J. Shanghai Jiao Tong Univ. (Sci.), 2024
https://doi.org/10.1007/s12204-024-2761-9

Ground-Glass Lung Nodules Recognition Based on CatBoost Feature Selection and Stacking Ensemble Learning

MIAO Jun[1*] (苗　军), *CHANG Yiru*[1] (常艺茹), *CHEN Chen*[2] (陈　辰), *ZHANG Maoxuan*[1] (张茂炫), *LIU Yan*[3] (刘　艳), *QI Honggang*[3] (齐洪钢), *GUO Zhijun*[4] (郭志军), *XU Qian*[5*] (徐　倩)
(1. School of Computer Science, Beijing Information Science and Technology University, Beijing 100101, China; 2. Pathological Information Engineering Technology Center, Jinan Supercomputing Technology Research Institute, Jinan 250100, China; 3. School of Computer Science and Technology, University of the Chinese Academy of Sciences, Beijing 100049, China; 4. Department of Radiology, Tianjin Kanghui Hospital, Tianjin 300385, China; 5. Department of Radiology, Huabei Petroleum General Hospital, Renqiu 062550, Hebei, China)

Abstract: Aimed at the issue of high feature dimensionality, excessive data redundancy, and low recognition accuracy of using single classifiers on ground-glass lung nodule recognition, a recognition method based on CatBoost feature selection and Stacking ensemble learning was proposed. First, the method uses a feature selection algorithm to filter important features and remove features with less impact, achieving the effect of data dimensionality reduction. Second, random forests classifier, decision trees, K-nearest neighbor classifier, and light gradient boosting machine were used as base classifiers, and support vector machine was used as meta classifier to fuse and construct the ensemble learning model. This measure increases the accuracy of the classification model while maintaining the diversity of the base classifiers. The experimental results show that the recognition accuracy of the proposed method reaches 94.375%. Compared to the random forest algorithm with the best performance among single classifiers, the accuracy of the proposed method is increased by 1.875%. Compared to the recent deep learning methods (ResNet+GBM+Attention and MVCSNet) on ground-glass pulmonary nodule recognition, the proposed method's performance is also better or comparative. Experiments show that the proposed model can effectively select features and make recognition on ground-glass pulmonary nodules.
Keywords: ground-glass pulmonary nodule, feature selection, ensemble learning
CLC number: TP181, R319　　**Document code:** A

0 Introduction

In recent years, lung cancer has become the cancer with the highest incidence rate and mortality, known as the "king of cancer". In the early stage of lung cancer, lung lesions often manifest in the form of nodules, among which ground-glass nodules (GGNs) are the main manifestation of early lung cancer. In clinical practice, doctors judge the condition of lung lesions by observing the imaging features of pulmonary nodules in computed tomography (CT)[1]. However, doctors are affected by various factors during the film reading process, resulting in missed and misdiagnosed cases[2]. Therefore, it is necessary to use computer aided design (CAD) based on medical imaging to assist doctors in the diagnosis of lung cancer. CAD has fast calculation speed, can perform accurate quantitative calculations, and never gets tired. It can not only reduce the workload of doctors but also improve the diagnostic ability of lesions. CAD diagnostic technology mainly follows the following steps: ① candidate lung nodule segmentation; ② feature extraction and optimization selection of pulmonary nodules; ③ classification and recognition of pulmonary nodules.

The feature dimension of ground-glass pulmonary nodules is too high, and there is too much redundant data. In the process of extracting and optimizing the selection of pulmonary nodule features, it is necessary to remove useless features and retain pulmonary nodule features that are beneficial for nodule classification as much as possible. Gao et al.[3] used the least absolute shrinkage and selection operator (LASSO) to screen the omics features of ground-glass nodules. Wan et al.[4] used the chi-square test to screen the pathological

Received: 2023-12-04　**Accepted:** 2023-12-25
Foundation item: the National Natural Science Foundation of China (No. 62271466), the Natural Science Foundation of Beijing (No. 4202025), the Tianjin IoT Technology Enterprise Key Laboratory Research Project (No. VTJ-OT20230209-2), and the Guizhou Provincial Sci-Tech Project (No. ZK[2022]-012)
*E-mail: jmiao@bistu.edu.cn, xq200309@sina.com

characteristics of ground-glass pulmonary nodules in patients. Cai et al.[5] used Spearman correlation analysis and Lasso regression analysis for feature dimensionality reduction. Liu[6] used principal component analysis (PCA)[7] and the method of removing low variance features for dimensionality reduction analysis. Dai et al.[8] used a hybrid frog jumping algorithm for feature selection, while Darabi et al.[9] integrated the minimum redundancy maximum correlation algorithm and genetic algorithm to select feature subsets. However, these algorithms are only improved at the data level, and their evaluation criteria are independent of specific learning algorithms, so the selected feature subset may not be optimal for different classifiers.

In the process of pulmonary nodule classification and recognition, Li et al.[10] used support vector machine (SVM) algorithm to construct a classifier for identifying benign and malignant pulmonary nodules. Miao et al.[11] used gradient boosting trees for predicting hepatitis C infection. Wu and Zhang[12] used four classic algorithms, namely logistic regression, artificial neural network, SVM and AdaBoost, to identify ground-glass nodules in lung adenocarcinoma. These models have performed well in most fields, but they are all based on a single classifier recognition model. For complex data, the recognition accuracy of a single classifier is often low. To address the aforementioned issues, many scholars have initiated research on ensemble learning. Chang et al.[13] proposed a SVM ensemble algorithm based on grouped features. Pang et al.[14] used Boosting[15] method to predict the survival of rectal adenocarcinoma. Che et al.[16] used Bagging integration to complete disease detection. Ensemble learning has been successfully applied in medical research, but some more mature ensemble learning methods, such as the Boosting method, have strong dependencies between learners, limited generalization ability in ground-glass lung nodule recognition, and small data volume of ground-glass lung nodules, which may lead to overfitting in other ensemble learning method components. In addition, some of the latest deep learning methods, such as ResNet+GBM+Attention[17] and MVCSNet[18], have also been introduced into lung nodule classification and recognition tasks.

In response to the above problems, this paper proposes a method based on gradient boosting (CatBoost) feature selection and Stacking ensemble learning. This method combines the data level and the model level to improve the accuracy of the identification of ground-glass nodules. At the feature data level, this paper adopts the CatBoost feature selection algorithm to construct the optimal feature subset for the problem of redundancy in ground-glass nodule feature data. At the model level, a Stacking ensemble learning model is proposed, which integrates four base classifier models: random forest (RF), decision tree (DT), K-nearest neighbor (KNN), and LightGBM, and the probability values output by the first layer of base classifiers are used as the input of the second layer learner to achieve the identification of ground-glass nodules. Through 5-fold cross-validation on the lung CT dataset provided by the General Hospital of North China Petroleum Administration, comparative analysis of the proposed model with other feature selection methods and classifier models is conducted. The experimental results show that the CatBoost feature selection and Stacking ensemble learning method proposed in this paper outperforms the mainstream ground-glass nodule identification methods in terms of classification accuracy.

1 Related Feature Selection and Ensemble Learning Methods

1.1 CatBoost Feature Selection Method

The CatBoost model provides various calculation methods for feature importance, as follows:

(1) Prediction-Values-Change. The influence of each input feature on the predicted value is considered separately. The basic idea is to use the definition of derivative for reference: if the value of important features changes, the predicted value will change greatly. The advantage is that only one model needs to be trained in the process of model training, and the calculation efficiency is high. The disadvantage is that the feature selection deviation can construct extreme examples with high feature importance but little impact on actual indicators. CatBoost's base model uses a symmetric tree structure, which feeds each feature into the tree to calculate its importance. The importance of each feature is defined as

$$I = \sum_{N_\mathrm{t}, N_\mathrm{l}} c_1(v_1 - \bar{v})^2 + c_2(v_2 - \bar{v})^2, \tag{1}$$

$$\bar{v} = \frac{c_1 v_1 + c_2 v_2}{c_1 + c_2}, \tag{2}$$

where N_t is the number of tress, N_l is the number of leaves, c_1, c_2 are respectively the sample weights corresponding to the left and right leaf nodes of the corresponding leaf node, and v_1, v_2 are the values related to the leaf node model.

(2) Loss-Function-Change. Consider the impact of each eigenvalue on the loss function separately. The core idea is to retrain a model without this feature according to the given data set, compare the difference between the original and new models in the loss function, and obtain the feature importance. The advantage is that the selection is unbiased, and the retrained model completely eliminates the influence of current features so that the evaluation is more accurate. The disadvantage is high time complexity. The definition of the

feature importance is shown in

$$I = |(f(E_i(v)) - \text{best\_}f)| - |(f(v) - \text{best\_}f)|, \quad (3)$$

where $E_i(v)$ represents the expected value of the model after removing the ith feature, v is the value of the model that contains the ith feature, f is the loss function, and best_f represents the best performance of the model. The first term of Eq. (3) is the loss of the model without a certain feature, and the second term is the loss of the model with a certain feature. If the value of Eq. (3) is negative, that is, the loss value of the second term containing a certain feature is greater than the loss value of the first term without a certain feature, this result means that the introduction of the feature brings an increase in loss and is not a good feature. In other words, the larger the value of Eq. (3), the better the feature, and the smaller the value, the worse the feature. Therefore, when using the Loss-Function-Change importance evaluation index for feature selection, features with negative importance should be removed to eliminate the negative impact of these features on subsequent models.

In response to the issue of identifying ground-glass pulmonary nodules, this article needs to study which features are extracted from ground-glass pulmonary nodules and which feature importance measurement methods are used to achieve the best possible feature selection and separate recognition performance.

1.2 Stacking Ensemble Learning Methods

Stacking an integrated learning method is an integrated classification strategy that combines multiple single-models. The Stacking model is usually designed with a 2-layer structure, as shown in Fig. 1.

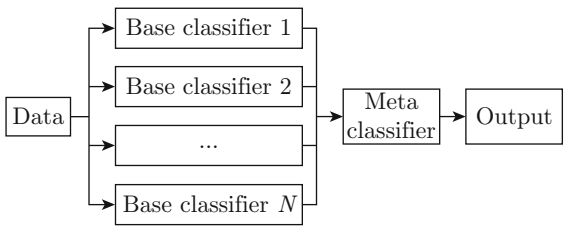

Fig. 1 Stacking model structure.

During the training process of the Stacking ensemble model, the original dataset D is first evenly divided into k mutually exclusive subsets: $D = \{D_1, D_2, \cdots, D_k\}$. Then, select one subset as the test set in sequence and the union of the remaining $(k-1)$ subsets as the training set. Next, train and test each primary learner $h_1, h_2, \cdots, h_n$ in the Stacking ensemble model and obtain the outputs of n base classifiers $T = \{T_1, T_2, \cdots, T_n\}$. This process is also known as k-fold cross-validation.

Finally, the output result of the base classifier obtained from the above process is input as new features $D' = T = \{T_1, T_2, \cdots, T_n\}$ into the meta classifier h' to obtain the final output result.

In response to the issue of identifying ground-glass pulmonary nodules, this article needs to study which base classifiers are specifically integrated under the ensemble learning framework and which meta classifiers are used to achieve the best possible recognition performance.

2 Feature Extraction and Ensemble Learning of Ground-Glass Pulmonary Nodules

This article proposes a ground-glass lung nodule recognition method based on CatBoost feature selection and Stacking ensemble learning, targeting the characteristics of small quantity, high feature dimension, and large amounts of redundant data in ground-glass lung nodules. It mainly involves the following work steps:

(1) Data preparation. Randomly select a certain number of lung nodules from the collected data as the dataset for this study and divide the training and testing sets.

(2) Feature extraction. Extract imaging omics features from experimental data of ground-glass pulmonary nodules for subsequent feature selection.

(3) Feature selection. The feature dimension of ground-glass pulmonary nodules is high, and redundant features may affect the model construction effect. Therefore, this article uses CatBoost to select features from the extracted features.

(4) Ensemble learning methods. We have designed a Stacking ensemble learning method that integrates multiple heterogeneous base classifiers, which can avoid overfitting on small datasets and improve model classification accuracy.

2.1 Evaluation Indicators

In evaluating medical image classification methods, using only a single accuracy to assess the performance of classifiers has certain limitations. Therefore, it is necessary to comprehensively evaluate the Accuracy and Recall of positive and negative samples. This article focuses on the recognition of ground-glass pulmonary nodules, which is a binary classification task. Therefore, this article selects Accuracy, Sensitivity, Specificity, and F1 score as evaluation indicators. The F1 evaluation index involves Precision and Recall evaluation indicators, where true positive (TP) is the number of correctly classified ground-glass lung nodules, false positive (FP) is the number of samples classified as ground-glass lung nodules, true negative (TN) is the number of samples correctly classified as ground-glass lung nodules, and false negative (FN) is the number of samples classified as ground-glass lung nodules. The

above evaluation indicators are shown in

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \tag{4}$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{5}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}, \tag{6}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{7}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{8}$$

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{9}$$

2.2 Data Collection and Labelling of Ground-Glass Pulmonary Nodules

We collected and organized lung CT data from General Hospital of North China Petroleum Administration, resulting in a dataset containing 500 lung CT images. Five experienced radiologists annotated this dataset, and everyone was trained using a unified standard before annotation. The dataset is first independently labelled by three doctors, recording the diameter, center coordinates, and type of pulmonary nodules. Next, the deputy chief physician checks the labelling results of the first round and provides the conclusion. Finally, the chief physician verifies the labelling results of the first two rounds and provides the final determination. This article visualizes 2 416 nodules labelled as ground-glass pulmonary nodules and non-ground-glass pulmonary nodules. Among them there are 369 ground-glass pulmonary nodules, 1 559 non-ground-glass pulmonary nodules, and 488 uncertain types. This article selected 431 non-ground-glass pulmonary nodules and 369 ground-glass pulmonary nodules as the dataset for the experiment and divided them into training and testing sets at a $4:1$ ratio.

2.3 Feature Extraction of Ground-Glass Pulmonary Nodules

Firstly, feature extraction is performed on the experimental data, as shown in Fig. 2. The specific steps are as follows:

(1) Extract the nodule area (3D data block) from the 3D image data based on the coordinate information of the nodules in the doctor label file.

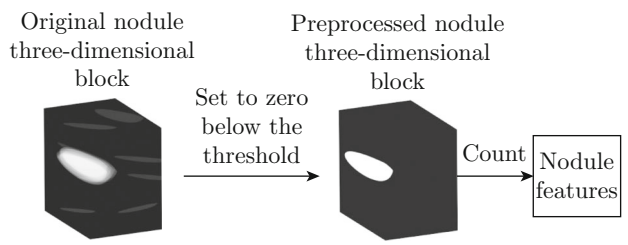

Fig. 2 Radiomics feature extraction.

(2) Set the grayscale values below 60 on the lung nodule data block obtained in Step (1) to 0.

(3) Calculate the imaging omics features on the data block.

Extract 45 radiomics features for each lung nodule, including 15 first-order features, such as energy, maximum, and variation. Based on the principle of the gray level co-occurrence matrix, 20 texture feature values, such as contrast, entropy, and dissimilarity, were extracted. Based on the principle of gray-level co-occurrence matrix, 10-feature information was extracted, such as short run emphasis (SRE), long run emphasis (LRE), low gray level run emphasis (LDLRE) and short run high gray level emphasis (SRHGLE).

2.4 Feature Selection Study of Ground-Glass Pulmonary Nodules

The data volume of ground-glass pulmonary nodules is small. The CatBoost algorithm has good classification performance and strong robustness for small sample datasets. Moreover, the CatBoost algorithm is an embedded feature selection method based on tree models, which combines the advantages of filtering and wrapping, embedding feature selection into the model construction process. CatBoost feature selection depends on the importance of each feature to the model. Compared with general feature selection methods, using CatBoost for feature selection can better characterize the combination relationship between features and more effectively eliminate redundant and irrelevant features. In addition, in the literature review, most feature selection methods used for ground-glass pulmonary nodules are filtering or wrapping, and embedded feature selection methods are rarely used. Therefore, this article uses the CatBoost-based feature selection method.

Based on the two feature importance calculation methods of Prediction-Values-Change and Loss-Function-Change provided by Catboost in Subsection 1.1, 45 radiomics features extracted in Subsection 2.3 are modeled. Through experiments, this article ultimately chooses Loss-Function-Change as the calculation method for feature importance. The reason is that the two selected features were fed into the model, and the recognition accuracy was 91.250% and 94.375%, respectively. The experimental results showed that using the Loss-Function-Change calculation method significantly improved the accuracy. In addition, for the dataset of ground-glass pulmonary nodules in this article, the number of features and samples is relatively small. Therefore, when using the unbiased Loss-Function-Change calculation method in this article, the effectiveness of feature selection can be improved, thereby enhancing the model's performance.

2.5 Research on Ensemble Learning of Ground-Glass Pulmonary Nodules

Stacking ensemble learning can effectively improve the robustness of models. Its classifiers can be

heterogeneous or homogeneous, and the models obtained by combining stacked ensemble learning methods have strong generalization ability. In addition, the Stacking algorithm also uses cross-validation to reduce the risk of overfitting, and it can also have good results when there is insufficient data, effectively improving the robustness of the model. Therefore, this article proposes a ground-glass lung nodule recognition method based on the Stacking ensemble learning model. In the research of Stacking algorithm configuration, the model recognition effect is different for different configurations of meta classifiers and base classifiers, so selecting an appropriate single model for fusion is the core of affecting the Stacking integration model.

In order to improve the recognition performance of Stacking, this article introduces LightGBM, KNN, DT, RF, and SVM into the Stacking integrated model. The LightGBM algorithm usually performs well on multiple datasets and generally improves the accuracy of data classification. The KNN algorithm has mature theory and wide applications. Its model is easy to understand and performs well without excessive parameter tuning. The DT model has interpretability and fast classification speed. The RF is relatively stable and not prone to overfitting, with the advantage of high parallelism. The SVM algorithm projects data onto a high-dimensional feature space based on kernel functions and constructs a maximum interval hyperplane, which can reduce overfitting. The above five classifiers meet the principles of diversity and independence when selecting classifiers for the Stacking model, and they perform well in data analysis and evaluation.

In recognition of ground glass pulmonary nodules, in order to verify the accuracy and differences among various models, as well as the optimal model construction method, this paper compares and analyzes the classifiers (LightGBM, KNN, DT, RF, and SVM) using the evaluation indicators in Subsection 2.1, namely Accuracy, Sensitivity, Specificity, and F1 score. The results are shown in Table 1.

Table 1 Experimental results for the different classifiers

Classifier	Accuracy	Sensitivity	Specificity	F1
LightGBM	0.893 75	0.851 35	0.930 23	0.881 12
KNN	0.893 75	0.824 32	0.953 49	0.877 70
DT	0.868 75	0.783 78	0.941 86	0.846 72
RF	0.925 00	0.891 89	0.953 49	0.916 67
SVM	0.918 75	0.905 41	0.930 23	0.911 57

According to the above analysis in the table, RF has the highest accuracy, and SVM has the most heightened sensitivity, while LightGBM, KNN, and DT classifiers have similar classification indicators. Therefore, LightGBM, KNN, DT, and RF are selected as the base classifiers to meet the "accuracy but difference" of the base model. Choosing a simple classifier for meta classifiers can prevent overfitting, while SVM performs well under various indicators and has good learning ability. However, RF has the highest accuracy. Therefore, experiments were conducted using RF and SVM as meta classifiers, respectively. Through experiments, it was found that the final test accuracy of using RF as the meta classifier is 88.75%, while the final test accuracy of using SVM as the meta classifier is 94.375%, which is much higher than that of RF. Therefore, SVM is chosen as the meta classifier in this article.

In order to analyze the performance of each classifier more intuitively, the receiver operating characteristic (ROC) is used for representation. Figure 3 shows the ROC and area under curve (AUC) of ground glass pulmonary nodule recognition results. The horizontal axis represents the probability of false positives, which is 1−specificity, and the vertical axis represents the probability of true positives, which is sensitivity. ROC represents the performance of the classifier. As shown in Fig. 3, the ROC of DT is relatively small, indicating that a single DT model has a weak recognition ability for ground glass pulmonary nodules. However, the underlying idea of ensemble learning is that even if one weak classifier obtains an incorrect prediction, other weak classifiers can still correct the error. Through experiments, it was found that the accuracy of the Stacking model without DT in the base model is 90.625%, while the accuracy of the Stacking model with DT added to the base model is 94.375%. The above results show that even if a single DT has weak performance, adding it to the stacking base model can improve the entire ensemble learning model.

Therefore, this article uses the above five classifiers in the Stacking ensemble learning model, uses LightGBM,

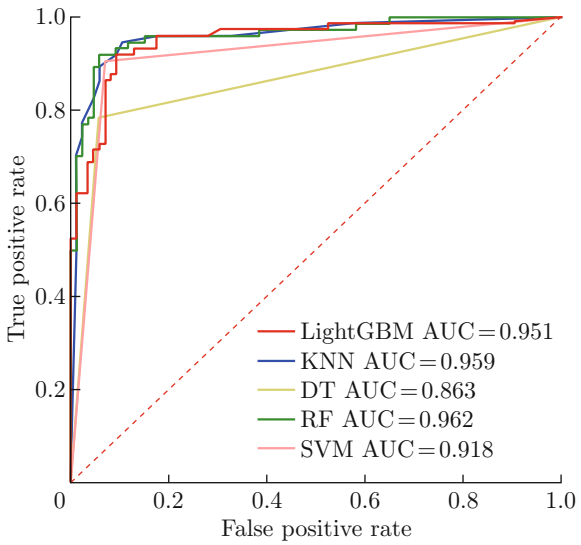

Fig. 3 ROC of different classifier test results.

KNN, DT, and RF as the base classifiers, and uses SVM as the meta classifier to achieve recognition of ground-glass pulmonary nodules.

3 Experiments and Result Analysis

3.1 Experimental Design

3.1.1 *Model Training*

The Stacking model proposed in this article has a two-layer structure: The first layer consists of LightGBM, KNN, DT, and RF; the second layer is composed of SVM. Taking LightGBM in the first layer as an example, based on five-fold cross-validation, this paper divides the training set into five parts, with one part serving as the test set and the rest as the training set. The training and testing sets are repeatedly partitioned 5 times to obtain the predicted values of 5 labels. The training of other base classifiers such as KNN, DT, and RF is similar. After training all base classifiers, construct a new training dataset and a new testing dataset. The new training set is composed of a combination of predicted label values from various base classifiers. In contrast, the new test set is composed of the mean of the test label values of each learner from the original test dataset. Four base classifiers output four predicted values of the original dataset labels, combine these four predicted values as the input dataset for the second layer SVM model, and train and test the second layer model. The above steps complete the training of the integrated model.

The flowchart of the recognition method for ground-glass pulmonary nodules based on CatBoost feature selection and Stacking ensemble learning is shown in Fig. 4.

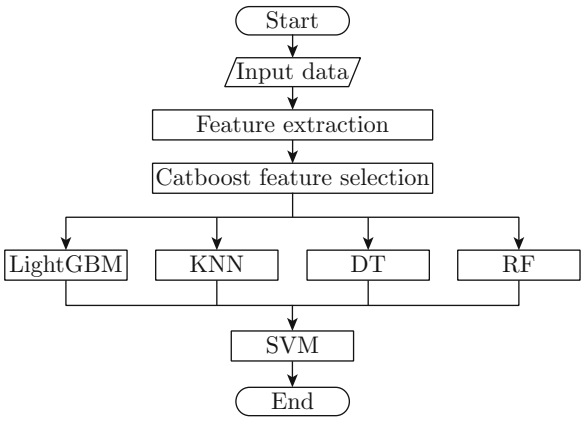

Fig. 4 CatBoost-Stacking flowchart.

3.1.2 *Model Parameter Setting and Optimization*

The selection and setting of parameters will significantly impact the final recognition effect of the model, so reasonable selection of parameters is the key to achieving model optimization. This article uses the grid search method to evaluate and select model parameters when optimizing parameters.

(1) CatBoost feature selection. When performing feature selection on Catboost, the learning_rate is 0.5, the maximum number of iterations is 100, and the depth is 5. Finally, the importance of 45-dimensional features in the original data is obtained.

(2) Stacking ensemble learning. In terms of parameter selection for the classifier, the LightGBM learner adopts n_ Estimators whose number of iterations is 1 000. The weight of the nearest neighbor samples of each sample in the KNN learner is Uniform. The evaluation criterion used by the DT learner for splitting nodes is the Gini index, and the policy splitter for splitting nodes is specified as the best splitting strategy. The RF learner uses a forest with 800 trees and a maximum depth of 100 trees. SVM serves as a classifier, using a radial basis function kernel with a regularization coefficient of 1 and a hyperparameter of 0.5. The above parameters have been experimentally verified to achieve the best results.

The software and hardware environment used in this experiment is as follows: Intel Core i7-7820X CPU; A NVIDIA GeForce RTX 2080Ti graphics card with 11 GB of memory; Windows 10 operating system; Pytorch 1.9.1 deep learning framework; Python 3.7 programming language; integrated development tool PyCharm.

3.2 Feature Selection Analysis and Evaluation of Ground-Glass Pulmonary Nodules

CatBoost feature selection was performed on the 45 radiomics features extracted above, and the original data was standardized and modeled using the CatBoost method to obtain the importance of the 45-dimensional features of the original data, as shown in Fig. 5. From Fig. 5, it can be seen that dissimilarity and integrated memory controller (imc2) are the two most important features, both of which are texture feature values described based on the principle of the gray-level co-occurrence matrix. Therefore, it can be seen that the ground-glass and non-ground-glass lung nodule datasets used in this article exhibit different shapes and possess different imaging information on the high-resolution CT lung window. Some image information is difficult to detect with the naked eye, and the gray-level co-occurrence matrix can reflect the comprehensive information of the image's gray level regarding direction, adjacent interval, change amplitude, and other factors. It can analyze the texture features of the image in the global domain, provide more lesion information, and greatly assist in the recognition of ground-glass pulmonary nodules. The extracted 10-feature information based on the principle of the gray-level co-occurrence matrix has a relatively small impact on the classification importance of ground-glass. The gray-level run-length matrix of an image reflects the image's grayscale in terms of direction, adjacent

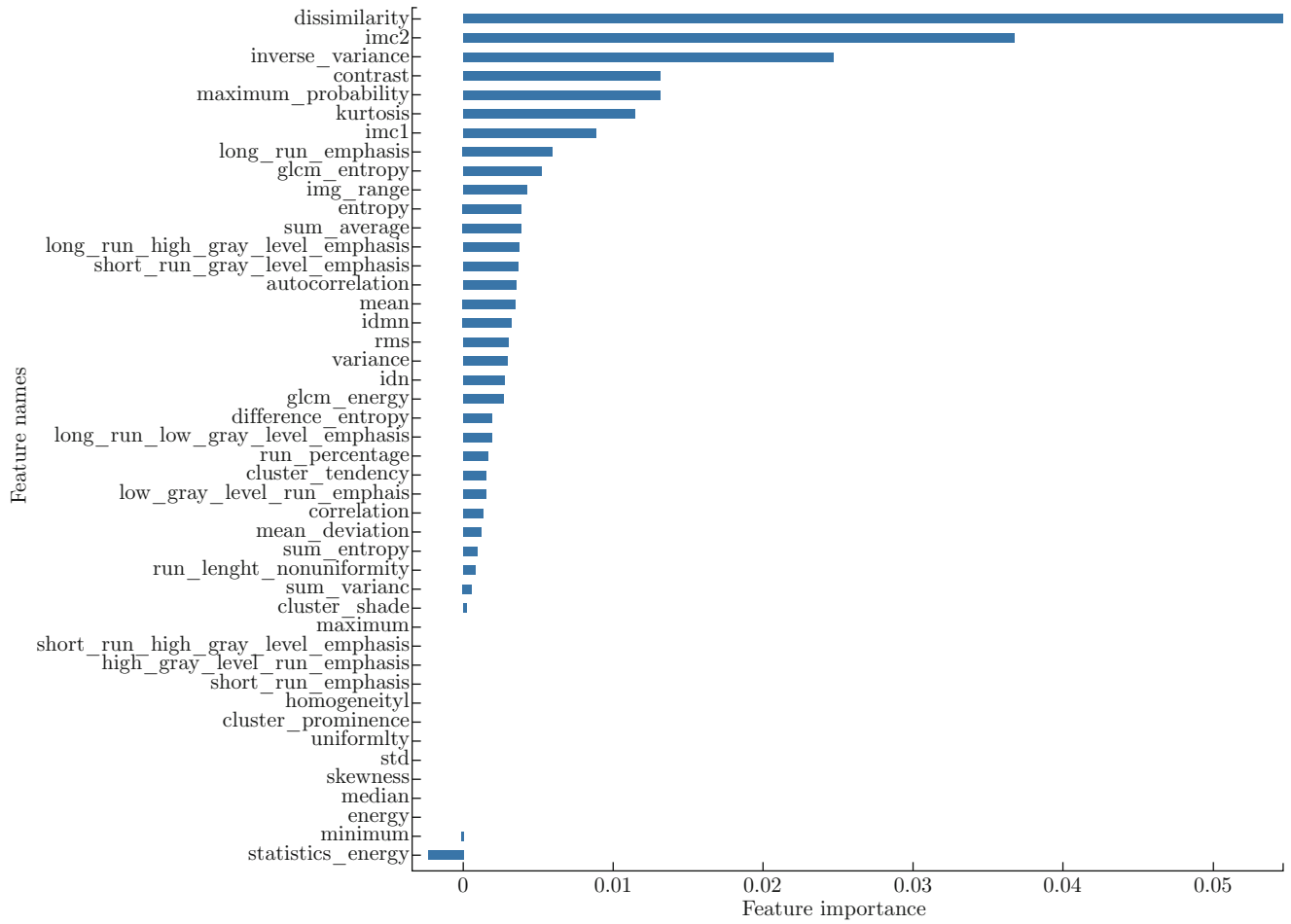

Fig. 5 Feature importance map based on the calculation method of Loss-Function-Change.

intervals, and amplitude of change. However, for pulmonary nodules, this statistical information is less helpful in distinguishing different types of nodules. Therefore, this article chooses to send the top 60% of important features into subsequent models for nodules recognition.

In order to compare the superiority of the CatBoost feature selection method in the classification of ground-glass pulmonary nodules, original features, the Lasso feature selection algorithm[3], chi-square test feature selection algorithm[4], Spearman correlation analysis and Lasso regression analysis feature selection algorithm[5], PCA[7], as well as feature selection algorithms for removing low variance[6], Relief[19] feature selection algorithm[20], and mutual information feature selection algorithm[21] were used, respectively. Compared with the CatBoost feature representation algorithm, the experimental results are shown in Table 2.

Table 2 Effects of different feature choices

Feature expression algorithm	KNN		LightGBM		RF		DT		Stacking	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Original	0.900 00	0.85	0.893 75	0.88	0.918 75	0.90	0.868 75	0.85	0.925 00	0.91
Lasso[3]	0.893 75	0.87	0.868 75	0.85	0.925 00	0.92	0.868 75	0.85	0.912 50	0.91
Chi-square test[4]	0.875 00	0.81	0.893 75	0.88	0.912 50	0.90	0.825 00	0.78	0.900 00	0.91
Spearman correlation analysis and Lasso regression[5]	0.825 00	0.80	0.868 75	0.85	0.893 75	0.88	0.850 00	0.82	0.906 25	0.92
PCA[7]	0.818 75	0.79	0.831 25	0.85	0.837 50	0.80	0.837 50	0.80	0.843 75	0.86
Remove low variance[6]	0.893 75	0.88	0.900 00	0.89	0.912 50	0.90	0.868 75	0.85	0.918 75	0.92
Relief[19]	0.881 25	0.82	0.900 00	0.89	0.893 75	0.87	0.850 00	0.79	0.906 25	0.90
Mutual information[21]	0.868 75	0.85	0.906 25	0.90	0.881 25	0.90	0.812 50	0.76	0.925 00	0.92
CatBoost	0.918 75	0.85	0.893 75	0.88	0.925 00	0.91	0.868 75	0.85	0.943 75	0.93

The experimental results show that the CatBoost feature selection algorithm proposed in this paper not only has a certain dimensionality reduction effect but also extracts features that are more conducive to the recognition task of ground glass pulmonary nodules, and all indicators of the model have been improved.

3.3 Evaluation of Ensemble Learning Algorithms for Ground-Glass Pulmonary Nodules

In order to verify the effectiveness of the CatBoost-based Stacking method in the classification of ground-glass pulmonary nodules, this paper compared it with SVM, voting merging[22], and Bagging algorithms. The classifiers used in voting to merge ensemble learning are RF, DT, KNN, and LightGBM, using soft voting strategy for decision-making. The Bagging algorithm uses a grid search method to optimize the number of sub-models and ultimately obtains the optimal number of base classifiers of 400. In addition, this article also compares the experimental performance of the proposed method with some classic and latest deep learning methods, including ResNet18[23], AlexNet[24], DenseNet[25], ResNet+GBM+Attention[17], and MVCSNet[18]. The experimental results are shown in Table 3.

Table 3 Performance comparison of different classification models

Classifier	Accuracy	Training time/s	Number of model parameters
SVM	0.918 75	16.414	8.424×10^3
Voting-merging[22]	0.856 25	20.764	5.022×10^4
Bagging	0.881 25	19.07	1.674×10^4
ResNet18[23]	0.927 80	1 792	8.298×10^6
AlexNet[24]	0.833 30	2 328	2.512×10^8
DenseNet[25]	0.861 10	1 924	1.124×10^7
ResNet+GBM+Attention[17]	0.938 40	2 402	8.320×10^7
MVCSNet[18]	0.952 10	2 774	9.987×10^7
Proposed method	0.943 75	25.793	5.086×10^4

The calculation method for the number of parameters in the above table is the sum of the SVM model parameters and the parameters of RF, DT, and LightGBM. Among them, the parameter quantity of RF, DT, and LightGBM is the product of the number of training samples and the number of sample features. In contrast, the parameter quantity of SVM is the product of the number (N_1) of support vectors and the sum of the number (N_2) of vector features and the number (N_3) of weight coefficients. The above calculation formula is shown in Parameter quantity $= N_1(N_2 + N_3)$.

The experimental results show that the Stacking method used in this article integrates the advantages of heterogeneous classifiers, with a recognition accuracy of 94.375%, which is 8.75 percentage points higher than the voting merging method. Compared with the Bagging algorithm, the proposed method's accuracy is 6.25 percentage points higher, which fully demonstrates the superiority of the Stacking ensemble learning model compared to a single classifier. Compared with deep learning models ResNet18[23], AlexNet[24], DenseNet[25], and ResNet+GBM+Attention[17], the accuracy of the method proposed in this paper is also higher than theirs. Compared to traditional models, deep models have strong expressive power, but they require more training data to avoid overfitting and ensure good performance on the dataset. This may not be suitable for small sample datasets in this article. Compared with the recently proposed model MVCSNet[18] in the field of ground-glass lung nodule recognition, its accuracy is comparable. The reason for the analysis is as follows: The MVCSNet model introduces an improved self-attention mechanism, which can model the global space well and improve recognition performance but also has high computational complexity. The training time and parameter count of the model proposed in this article are much smaller than those of the MVCSNet model, indicating that the method proposed in this article can achieve equivalent recognition performance with less time and memory costs. It can be inferred that the Stacking ensemble learning method used in this paper has good applicability for ground-glass pulmonary nodules with small data volume and high feature dimensions.

In addition, in this article the highly accurate identification methods for ground-glass pulmonary nodules proposed in recent years are also validated, as shown in Table 4. The experimental results show that compared to the comparative methods, the accuracy of our method has increased by 5.000, 3.282, and 3.125 percentage points, respectively, and it is also higher than existing research results in sensitivity and F1 indicators.

In order to verify that the base classifier of the

Table 4 Comparison with existing research results

Identification method	Accuracy	Sensitivity	F1
Lasso+Logistic regression[3]	0.893 75	0.86	0.88
Spearman+Lasso+SVM[5]	0.910 93	0.90	0.90
Remove low variance+SVM[6]	0.912 50	0.90	0.90
Proposed method	0.943 75	0.92	0.93

algorithm proposed in this article is the optimal combination, comparative experiments were conducted on different combinations of base classifiers, and the results of each experiment are shown in Table 5.

Table 5 Comparative experiments on different combinations of base classifiers

Classifier	Accuracy	F1
LightGBM	0.893 75	0.881 12
KNN	0.893 75	0.877 70
DT	0.868 75	0.846 72
RF	0.925 00	0.916 67
RF+DT+KNN	0.918 75	0.911 56
LightGBM+DT+RF	0.906 25	0.896 55
LightGBM+KNN+RF	0.918 75	0.910 34
LightGBM+KNN+DT	0.912 50	0.904 11
Proposed method	0.943 75	0.930 00

The above table shows that the accuracy of ground-glass lung nodule recognition after integrating four classifiers is higher than that of a single classifier, and the accuracy decreases to varying degrees when any base classifier is reduced. From this, it can be seen that LightGBM, KNN, DT, and RF are the best combinations as the base classifiers for the model in this paper.

4 Conclusion

This article proposes a ground-glass lung nodule recognition model based on CatBoost feature selection and Stacking ensemble learning in the recognition method of ground-glass lung nodules. In response to the possibility of feature redundancy in the extracted medical images, which may affect the recognition results of the classifier, this paper uses the CatBoost algorithm for modeling and then obtains the importance of each feature. The selected features with higher priority are combined into a new dataset and sent to the subsequent recognition model. In addition, in terms of selecting recognition models, this article has modeled and experimented with popular classifiers such as LightGBM, KNN, DT, RF and SVM. Their performance varies greatly from the experimental results, indicating that a single classifier has certain limitations on different datasets. Therefore, this article proposes a recognition method for ground-glass lung nodule images based on Stacking ensemble learning. This method first analyzes a single classifier and builds a Stacking evaluation model with LightGBM, KNN, DT, and RF as the base classifiers and SVM as the meta classifiers. Then, the parameters of the built evaluation model are set. The experimental results show that the method based on CatBoost feature selection and Stacking ensemble learning adopted in this paper is superior to the current mainstream feature selection methods and single classification algorithms. Standard ensemble learning models and deep learning models achieve an accuracy of 94.375% in recognition, verifying the effectiveness of the proposed method in identifying ground-glass pulmonary nodules.

In our next research, we will refer to models such as Transformer to study better-performing lung nodule recognition models.

Conflict of Interest The authors declare that they have no conflict of interest.

References

[1] AGGARWAL P, VIG R, SARDANA H K. Semantic and content-based medical image retrieval for lung cancer diagnosis with the inclusion of expert knowledge and proven pathology [C]//*2013 IEEE Second International Conference on Image Information Processing*. Shimla: IEEE, 2013: 346-351.

[2] WANG X, MA D. Advances in computer-aided diagnosis in pulmonary nodules [J]. *Chinese Journal of Radiology*, 2006, **40**(4): 443-445 (in Chinese).

[3] GAO L, YU X X, KANG B, et al. Predictive value of CT-based radiomics nomogram for the invasiveness of lung pure ground-glass nodules [J]. *Journal of Shandong University* (*Health Science*), 2022, **60**(5): 87-97 (in Chinese).

[4] WAN H Y, LI J, WANG B, et al. Establishment of prediction model for isolated pulmonary benign or malignant nodule by Bayesian network [J]. *Journal of Chinese Oncology*, 2022, **28**(5): 380-384 (in Chinese).

[5] CAI J H, DUAN S F, YUAN H, et al. Machine learning in differentiating pulmonary invasive adenocarcinoma from non-invasive adenocarcinoma manifested as pure ground-glass nodule [J]. *Chinese Journal of Medical Imaging Technology*, 2020, **36**(3): 405-410 (in Chinese).

[6] LIU X F. The clinical value of CT radiomics in the diagnosis of ground-glass pulmonary nodules [D]. Wuhu: Wannan Medical College, 2021 (in Chinese).

[7] MAĆKIEWICZ A, RATAJCZAK W. Principal components analysis (PCA) [J]. *Computers & Geosciences*, 1993, **19**(3): 303-342.

[8] DAI Y Q, GUO X Y, WANG M, et al. Feature selection of high-dimensional biomedical data based on shuffled frog leaping algorithm [J]. *Application Research of Computers*, 2021, **38**(4): 1062-1068 (in Chinese).

[9] DARABI N, REZAI A, HAMIDPOUR S S F. Breast cancer detection using RSFS-based feature selection algorithms in thermal images J]. *Biomedical Engineering*: *Applications*, *Basis and Communications*, 2021, **33**(3): 2150020.

[10] LI Y F, LUO Y, GUO L, et al. Radiomics analysis and machine learning for classification of benign and malignant pulmonary nodules [J]. *Radiologic Practice*, 2021, **36**(4): 464-469 (in Chinese).

[11] MIAO X F, LIU M, JIANG Y. Hepatitis C prediction based on machine learning algorithms [J]. *Journal of Jilin University* (*Information Science Edition*), 2022, **40**(4): 638-643 (in Chinese).

[12] WU T F, ZHANG R S. Research on the application of machine learning in the malignant grinding glass density nodules of lung [J]. *Journal of Guangzhou University* (*Natural Science Edition*), 2018, **17**(3): 33-39 (in Chinese).

[13] CHANG T T, LIU H W, FENG J. Support vector machine ensemble learning algorithm research based on heterogeneous data [J]. *Journal of Xidian University*, 2010, **37**(1): 136-141 (in Chinese).

[14] PANG L, LAN W X, WANG Q Q, et al. Machine learning-based survival prediction model for colorectal adenocarcinoma cancer [J]. *Modern Preventive Medicine*, 2023, **50**(2): 227-232 (in Chinese).

[15] BARTLETT P, FREUND Y, LEE W S, et al. Boosting the margin: A new explanation for the effectiveness of voting methods [J]. *The Annals of Statistics*, 1998, **26**(5): 1651-1686.

[16] CHE X J, YU Y J, LIU Q L, et al. Enhanced Bagging ensemble learning and multi-target detection algorithm [J]. *Journal of Jilin University* (*Engineering and Technology Edition*), 2022, **52**(12): 2916-2923 (in Chinese).

[17] KUANG J, HONG M J, LIU X C, et al. Classification of pulmonary nodules based on attention mechanism [J]. *Computer Applications and Software*, 2022, **39**(1): 163-167 (in Chinese).

[18] ZHU Q K, WANG Y Q, CHU X P, et al. Multi-view coupled self-attention network for pulmonary nodules classification [M]//Computer vision – ACCV 2022. Cham: Springer, 2022: 37-51.

[19] KIRA K, RENDELL L. The feature selection problem: Traditional methods and a new algorithm [C]//*10th National Conference on Artificial Intelligence*. San Jose: AAAI, 1992: 129-134.

[20] HE X Y, GONG J, WANG L J, et al. Feature selection based on feature vectorization on computer tomography scan of pulmonary nodules [J]. *Application Research of Computers*, 2018, **35**(8): 2544-2548 (in Chinese).

[21] WANG J, ZHANG X L, ZHAO J J. Feature selection algorithm for diagnostic model of solitary pulmonary nodules [J]. *China Sciencepaper*, 2014, **9**(10): 1201-1205 (in Chinese).

[22] DIMITRIADOU E, WEINGESSEL A, HORNIK K. Voting-merging: An ensemble method for clustering [M]//Artificial neural networks — ICANN 2001. Berlin, Heidelberg: Springer, 2001: 217-224.

[23] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition [C]//*2016 IEEE Conference on Computer Vision and Pattern Recognition*. Las Vegas: IEEE, 2016: 770-778.

[24] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks [J]. *Communications of the ACM*, 2017, **60**(6): 84-90.

[25] HUANG G, LIU Z, VAN DER MAATEN L, et al. Densely connected convolutional networks [C]//*2017 IEEE Conference on Computer Vision and Pattern Recognition*. Honolulu: IEEE, 2017: 2261-2269.