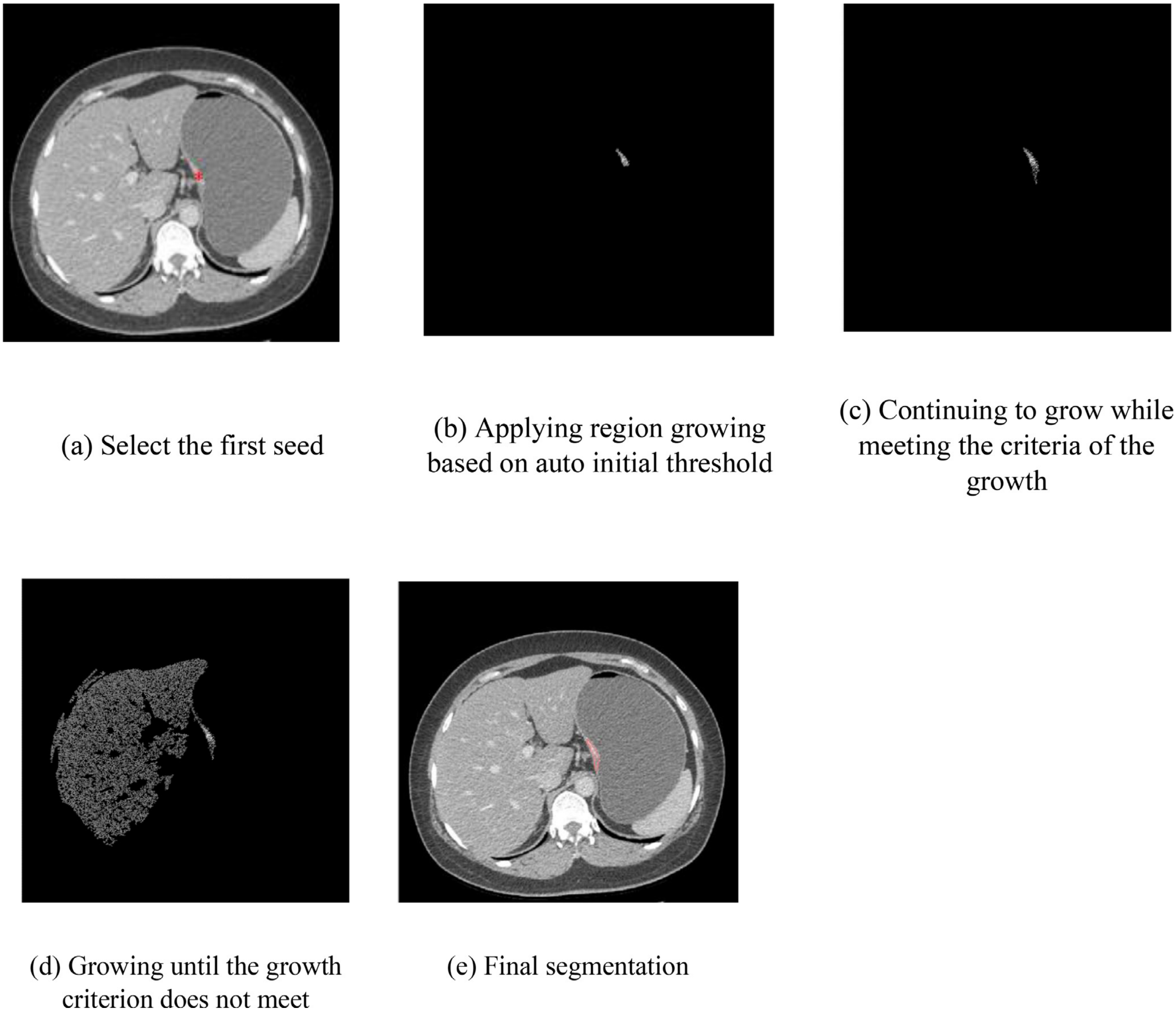
After computing 2D features of all segmented tumor regions in N involved CT image slice, CAD scheme computes each 3D feature (F_{3D}^k) as

$$F_{3D}^k = \sum_{i=1}^{N} w_i \times F_{2D}^k \tag{1}$$

where w_i is the ratio of the segmented tumor volume on a ith slice to the whole tumor volume segmented on all N involved CT slices. The segmented tumor volume on a i th slice is computed by multiplying the segmented region size (2D) to the CT slice thickness. Finally, all 315 computed 3D feature values are normalized between 0 to 1 to reduce case-based reliance and weight all features evenly.

2.4. Feature dimensionality reduction using random projection algorithm

Since the initial feature pool contains 315 image features, many of them can be redundant (highly correlated) or irrelevant (with



(a) Select the first seed

(b) Applying region growing based on auto initial threshold

(c) Continuing to grow while meeting the criteria of the growth

(d) Growing until the growth criterion does not meet

(e) Final segmentation

Fig. 2. The process of 2D tumor region segmentation.

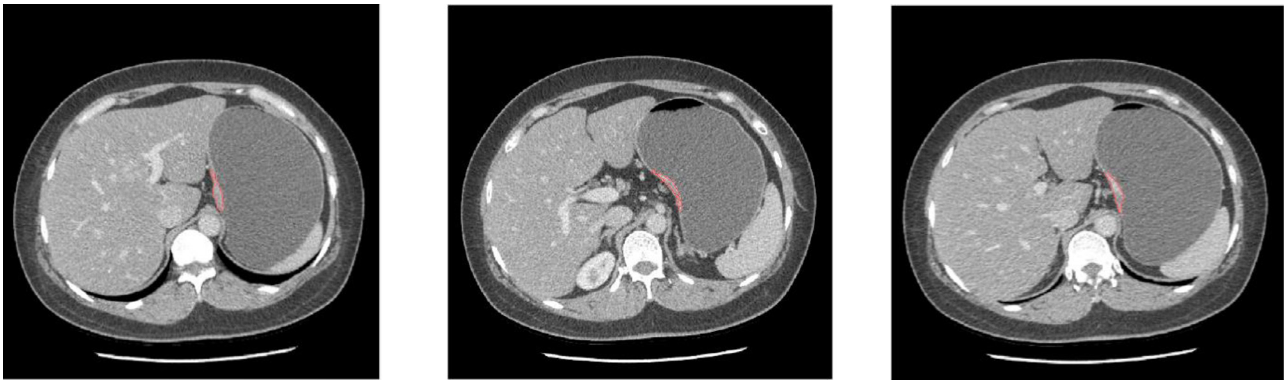


Fig. 3. An Example of 3D segmentation of a lesion in 3 different slices.

lower performance). Hence, selecting a small set of optimal features to reduce the feature dimension and enhance learning accuracy is vital. In this study, we investigate and apply a novel image feature regeneration method of the Random Projection Algorithm (RPA). Theoretic analysis has indicated that the RPA has advantages for its simplicity, high performance, and robustness compared to other feature reduction methods; however, empirical results are sparse [19]. Meanwhile, RPA has been investigated and tested in many engineering applications such as text [20] and face and object recognition [21] and yielded comparable results to conventional feature regeneration methods like principal component analysis (PCA) [22]. Nevertheless, the advantage of employing RP

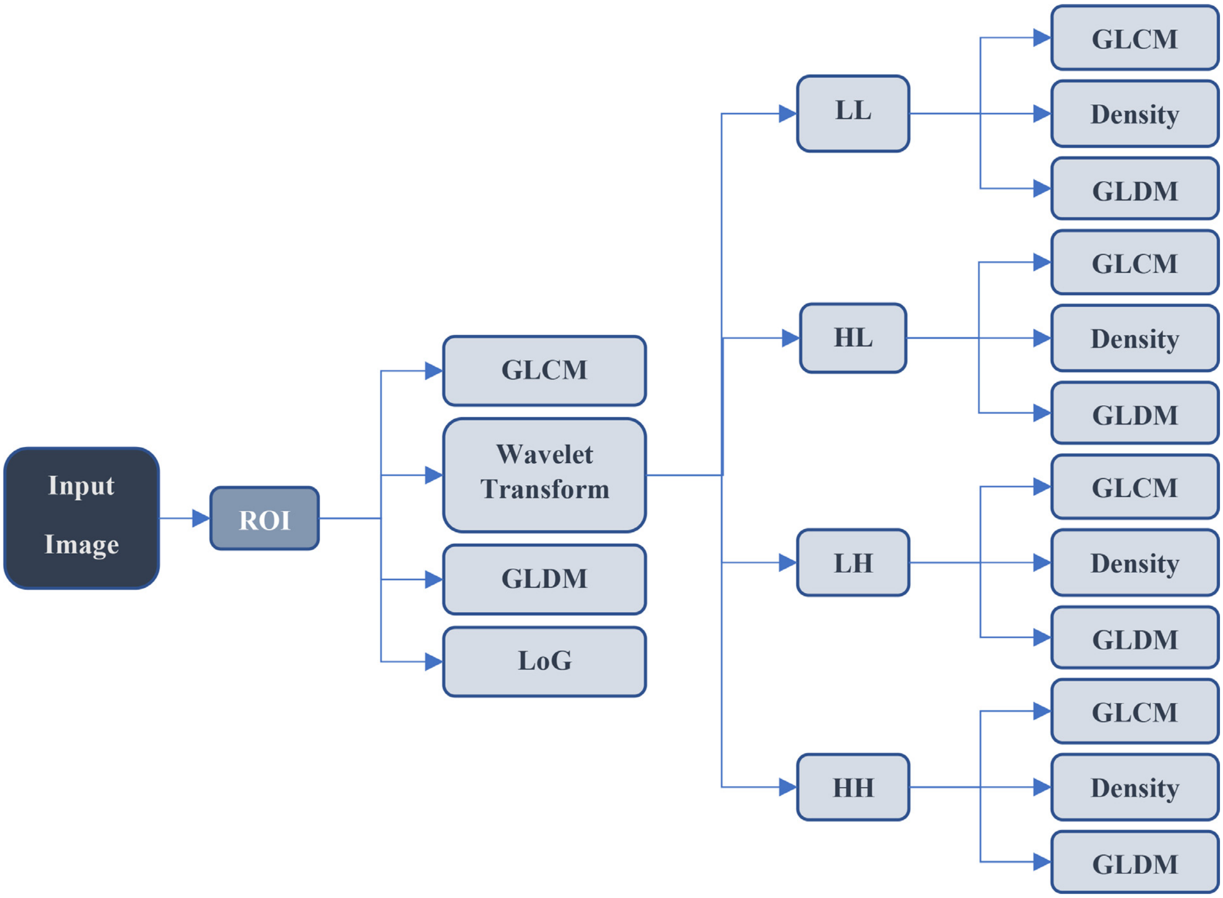


Fig. 4. Diagram of feature extraction process.

methods over their alternative is that they generate more robust results and computationally inexpensive [19,23].

In this study, we will apply RPA to generate optimal features from the original large pool of radiomics features. Following is a brief introduction of the RPA method. By considering each case as a point in a k dimensional space, where k represents the number of features, the Euclidian distance between two points can be expressed as follows:

$$|M-N| = \sqrt{\sum_{i=1}^{k} (m_i - n_i)^2} \tag{2}$$

Regarding Formula (2), $M = (m_1,..., m_k)$, and $N = (n_1,...n_k)$ are two points in the k dimensional space. Likewise, the volume of a sphere with radius r and volume of V in k dimensional space is defined as follows in Formula 3 [24]:

$$V(k) = \frac{r^k \pi^{\frac{k}{2}}}{\frac{k}{2}\Gamma\left(\frac{k}{2}\right)} \tag{3}$$

The normalization of the feature matrix between [0, 1] suggests that all data can be included in a sphere with a radius of 1. The important fact about a sphere with unit radius is that the more increase in dimension, the more reduction in the volume (Formula 4). Simultaneously, the possible distance between the two points remains at 2 [24].

$$\lim_{k\to\infty} \left(\frac{\pi^{\frac{k}{2}}}{\frac{k}{2}\Gamma\left(\frac{k}{2}\right)} \right) \cong 0 \tag{4}$$

Additionally, according to the theory of the heavy-tailed distribution, for a case like $M = (m_1,..., m_k)$ in the space of features, considering features independent with an acceptable approximation, or almost perpendicular variables mapping to different axes, with $E(m_i) = p_i$, $\sum_{i=1}^{k} p_i = \mu$ and $E|(m_i - p_i)^d| \le p_i$ for $d = 2, 3, ...,$ $t^2/6\mu$, then, a probability can be computed using Formula 5 [24]:

$$prob\left(\left| \sum_{i=1}^{k} m_i - \mu \right| \ge t \right) \le Max\left(3e^{\frac{-t^2}{12\mu}}, 4 \times 2^{\frac{-t}{e}} \right) \tag{5}$$

The more the value of t increases, the less chance of a point be out of that distance. Thus, M should be focused around the mean value. In particular, according to Formula 4 and 5, with a satisfactory estimation, all data are contained in a sphere of unit size, and they are focused around their mean value. As a result, if the dimension increases, the volume of the sphere would close to zero. Therefore, the difference between the cases is not enough for accurate classification.

According to the above analysis, the larger the initial feature vector size, the bigger the space dimension is. Hence, most of the data is focused around the center, which leads to less difference between the features. Consequently, to reduce the feature dimension, a powerful technique is the one that reduces the dimensionality of features while preserves the distance between the points, indicating rough preservation of the vast amount of information. If we implement a conventional feature selection method and choose a d-dimensional sup-space of the initial feature vector randomly, it is expected that all the projected distances in the new space are within a determined scale-factor of the initial k-dimensional space . Thus, it is probable that after removing the redundant features, the accuracy would not increase due to the fact that the divergence between the points is not significant enough to consider as a robust model.

To address the concern discussed above and to optimize the feature space, Johnson-Lindenstrauss Lemma's theory can be applied in RPA [25]. This theory states that for any $0 < \varepsilon < 1$, and for any number of cases as t, which are like the points in k-dimensional space (R^k), if assuming d as a positive integer, Formula 6 can be used to compute this integer number [25]:

$$d \geq 4\frac{\ln t}{\left(\frac{\epsilon^2}{2} - \frac{\epsilon^3}{3}\right)} \tag{6}$$

Afterward, for any set W of t points in R^k, for all z, $w \in W$, it is revealed that there is a map, or random projection function like f: $R^k \rightarrow R^d$, which keeps the distance determined by Formula 7 [25]:

$$(1-\in)|z-w|^2 \leq |f(z)-f(w)|^2 \leq (1+\in)|z-w|^2 \tag{7}$$

The above approximation also can be achieved from Formula 8 as follows [25]:

$$\frac{|f(z)-f(w)|^2}{(1+\in)} \leq |z-w|^2 \leq \frac{|f(z)-f(w)|^2}{(1-\in)} \tag{8}$$

As demonstrated in Formula 8, the distance between the set of points in the lower-dimension space is roughly close to the distance in high-dimensional space. The Lemma theory declares that it is feasible to project a set of points from a high-dimensional space into a lower-dimensional space, as the distances between the points are approximately preserved.

As a result, the above analysis suggests that if the initial set of features are projected into space with a lower-dimensional subspace using the random projection method, the distances between points are preserved under better contrast. Hence, it may improve the classification accuracy between the features of two classes representing cases either with or without PM under low risk of overfitting ML models.

In this study, we also investigate whether using RPA can yield a better result in comparison to several commonly used feature dimensionality reduction methods used in the medical imaging informatics field, including principal component analysis (PCA) [26], least absolute shrinkage, and selection operator (LASSO) [27], maximum relevance and minimum redundancy (MRMR) [28], and recursive feature elimination (RFE) [29]. All extracted features in the above section are fed into the methods of RPA, PCA, LASSO, MRMR, and RFE. Each method generates 20 optimal features out of the initially large pool of 315 features.

2.5. Machine learning model

To classify between the study cases with or without PM, we build a multi-feature fusion-based machine learning model. However, due to the unbalance of our dataset, which includes 121 PM cases and 38 non-PM cases, we apply a synthetic minority oversampling technique (SMOTE) algorithm [30] to rebalance the original image dataset. The advantages of using SMOTE to develop machine learning models in medical images have been well investigated and demonstrated in many previous studies (including those conducted by researchers in our lab) [31–33]. In this study, we apply the SMOTE method to generate 83 synthetic non-PM cases. Thus, the dataset is expanded to 242 cases, including 121 PM cases and 121 non-PM cases.

After addressing the imbalance dataset, we select and implement the Gradient Boosting Machine (GBM) to train an optimal machine learning model to predict the risk of advanced gastric cancer patients having PM. The GBM model is a popular machine learning algorithm that has proven effective at classifying complex datasets and often first in class with predictive accuracy [34]. Under a hyperparameter tuning, the GBM model is implemented to achieve a low computational cost and high robustness in detection results as well. Additionally, to decrease the case partition and feature selection (or generation) bias, we use a leave-one-case-out (LOCO) based cross-validation method to train and test the GBM model. In each LOCO cycle, PRA and SMOTE are embedded in the training process. Then, one case not involved in the training cycle is tested by the GBM model trained using all other cases in the dataset. The model produces a prediction score for each testing case ranging from 0 to 1. A higher score indicates a higher risk of PM. The prediction performance is evaluated using a receiver operating characteristic (ROC) method after discarding all SMOTE generated non-PM training samples. The areas under ROC curves (AUC) and overall prediction accuracy after applying an operating threshold ($T = 0.5$) on the GBM model generated prediction scores are used as two performance evaluation indices. Additionally, Cohen's Kappa coefficient value is also computed for evaluating the performance of the CAD scheme. High Cohen's Kappa coefficient value (ranging from zero to one) illustrates high robustness and less randomness in the predicted results [35,36].

Table 2
The performance comparison of five GBM models optimized using five different feature selection and reduction methods.

	Precision	Sensitivity	Specificity	Accuracy	AUC
LASSO	38.9%	31.1%	80.0%	65.8%	0.59 ± 0.013
PCA	38.5%	64.1%	65.5%	65.2%	0.58 ± 0.021
RFE	56.5%	62.5%	51.2%	56.9%	0.60 ± 0.020
MRMR	50.0%	32.7%	82.0%	64.5%	0.60 ± 0.017
RPA	65.8%	43.1%	87.1%	71.2%	0.69 ± 0.019

Table 3
The comparison of two GBM model performance between using 2D and 3D image features generated using the RPA method.

	AUC	Accuracy
2D features	0.66±0.017	68.4%
3D features	0.69±0.019	71.2%

In summary, Fig. 5 shows a complete flow chat of using our CAD scheme to process images, compute optimal features, and train the GBM model in which the RPA and SMOTE are embedded inside the LOCO process. In this study, the segmentation and feature extraction steps were performed using MATLAB R2019a package, and the feature reduction and classifications were done using Python 3.7.

3. Results

Fig. 6 presents five ROC curves generated by the GBM models embedded with five different feature reduction methods (LASSO, PCA, RFE, RPA, MRMR). Table 2 shows the performance comparison between using RPA and the other four feature selection methods. The AUC value and the overall prediction accuracy of the GBM model trained using RPA with 3D image features as input are 0.69±0.019 and 71.2%, respectively. Moreover, the precision, sensitivity, and specificity of the proposed method are 65.78%, 43.101%, and 87.12%, respectively. The results indicate that using RPA leads to generate an optimal image feature vector that can build a GBM model with significantly higher prediction accuracy ($p < 0.05$) than using the GBM models optimized using the other four feature optimization methods.

Fig. 7 shows two ROC curves, and Table 3 reports the prediction performance values to compare two GBM models trained using 2D features computed from the largest tumor region segmented

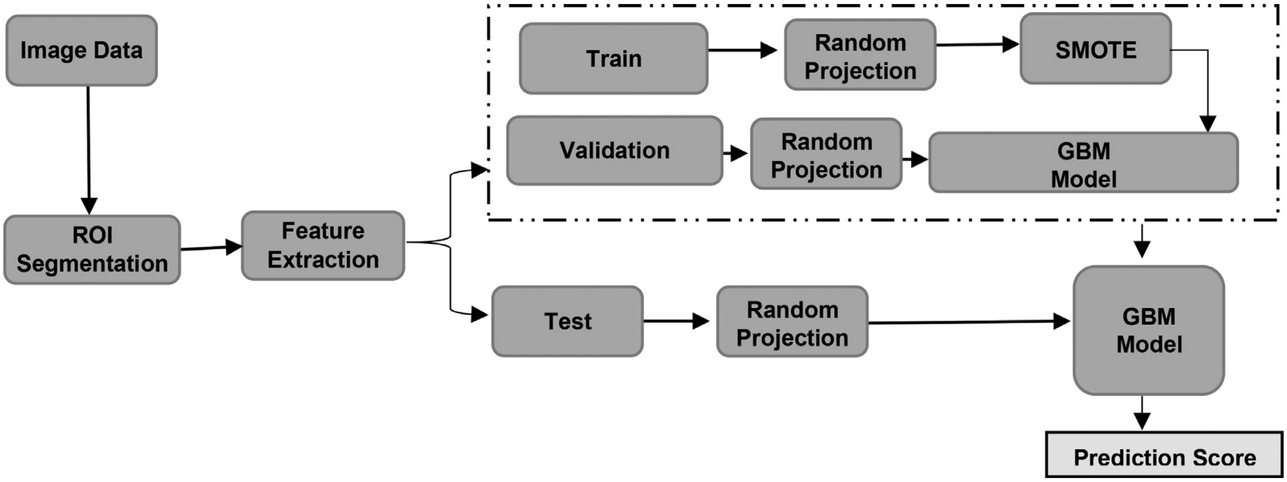


Fig. 5. The flowchart of the proposed CAD scheme.

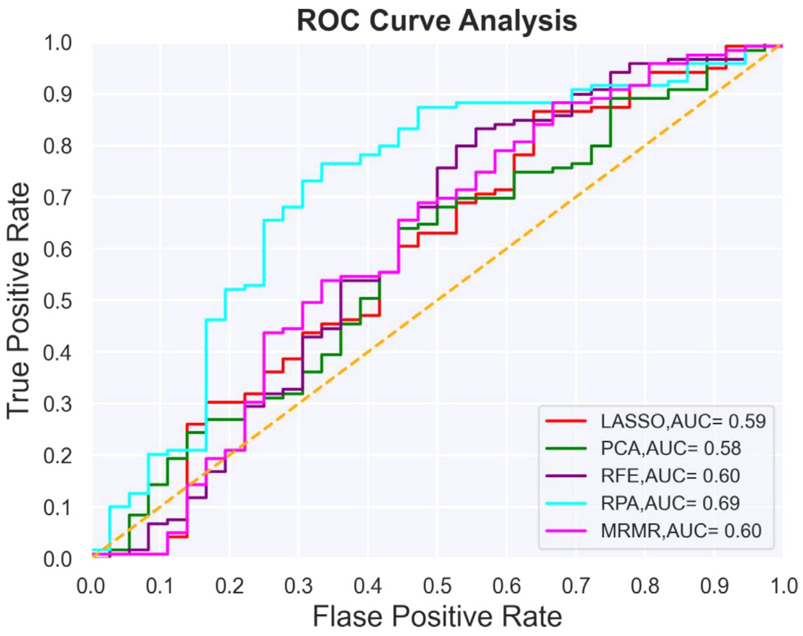


Fig. 6. Comparison of five ROC plots generated using GBM models optimized using five different feature selection or reduction methods.

from one CT image slice and the 3D features computed from the segmented tumor volumes. In these two GBM models, the RPA method is used to select and generate optimal features. The results demonstrate that using 3D image features yields significantly higher performance than using 2D features ($p < 0.05$) in predicting the risk of gastric cancer cases with PM.

In addition, we also build and compare several other types of ML models, including logistic regression, support vector machine (SVM), random forest, and decision tree. All models are trained and tested using the same LOCO cross-validation method embedded with RPA and SMOTE schemes. Table 4 and Fig. 8 present the results to compare the prediction performance of five ML models, which shows that GBM yields the highest accuracy than the other four ML models. However, AUC values between GBM, SVM, and logistic regression-based ML models are not statistically significantly different ($p > 0.05$).

Table 4
Comparison of prediction performance of five ML models.

	AUC value	Accuracy
SVM	0.66	64.55%
Logistic Regression	0.68	61.93%
Random Forest	0.63	69.03%
Decision Tree	0.56	65.16%
GBM	0.69	71.15%

4. Discussion

CT is the most popular imaging modality to detect and diagnose gastric cancer, and it may also provide a non-invasive alternative method to predict the risk of PM in advanced gastric cancer patients. Despite the potential advantages of using CT to detect or predict the risk of PM, the efficacy of radiologists in reading and interpreting CT images for PM detection is insufficient [37]. Although studies have suggested that developing and applying CAD schemes integrated with the radiomics concept and ML model is beneficial and may provide radiologists a second opinion to more accurately detect and diagnose different abnormalities [38], developing ML models using a large number of radiomics features and small training dataset remains a difficult task. In this study, we ex-

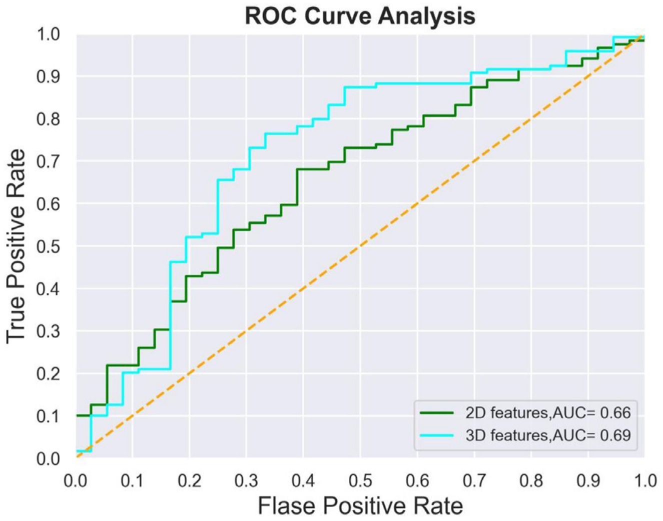


Fig. 7. Comparison of two ROC plots generated by two GBM models optimized using 2D and 3D features generated using the RPA method, respectively.

plore a new approach to develop a new CAD scheme or ML model with several unique characteristics and novel ideas in feature extraction and ML model optimization to improve accuracy in detecting advanced gastric patients with PM.

First, in a previous study conducted in this area, the authors performed manual segmentation of gastric cancer tumor regions from the single CT image slices [39]. However, manual segmentation of tumor regions is often inconsistent with large inter-observer variability due to the fuzzy boundary of the tumor regions, which makes the computed image features also inconsistent or not reproducible. Thus, the prediction accuracy may be affected or not robust. To solve this issue, we in our study developed an interactive CAD scheme with a graphical user interface (GUI) to initiate the segmentation of tumor regions from CT images. A user only needs to place an initial seed around the center of the tumor region that has the largest size in one CT slice. CAD scheme then segments tumor regions on all involved CT image slices automatically. The segmentation results can also be visualized by the human eyes on the GUI window. Although we have designed and installed a correction function icon in the GUI and the user can activate this function to order CAD scheme correcting the segmentation errors (if any), the results in this study show that the CAD scheme can achieve satisfactory results on automatically segmenting all 3,305 tumor regions from all 159 cases in our dataset.

Second, although the previous study [40] has reported developing a radiomics based ML model to detect and diagnose gastric cancer using CT images, in that study, the Authors used image features computed just from one manually selected CT image slice, which may not accurately represent image features of the entire tumor. To address this issue, we conduct the first study that develops and tests a new ML model using 3D image features. Our study results support our hypothesis that using 2D image features extracted from only one CT slice is not sufficient enough to represent the heterogonous characteristics of the tumors, while using 3D image features can yield significantly higher prediction performance. Specifically, in this study, we have performed 3D tumor segmentation and extracted 3D image features to detect or predict the risk of advanced gastric patients having PM. As shown in Table 3, the prediction performance of the GBM model trained using 3D features yields AUC=0.69±0.019 and an accuracy of 71.2%, which are significantly higher than the GBM model trained using 2D features with AUC=0.66±0.017 and the accuracy of 68.4% ($p < 0.05$), respectively.

Third, in developing CAD schemes to train ML models, identifying a small and efficient set of image features plays a critical role [41,42]; therefore, in previous studies, different feature dimensionality reduction methods have been investigated [43,44]. Although these studies made many improvements in optimizing the feature vectors, there is a significant challenge of achieving small feature vectors representing the complex and non-linear image feature space. In this study, we investigate the feasibility of applying the RPA to the medical imaging informatics field in optimizing the CAD scheme or ML model. Our study results show that RPA is a promising technique to reduce the dimensionality of a set of points lying in Euclidian space for very heterogeneous feature data, which commonly occurs in medical images and has advantages to achieve high robustness in classification and low risk of overfitting. Fig. 6 illustrates that the prediction performance of the GBM model embedded with RPA yields significantly higher performance than other GBM models embedded with the other four popular feature reduction methods (PCA, LASSO, MRMR, and RFE). As presented in Table 2, the AUC value after applying the RPA reached the highest prediction accuracy of 71.2% than the other four feature reduction methods. Moreover, the computed Cohen's Kappa coefficient value is 0.68, which indicates the reliability or robustness of the GBM model optimized using the RPA method.

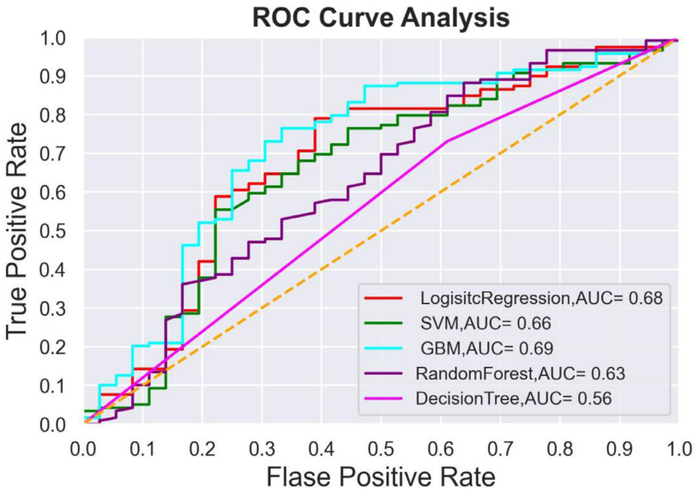


Fig. 8. Comparison of ROC plots of five ML models.

Fourth, since many ML models have been developed and used in medical imaging informatics or CAD fields, selecting which ML model can also be a challenging issue. In this study, we also compare the prediction performance of five popular ML models. The results show that many different ML models can yield very comparable performance, as shown in Table 4 and Fig. 8. However, comparing with the data presented in Table 2, we can find that selecting or generating optimal features plays a more critical role or contribution than choosing a different ML model. Thus, combing the above new observations of this study, we demonstrate that due to the very complicated distribution of radiomics features computed from medical images, RPA is a promising and more powerful technique applicable to generate optimal feature vectors for better training ML models used in CAD schemes of medical images.

Despite the encouraging results, we also notice some limitations in this study. First, the dataset used in this study is relatively small; hence to validate the results of this study, larger datasets are required before being tested in future prospective clinical studies. Second, although in this study, we have used synthetic data to balance the dataset and reduce the impact of an imbalanced dataset, using the SMOTE technique is just efficient for the low dimensional data, and it may not be appropriate or optimal for high dimensional data [45]. Third, in the initial pool of features, we only extracted a limited number of 315 statistics and textural features, which are much less than the number of features computed based on recently developed radiomics concepts and technology in other studies [46]. Thus, more texture features can be explored in future studies to increase the diversity of the initial feature pool, which may also increase the chance of selecting or generating more optimal features to significantly improve the accuracy of the ML model to predict the risk of PM. To overcome the above limitations, more studies and progress are needed in this field.

In summary, regardless of the above limitations, this is a valid proof-of-concept study that reveals a new and promising approach to identify and generate optimal feature vectors for training ML models implemented in CAD schemes of medical images. Since optimizing the feature vector is one of the critical steps of building an optimal ML model using the radiomics concept, the presented method in this study is not only limited to the detection of advanced gastric patients with PM, and it can also be beneficial for other medical imaging studies of developing ML models to detect different types of cancers or abnormalities in the future.

Declaration of Competing Interest

The authors declare that they have no competing interests.

Authors' contributions

SM Conceptualized the idea, performed formal Analysis and investigated the Methodology and wrote the original draft. MH assisted with Data Curation. GD helped in Analysis. BZH and SL supervised the project. BZH was responsible for Funding Acquisition.All authors provided critical feedback and helped in reviewing and editing the final draft.

Acknowledgment

This study is supported in part by research grant R01 CA197150 from the National Cancer Institute. The authors also thank the support from the Stephenson Cancer Center, University of Oklahoma.

References

[1] F. Bray, et al., Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA 68 (6) (2018) 394–424.

[2] A. Biondi, et al., Neo-adjuvant chemo (radio) therapy in gastric cancer: current status and future perspectives, World J. Gastrointest. Oncol. 7 (12) (2015) 389.

[3] T. Fukagawa, et al., A prospective multi-institutional validity study to evaluate the accuracy of clinical diagnosis of pathological stage III gastric cancer (JCOG1302A), Gastric Cancer 21 (1) (2018) 68–73.

[4] F.-H. Wang, et al., The Chinese Society of Clinical Oncology (CSCO): clinical guidelines for the diagnosis and treatment of gastric cancer, Cancer Commun. 39 (1) (2019) 1–31.

[5] F. Coccolini, et al., Intraperitoneal chemotherapy in advanced gastric cancer. Meta-analysis of randomized trials, Eur. J. Surg. Oncol. (EJSO) 40 (1) (2014) 12–26.

[6] H. Ishigami, et al., Phase III trial comparing intraperitoneal and intravenous paclitaxel plus S-1 versus cisplatin plus S-1 in patients with gastric cancer with peritoneal metastasis: PHOENIX-GC trial, J. Clin. Oncol. 36 (19) (2018) 1922–1929.

[7] P. Lambin, et al., Radiomics: extracting more information from medical images using advanced feature analysis, Eur. J. Cancer 48 (4) (2012) 441–446.

[8] H.J. Aerts, et al., Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach, Nat. Commun. 5 (1) (2014) 1–9.

[9] Z.-Q. Sun, et al., Radiomics study for differentiating gastric cancer from gastric stromal tumor based on contrast-enhanced CT images, J. X Ray Sci. Technol. 27 (6) (2019) 1021–1031.

[10] L. Wang, et al., CT-based radiomics nomogram for preoperative prediction of No.10 lymph nodes metastasis in advanced proximal gastric cacner, Eur. J. Surg. Obcol. (2020), doi:10.1016/j.ejso.2020.11.132.

[11] B. Zheng, et al., Interactive computer-aided diagnosis of breast masses: computerized selection of visually similar image sets from a reference library, Acad. Radiol. 14 (8) (2007) 917–927.

[12] G. Danala, et al., Classification of breast masses using a computer-aided diagnosis scheme of contrast enhanced digital mammograms, Ann. Biomed. Eng. 46 (9) (2018) 1419–1431.

[13] R.R. Gundreddy, et al., Assessment of performance and reproducibility of applying a content-based image retrieval scheme for classification of breast lesions, Med. Phys. 42 (7) (2015) 4241–4249.

[14] A. Rajaei, L. Rangarajan, Wavelet features extraction for medical image classification, Int. J. Eng. Sci. 4 (2011) 131–141.

[15] D. Hazra, Texture recognition with combined GLCM, wavelet and rotated wavelet features, Int. J. Comput. Electr. Eng. 3 (1) (2011) 146.

[16] S. Mirniaharikandehei, et al., Developing a quantitative ultrasound image feature analysis scheme to assess tumor treatment efficacy using a mouse model, Sci. Rep. 9 (1) (2019) 1–10.

[17] N. Ahmadi, G. Akbarizadeh, Iris tissue recognition based on GLDM feature extraction and hybrid MLPNN-ICA classifier, Neural Comput. Appl. 32 (7) (2020) 2267–2281.

[18] F. Zhao, C.J. Desilva, Use of the Laplacian of Gaussian operator in prostate ultrasound image processing, in: Proceedings of the 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. Vol. 20 Biomedical Engineering Towards the Year 2000 and Beyond (Cat. No. 98CH36286), IEEE, 1998.

[19] E. Bingham, H. Mannila, Random projection in dimensionality reduction: applications to image and text data, in: Proceedings of the Seventh ACM SIGKDD International Conference On Knowledge Discovery And Data Mining, 2001.

[20] Q. Wang, et al., Hierarchical feature selection for random projection, IEEE Trans. Neural Netw. Learn. Syst. 30 (5) (2018) 1581–1586.

[21] M.L. Mekhalfi, et al., Fast indoor scene description for blind people with multiresolution random projections, J. Visual Commun. Image Represent. 44 (2017) 95–105.

[22] N.F.M. Suhaimi, Z.Z. Htike, Comparison of Machine Learning Classifiers for dimensionally reduced fMRI data using Random Projection and Principal Component Analysis, 2019 7th International Conference on Mechatronics Engineering (ICOM), IEEE, 2019.

[23] Xie, H., J. Li, and H. Xue, A survey of dimensionality reduction techniques based on random projection. arXiv preprint arXiv:1706.04371, 2017.

[24] C.C. Aggarwal, A. Hinneburg, D.A. Keim, On the surprising behavior of distance metrics in high dimensional space, International Conference on Database Theory, Springer, 2001.

[25] S. Dasgupta, A. Gupta, An elementary proof of a theorem of Johnson and Lindenstrauss, Random Struct. Algorithms 22 (1) (2003) 60–65.

[26] M. Pechenizkiy, A. Tsymbal, S. Puuronen, PCA-based feature transformation for classification: issues in medical diagnostics, in: Proceedings. 17th IEEE Symposium on Computer-Based Medical Systems, IEEE, 2004.

[27] R. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Statist. Soc. 58 (1) (1996) 267–288.

[28] H. Peng, F. Long, C. Ding, Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy, IEEE Trans. Pattern Anal. Mach. Intell. 27 (8) (2005) 1226–1238.

[29] X. Zeng, et al., Feature selection using recursive feature elimination for handwritten digit recognition, 2009 Fifth International Conference on Intelligent Information Hiding and Multimedia Signal Processing, IEEE, 2009.

[30] A. Fernández, et al., SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary, J. Artif. Intell. Res. 61 (2018) 863–905.

[31] K.J. Wang, et al., A hybrid classifier combining Borderline-SMOTE with AIRS algorithm for estimating brain metastasis from lung cancer: a case study in Taiwan, Comput. Methods Programs Biomed. 119 (2) (2015) 63–76.

[32] S. Yan, et al., Improving lung cancer prognosis assessment by incorporating synthetic minority oversampling technique and score fusion method, Med. Phys. 43 (6Part1) (2016) 2694–2703.
[33] F. Aghaei, et al., Applying a new quantitative global breast MRI feature analysis scheme to assess tumor response to chemotherapy, J. Magn. Reson. Imaging 44 (5) (2016) 1099–1106.
[34] R. Hu, X. Li, Y. Zhao, Gradient boosting learning of Hidden Markov models, in: 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings, IEEE, 2006.
[35] M.L. McHugh, Interrater reliability: the kappa statistic, Biochem. Med. 22 (3) (2012) 276–282.
[36] M. Heidari, et al., Improving the performance of CNN to predict the likelihood of COVID-19 using chest X-ray images with preprocessing algorithms, Int. J. Med. Inf. 144 (2020) 104284.
[37] R. Seevaratnam, et al., How useful is preoperative imaging for tumor, node, metastasis (TNM) staging of gastric cancer? A meta-analysis, Gastric Cancer 15 (1) (2012) 3–18.
[38] V.M. Gonçalves, M.E. Delamaro, F.d.L.d.S. Nunes, A systematic review on the evaluation and characteristics of computer-aided diagnosis systems, Rev. Bras. Eng. Bioméd. 30 (4) (2014) 355–383.
[39] S. Liu, et al., CT textural analysis of gastric cancer: correlations with immunohistochemical biomarkers, Sci. Rep. 8 (1) (2018) 1–9.
[40] R. Li, et al., Detection of gastric cancer and its histological type based on iodine concentration in spectral CT, Cancer Imaging 18 (1) (2018) 1–10.
[41] M. Kuhn, K. Johnson, An introduction to feature selection, in: Applied Predictive Modeling, Springer, 2013, pp. 487–519.
[42] M. Tan, J. Pu, B. Zheng, Optimization of breast mass classification using sequential forward floating selection (SFFS) and a support vector machine (SVM) model, Int. J. Comput. Assist. Radiol. Surg. 9 (6) (2014) 1005–1020.
[43] S. Khalid, T. Khalil, S. Nasreen, A survey of feature selection and feature extraction techniques in machine learning, 2014 Science and Information Conference, IEEE, 2014.
[44] G. Chandrashekar, F. Sahin, A survey on feature selection methods, Comput. Electri. Eng. 40 (1) (2014) 16–28.
[45] R. Blagus, L. Lusa, SMOTE for high-dimensional class-imbalanced data, BMC Bioinform. 14 (2013) 106 -106.
[46] T. Wang, et al., Correlation between CT based radiomics features and gene expression data in non-small cell lung cancer, J. X Ray Sci. Technol. 27 (5) (2019) 773–803.