

Received November 14, 2020, accepted December 17, 2020, date of publication December 21, 2020, date of current version December 31, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3046309

# Effective Breast Cancer Recognition Based on Fine-Grained Feature Selection

GUANGLI LI<sup>1</sup>, TIAN YUAN<sup>1</sup>, CHUANXU LI<sup>1</sup>, JIANWU ZHUO<sup>1</sup>, ZILIANG JIANG<sup>1,2</sup>, JINPENG WU<sup>2</sup>, DONGHONG JI<sup>3</sup>, AND HONGBIN ZHANG<sup>1,2</sup>

<sup>1</sup>School of Information Engineering, East China Jiaotong University, Nanchang 330013, China

<sup>2</sup>School of Software, East China Jiaotong University, Nanchang 330013, China

<sup>3</sup>Cyber Science and Engineering School, Wuhan University, Wuhan 430071, China

Corresponding authors: Hongbin Zhang (zhanghongbin@whu.edu.cn) and Tian Yuan (15072080165@163.com)


This work was supported in part by the National Natural Science Foundation of China under Grant 61762038 and Grant 61861016; in part by the Jiangxi Provincial Department of Science and Technology under Grant 20171BAB202023 and Grant 20192ACB21004; in part by the Key Research and Development Plan of Jiangxi Provincial Science and Technology Department under Grant 20171BBG70093, Grant 20192BBE50071, and Grant 20202BBEL53003; in part by the Humanity and Social Science Fund of Ministry of Education of China under Grant 20YJAZH142; in part by the Science and Technology Projects of Jiangxi Provincial Department of Education under Grant GJJ190323; and in part by the Humanity and Social Science Foundation of Jiangxi University under Grant TQ19101 and Grant TQ20108.

**ABSTRACT** Early detection and diagnosis of breast cancer are crucial to improve the survival rates of patients. Hence, pathologists and radiologists need a computer-aided diagnosis system to assist their clinical diagnoses effectively and efficiently. However, most breast cancer recognition models are faced with the sample scarcity problem, which results in serious overfitting and lowers recognition performance. To alleviate the sample scarcity problem, a simple, effective model called “refinement, correlation, adaptive” (RCA) for breast cancer recognition is proposed from the perspective of fine-grained feature selection. An innovative multiview efficient range-based gene selection algorithm is proposed to complete the first-layer feature “refinement,” which contributes to suppressing the noisy information in the original feature space. Then, more-discriminant but low-dimensional information among heterogeneous features is mined through the second-layer cross-modal “correlation” mining. Feature dimensions are reduced to a reasonable value that fits the sample size well and alleviates the overfitting problem. Finally, the last-layer decision-tree-guided “adaptive” feature selection is completed using the gradient boosting decision tree algorithm. The RCA model was validated on two well-known datasets. The experimental results demonstrate that the proposed RCA model can address the sample scarcity problem well. It outperforms state-of-the-art baselines, especially in terms of accuracy and the area of the Kivi diagram. The largest performance improvements of the metrics are 2.39% and 1121, respectively. Moreover, an online diagnosis system based on the RCA model is proposed. It provides rapid and effective breast cancer recognition, which should make clinical diagnoses more convenient and narrow the gap between theoretical research and practical application.

**INDEX TERMS** Breast cancer recognition, fine-grained feature selection, sample scarcity, efficient range-based gene selection, cross-modal correlation mining, gradient boosting decision tree.

## I. INTRODUCTION

Breast cancer, which is usually associated with women, is a leading cause of death. Early detection and diagnosis help to reduce the mortality of breast cancer and improve the quality of life. However, owing to the lack of adequate medical resources, many patients do not receive timely diagnosis and accurate treatment. Hence, to alleviate this

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

problem, pathologists and radiologists need a computer-aided diagnosis (CAD) system to assist their clinical diagnoses effectively and efficiently. In particular, higher recognition accuracy, a lower false positive rate (FPR), and a lower false negative rate (FNR) can be obtained with the application of state-of-the-art machine-learning technologies, including deep learning [1]–[3] computer vision, feature fusion [4], [5], feature selection [6], [7], and ensemble learning [8], [9]. It is known that medical image annotation with high quality usually has a very large economic cost. Meanwhile, ethical

issues or individual privacy greatly limits the number of available samples. Moreover, there is a large gap between the medicine field and the computer science field, and this also restricts the acquisition of high-quality samples. Hence, most breast cancer recognition models face the well-known sample scarcity problem, which can cause serious overfitting and lower recognition performance and can reduce the practicability of the recognition models.

To address the sample scarcity problem, some researchers have used state-of-the-art generative adversarial network models to generate completely novel samples [10], [11], which can alleviate the sample scarcity problem to a certain degree. However, experts in the medicine field usually question the authenticity of the generated samples. Other researchers have utilized multitask learning (MTL) frameworks that can share discriminant intermediate information among different tasks, such as recognition, segmentation, and detection, to address the sample scarcity problem. However, it is difficult to train an elaborate MTL framework.

In this work, a simple, effective, novel model called “refinement, correlation, adaptive” (RCA) is proposed. Comprehensive explanations are provided in Section II(C). RCA provides effective and efficient breast cancer recognition. The RCA model is derived from the perspective of fine-grained feature selection, which means that deeper pathological information can be acquired by multistage deep semantics mining. Hence, low-dimensional features with more-powerful discriminant ability (see the feature SG in Figures 5 and 6 in Section IV(C) - the distributions of different types of samples vary greatly) are generated to better fit the sample size and address the sample scarcity problem. Meanwhile, the RCA model requires fewer parameters. Therefore, unlike the above-mentioned methods, the proposed RCA model does not need any novel samples, and its training procedure is easier than the above-mentioned methods [1]–[11]. Conceptually and empirically, the RCA model makes the following four contributions.

(1) The RCA model for breast cancer recognition from the perspective of fine-grained feature selection is proposed. More-discriminant information is mined progressively. The RCA model outperforms state-of-the-art baselines. More importantly, it provides a holistic and versatile framework for image classification. Other well-known feature selection or cross-modal analysis methods can be absorbed into this framework seamlessly.

(2) An innovative multiview efficient range-based gene selection (MvERGS) algorithm for the first-layer feature refinement was designed. It helps suppress the noisy information in the original feature space and builds an important foundation for the subsequent feature selection stages. Owing to its scalability and robustness, the MvERGS algorithm can be used in some downstream research fields that require elaborate feature selection. Hence, it is a valuable by-product of the RCA model.

(3) The new features generated by the RCA model have lower dimensions, which fit the sample size well and can

address the overfitting problem. Lower dimensions also improve the running efficiency of the proposed model.

(4) The end-to-end online system for effective breast cancer recognition based on the proposed RCA model was optimized further. This can narrow the gap between theoretical research and practical application. Moreover, owing to its simplicity and transportability, the RCA model can be deployed on a normal workstation and needs fewer computing resources.

This article is organized as follows. In Section II, the related research works are described. After two correlated research fields, breast cancer recognition and feature selection, have been described, the motivations for this work are given. Then, the RCA model is discussed in detail in Section III. In Section IV, extensive experiments on two well-known mammographic datasets are described, and some important conclusions are drawn. Finally, in Section V, the current work is summarized, and planned future work is discussed.

## II. RELATED WORKS

### A. BREAST CANCER RECOGNITION

Extensive research has been done on breast cancer recognition in the medical image analysis field. Recently, with the great success of convolutional neural networks (CNNs) in computer vision fields, some researchers have started to fine-tune pretrained CNNs, such as InceptionV3 [12], AlexNet [13], visual geometry group (VGG) [14], and residual neural network (ResNet) models, for breast cancer recognition. Some researchers have used the bottleneck features of these pretrained CNNs to train traditional classifiers, including support vector machine (SVM),  $k$ -nearest neighbors (KNN), logistic regression (LR), and decision trees, which can take full advantage of the implicit characteristics of the traditional classifiers. For example, Silva *et al.* [15] trained an SVM-based classifier for breast disease diagnosis. Cong *et al.* [16] proposed an ensemble-learning method that integrates a group of traditional classifiers, including KNN, SVM, and naive Bayes, to complete breast cancer recognition. Although the fine-tuning method is easy, it cannot achieve satisfactory performance owing to insufficient training samples. Transfer learning [17], [18] may be a good choice to overcome this problem. It is a relatively efficient method for breast cancer recognition.

Some novel models or methods, such as conditional infilling generative adversarial networks [10], domain adaptation [19], [20], and few-shot learning strategies [21], [22], have been proposed to alleviate the sample scarcity problem. Other researchers have attempted to integrate many CNN-based models to improve the final performance. For example, multiview deep ResNet [23] and context-aware CNNs [24] were proposed in this research direction. This method makes full use of multiple CNNs, especially for heterogeneous networks. Satisfactory performance improvement can be observed after training. However, training so

many deep-learning-based models requires more computing resources, and this is inconvenient for a normal workstation. For this reason, some researchers have attempted to optimize recognition models from two perspectives.

In the first, Shen *et al.* [25] used lesion patches to train an initial patch-based breast cancer recognition model. Then, the pathological knowledge learned by the patch-based model was transferred to whole mammograms. This method can fully use the limited lesion patches to improve the final recognition performance. Certainly, it requires extra lesion annotations.

In the second, some researchers have combined a group of correlated medical image analysis tasks, such as breast cancer recognition, lesion segmentation, and lesion localization, into a holistic multitask learning framework. The implicit complementary information hidden among different tasks can be fully used to improve recognition performance. For example, Chiranji *et al.* [26] designed a multitask learning framework that consists of lesion area segmentation and breast cancer classification based on fuzzy C-means clustering and a fuzzy SVM. The two tasks complement each other to boost the final performance. However, a relatively complicated network should be designed first, and the parameters of the multitask learning framework should be tuned carefully.

In summary, breast cancer recognition has attracted increasing attention in the fields of computer vision and medical image analysis. The above-mentioned deep-learning-based methods [1]–[3] have greatly promoted the research progress. However, because of sample scarcity, overfitting is still one of the most important issues that researchers must address. Meanwhile, ways to make a trade-off between effectiveness and practicality possible have also attracted attention.

## B. FEATURE SELECTION

A feature selection algorithm is effective for medical image analysis. It helps with the sample scarcity problem to a certain degree. It can also improve the real-time efficiency owing to lower dimensions. The feature selection research field includes single-modality feature selection and multimodality feature selection. Here, the two directions are reviewed briefly.

Ji *et al.* [27] used the maximum relevance and minimal redundancy feature selection algorithm with weight reinforcement to classify rheumatoid arthritis medical images. Veeramuthu *et al.* [28] used a spatial gray-level difference algorithm and correlation-based feature selection method to select the most-important features for brain tumor recognition. Sudha *et al.* [29] used the improved lion optimization algorithm to choose feature subsets more efficiently and classify breast cancer with excellent accuracy. Kumar *et al.* [30] proposed a particle swarm optimization (PSO)-based rough set feature selection technique for obtaining a minimal set of relevant features. The selected features are applied to the classification of multiclass

motor imagery. Zhu *et al.* [31] combined feature selection and subspace learning methodological approaches to complete feature selection in a unified framework. Specifically, Mourragui *et al.* [20] utilized two subspace learning methods — linear discriminant analysis and locality preserving projection — that have proved their effectiveness in a variety of fields to select class-discriminative and noise-resistant features for Alzheimer’s disease (AD) diagnosis. In short, the above-mentioned single-modality feature selection methods can refine the original image features and improve the final recognition performance. However, the correlations among heterogeneous image features are not fully utilized.

Zhang and Shen [32] proposed a general methodology, multimodal multitask learning, to realize feature selection and multimodal feature fusion simultaneously. Zhou *et al.* [33] proposed a novel latent representation learning method that can utilize intermodality association for multimodality AD diagnosis. Kumar *et al.* [34] proposed a model that can improve the fusion of the complementary information in multimodality positron-emission-computed tomography with a supervised CNN that learns to fuse complementary information for multimodality medical image analysis. Zheng *et al.* [35] proposed a multimodality stacked deep polynomial network to fuse multimodality neuroimaging data and learn more-discriminative and more-robust feature representations for AD classification. Zu *et al.* [36] proposed a multimodality method for AD diagnosis. It completes adaptive feature selection and local similarity learning simultaneously. Hence, a similarity matrix is obtained for classification by considering heterogeneous modalities. Zhang *et al.* [37] trained two independent CNNs by multimodal medical images. The correlations among the two CNNs were used for neuropsychological diagnosis.

The above-mentioned feature selection algorithms [6], [7] have played important roles in the medical image analysis field. More-discriminant features are selected for different kinds of task. Nevertheless, only a few studies have been conducted to develop a breast cancer recognition model from the perspective of fine-grained feature selection. This new research direction was the focus of this study.

## C. RESEARCH MOTIVATIONS

Based on the above-mentioned analysis, the research motivations in this study can be divided into three because the RCA model is made up of three key components. The first is derived from the perspective of feature refinement. Because of noisy interference and high dimensions, the original image features are not “clean” and cannot accurately depict the lesion areas in whole mammograms. Hence, one goal is to refine the original features and use the retained discriminant information for breast cancer recognition. The second motivation is derived from deep-level cross-modal correlation mining. Features are “cleaner” after feature refinement. However, different visual features usually point to the same or similar pathological semantics in whole mammograms. Deep-level cross-modal pathological information that hides

among heterogeneous features can be mined and utilized to characterize whole mammograms properly. This further reduces the feature dimensions, which fit the sample size well and suppress the overfitting problem. The last motivation comes from the built-in adaptive feature selection function of the gradient boosting decision tree (GBDT) algorithm. A decision tree can guide a model to make the final feature selection adaptively and to boost recognition performance.

In summary, the RCA model is based on fine-grained feature selection, which combines feature refinement, cross-modal correlation mining, and adaptive feature selection into a holistic and versatile framework. This model uses low-dimensional but discriminant features to complete effective and efficient breast cancer recognition. Hence, it addresses the sample scarcity problem well. Moreover, it is a lightweight model deployed on a normal workstation. This helps to narrow the gap between theoretical research and practical application.

### III. PROPOSED BREAST CANCER RECOGNITION MODEL

#### A. MAIN FRAMEWORK

The proposed RCA model consists of several key components, including the first-layer feature selection (feature refinement), second-layer feature selection (cross-modal correlation mining), last-layer feature selection (decision tree guided adaptive feature selection), and breast cancer classification. The main framework of the RCA model is illustrated in Figure 1.

Seven well-known image features — SIFT (S), GIST (G), HOG (H), LBP (L), DenseNet (D), ResNet (R), and VGG16 (V) — are extracted to characterize entire mammograms from diverse visual perspectives, including shape, texture, and deep-level semantics. A fine-grained feature selection idea is proposed. A novel MvERGS algorithm is designed to complete feature refinement, which is defined as the first-layer feature selection. The refined features are denoted as the symbols  $\tilde{S}$ ,  $\tilde{G}$ , and  $\tilde{H}$ . Then, the implicit cross-modal correlations among the refined features are fully mined through the method of discriminant correlation analysis (DCA), which is defined as the second-layer feature selection. The cross-modal correlations are denoted as SG, SH, SV, and GV. For example, SG represents the cross-modal correlations among the S and G features. Finally, the decision tree embedded in the GBDT algorithm is used to guide the last-layer feature selection adaptively. The final selected feature is absorbed into the GBDT classifier to complete breast cancer recognition.

#### B. MODEL DETAILS

As illustrated in Figure 1, the RCA model contains three feature selection layers. The MvERGS algorithm is used to complete the first-layer feature selection. An attempt is made to refine the original image features and suppress the noisy information. The MvERGS algorithm derives from the traditional efficient range-based gene selection (ERGS)

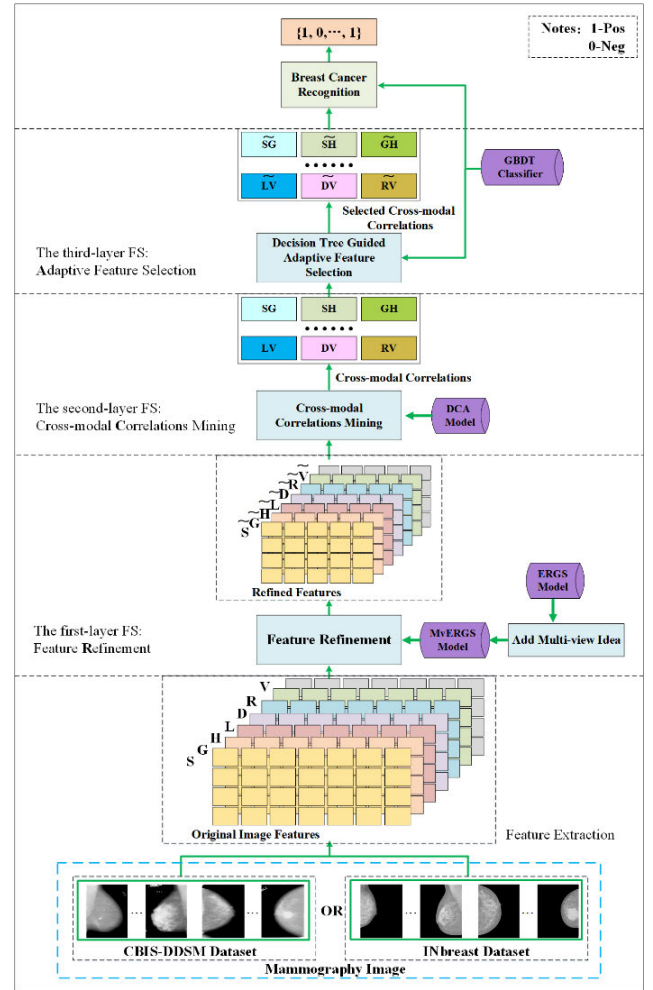


FIGURE 1. Main framework of the RCA model.

algorithm [38], but it is a further development of that algorithm. The new algorithm refines the original image features from two complementary views. Hence, more-discriminant but low-dimensional features are retained to handle the sample scarcity problem better.

The mammographic dataset is denoted as  $Mamm = \{x_1, x_2, \dots, x_n\}$ . It contains  $n$  samples. The feature set is denoted as  $F = \{f_1, f_2, \dots, f_d\}$ . It contains  $d$  features. The class label set is denoted as  $C = \{c_1, c_2\}$ . Moreover,  $\mu_{ij}$  and  $\sigma_{ij}$  represent the mean and standard deviation of feature  $f_i$  on category  $c_j$ , respectively. The effective range of feature  $f_i$  on category  $c_j$  is

$$R_{ij} = [r_{ij}^-, r_{ij}^+] = [\mu_{ij} - (1 - p_j) \gamma \sigma_{ij}, \mu_{ij} + (1 - p_j) \gamma \sigma_{ij}] \quad (1)$$

Here,  $r_{ij}^-$  and  $r_{ij}^+$  represent the lower and upper bounds of the effective range of feature  $f_i$  on category  $c_j$ , thus the overlapping (OA) area of category  $c_j$  is  $OA_j = r_{j1}^+ - r_{j2}^-$  respectively,  $p_j$  is the probability of category  $c_j$ , and  $\gamma$  is determined statistically by Chebyshev's inequality: 1.732.

1) FIRST VIEW OF THE MVERGS ALGORITHM

The proportion of feature overlapping regions in the effective range is considered. It is a relative value to measure the advantages and disadvantages of the original features, and it is not affected by the number of samples and the absolute size of the effective range. The smaller the proportion, the better the feature. This means that this feature can effectively distinguish heterogeneous samples. However, the larger the proportion, the worse the feature. This means that greater confusion can be observed among heterogeneous samples by applying this feature. Thus, the overlapping area ratio (OAR) of feature  $f_i$  is calculated based on the effective range in all class sets using the following equation.

$$E-OAR_i = \sum_{j=1}^{l-1} \sum_{k=j+1}^l \sum_{m=j,k} \frac{\varphi_i(j,k)}{R_{im}} \quad (i = 1, 2, \dots, d) \quad (2)$$

$$\varphi_i(j,k) = \begin{cases} r_{ij}^+ - r_{ij}^-, & \text{if } r_{ij}^+ > r_{ij}^- \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

Because the dataset includes two categories,  $E-OAR_i = OA_i / (r_{i1}^+ - r_{i1}^-) + OA_i / (r_{i2}^+ - r_{i2}^-)$ . The  $E-OAR_i$  of each feature  $f_i$  is normalized, and its weight  $Ew_i$  is calculated for all samples:

$$Ew_i = 1 - \frac{E-OAR_i}{\text{Max}\{E-OAR_t | t = 1, 2, \dots, d\}} \quad (4)$$

Next,  $\text{DiffLabel}_i(x_j) = \{x_p | x_p \in k-Neighbor_i(x_j) \wedge \text{Label}(x_p) \neq \text{Label}(x_j)\}$  is denoted as the sample set in which the label is different from sample  $x_j$  on feature  $f_i$ . Moreover,  $k-Neighbor_i(x_j)$  is the  $k$ -nearest neighbor of sample  $x_j$  on feature  $f_i$ . Equation (5) is used to calculate the overlapping regions based on the proportion of heterogeneous samples in the neighbors:

$$k-OAR_i(x_j) = \frac{|\text{DiffLabel}_i(x_j)|}{k} \quad (5)$$

Here,  $k-OAR_i(x_j)$  represents the proportion of samples with different class labels from  $x_j$  among the  $k$  neighbors of sample  $x_j$  on feature  $f_i$ . When the overlapping region of feature  $f_i$  is calculated, Equation (6) is used to calculate the average OAR of the sample space on feature  $f_i$  based on the proportion of heterogeneous samples in the neighbors.

$$Ak-OAR_i = \frac{\sum_{j=1}^n k-OAR_i(x_j)}{n} \quad (6)$$

The greater  $Ak-OAR_i$ , the greater the confusion degree, and the weaker the discriminant ability of the features. The weight  $Kw_i$  of feature  $f_i$  is calculated based on  $NAk-OAR_i$ :

$$Kw_i = 1 - \frac{NAk-OAR_i}{\text{Max}\{Ak-OAR_t | t = 1, 2, \dots, d\}} \quad (7)$$

Next,  $Kw$  and  $Ew$  are considered comprehensively. Then, the parameter  $\alpha$  that belongs to  $[0, 1]$  is used to tune these two

weights and obtain the following feature weight:

$$N-Fw_i = \frac{Fw_i}{\text{Max}\{Fw_t | t = 1, 2, \dots, d\}} = \frac{\alpha Ew_i + (1-\alpha) Kw_i}{\text{Max}\{Fw_t | t = 1, 2, \dots, d\}} \quad (8)$$

According to Equation (9), the parameter  $\theta$  is used as a threshold to perform feature selection.

$$FR_1 = \{N-Fw_i | N-Fw_i > \theta\} \quad (9)$$

where  $N-Fw_i$  is proportional to the importance of the corresponding feature. If it is greater than  $\theta$ , the corresponding feature is chosen and moved into a queue  $FR_1$ . A new refined image feature,  $F_m$ , is generated based on  $FR_1$ .

2) SECOND VIEW OF THE MVERGS ALGORITHM

The overlapping region in the effective range, which is an absolute value and derives from the traditional ERGS algorithm, is considered. Similar to the above-mentioned first view, the smaller the overlapping region, the better the feature. Hence, this view is a useful complement to the first view. The weight  $w_i$  of feature  $f_i$  is calculated, and weight matrix  $W$  is obtained. Then, Equation (10) is used to normalize matrix  $W$  and obtain  $N-W_i$ :

$$N-W_i = \frac{W_i}{\text{Max}\{W_t | t = 1, 2, \dots, d\}} \quad (10)$$

Similar to Equation (9), the parameter  $\theta$  is used as a threshold to perform feature selection and generate the refined feature  $F_e$ .

Finally, the two complementary views of the MVERGS algorithm are combined by concatenating the  $F_m$  and  $F_e$  features. Thus, the new feature  $F'$  is obtained. It contains the key components of the original features. This ensures the integrity of the effective information in the original feature space (see Figures 5 and 6 in Section IV (C)). Moreover, because of its low dimension, the  $F'$  feature can also improve the real-time efficiency of the recognition model and build a strong foundation for the subsequent cross-modal correlation mining. Most importantly, the  $F'$  feature fits the sample size well after dimension reduction, thereby alleviating the data scarcity problem. In summary, the novel ERGS algorithm has two apparent advantages. First, it employs two complementary views to refine the original features and improves their discriminant abilities. This proves the scalability of the MVERGS algorithm, which means that more views can be absorbed into this algorithm to improve its effectiveness further. Second, it only processes the bottom feature components and does not depend on the high-level visual content. This proves the robustness of the MVERGS algorithm, showing that it can effectively process any feature in any research field. Hence, the MVERGS algorithm can be used in some downstream research fields that need elaborate feature selection. These advantages are demonstrated in Section V.

As described above, the MvERGS algorithm tries to refine the original features. Noisy information is suppressed to a certain degree. Hence, the refined features contain much valuable pathological semantics information. However, each feature characterizes mammography images from its own visual perspective. Therefore, it is insufficient to improve the final recognition performance. The deep-level cross-modal pathological correlations hidden among heterogeneous image features, which can characterize the lesion areas properly, have not been explored. The visual appearance of breast masses, including texture, shape, color, and edge, usually point to the same or a similar lesion area. Therefore, heterogeneous image features contain plentiful cross-modal correlations (heterogeneous features can be regarded as different “modalities,” so cross-modal correlation mining can be performed to generate more-effective features), which helps to improve the final recognition performance. Hence, after the first-layer feature refinement, the second-layer feature selection is implemented through the DCA method. The aim is to identify the cross-modal pathological correlations among heterogeneous image features to characterize whole mammograms more accurately.

The number of mammography images is  $n$ , and the number of categories is  $C = 2$ . The dimension of the feature vector  $\mathbf{X}$  (or  $\mathbf{Y}$ ) is decomposed into  $c$  parts so that sample  $m_i$  belongs to the  $i$ th category — that is,  $m = \sum_{i=1}^c m_i$ .

The average value of the  $j$ th sample  $x_{ij}$  of the  $i$ th image and the average value of all samples  $\bar{x} = \frac{1}{m} \sum_{i=1}^c \sum_{j=1}^{m_i} x_{ij} = \frac{1}{m} \sum_{i=1}^c m_i \bar{x}_i$  are calculated. Then, Equation (11) is used to calculate interclass divergence matrix  $\mathbf{S}_{bx}$ .

$$\mathbf{S}_{bx(p \times p)} = \sum_{i=1}^c m_i (\bar{x}_i - \bar{x}) (\bar{x}_i - \bar{x})^T = \boldsymbol{\varphi}_{bx} \boldsymbol{\varphi}_{bx}^T \quad (11)$$

where  $\boldsymbol{\varphi}_{bx(p \times c)} = [\sqrt{m_1}(\bar{x}_1 - \bar{x}), \dots, \sqrt{m_c}(\bar{x}_c - \bar{x})]$ . The interclass divergence matrix is diagonalized using Equation (12):

$$\mathbf{P}^T (\boldsymbol{\varphi}_{bx}^T \boldsymbol{\varphi}_{bx}) \mathbf{P} = \hat{\boldsymbol{\Lambda}} \quad (12)$$

Here,  $\mathbf{P}$  is an orthogonal eigenvector matrix, and  $\hat{\boldsymbol{\Lambda}}$  is a diagonal matrix with eigenvalues in descending order. In addition,  $\mathbf{Q}_{(c \times c)}$  contains  $r$  feature vectors, and there is a correlation with the  $r$  largest nonzero features of matrix  $\mathbf{P}$ ; therefore,

$$(\boldsymbol{\varphi}_{bx} \mathbf{Q})^T \mathbf{S}_{bx} (\boldsymbol{\varphi}_{bx} \mathbf{Q}) = \boldsymbol{\Lambda}_{(r \times r)} \quad (13)$$

Then, conversion matrix  $\mathbf{W}_{bx} = \boldsymbol{\varphi}_{bx} \mathbf{Q} \boldsymbol{\Lambda}^{-\frac{1}{2}}$  is used to unitize  $\mathbf{S}_{bx}$  and reduce feature matrix  $\mathbf{X}$  from  $p$  to  $r$ . (It is the same for feature matrix  $\mathbf{Y}$ .)

$$\mathbf{X}'_{(r \times m)} = \mathbf{W}_{bx}^T \mathbf{X}_{(p \times m)} \quad (14)$$

where  $\mathbf{I} = \mathbf{W}_{bx}^T \mathbf{S}_{bx} \mathbf{W}_{bx}$  is the interclass divergence matrix, and the singular value decomposition technique is used to

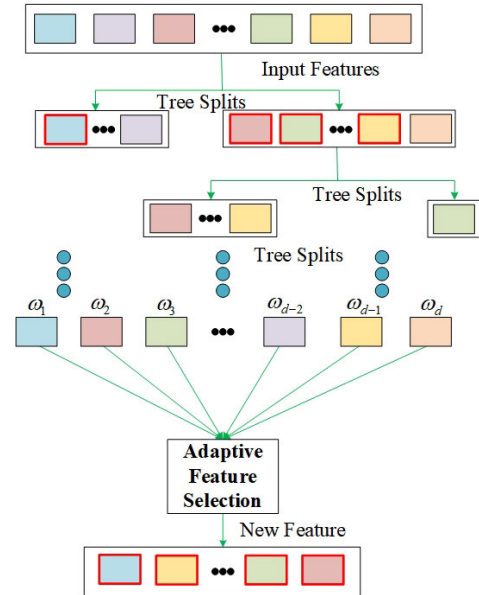


FIGURE 2. Workflow of the decision-tree-guided adaptive feature selection.

diagonalize interclass covariance matrix  $\mathbf{S}'_{xy}$ :

$$(\mathbf{U} \sum^{-\frac{1}{2}})^T \mathbf{S}'_{xy} (\mathbf{V} \sum^{-\frac{1}{2}}) = \mathbf{I} \quad (15)$$

Finally, covariance matrix  $\mathbf{S}'_{xy}$  is used to map the original image features  $\mathbf{X}$  and  $\mathbf{Y}$  to the intermediate space:

$$\mathbf{X}'' = \mathbf{W}_{cx}^T \mathbf{X}' = \underbrace{\mathbf{W}_{cx}^T \mathbf{W}_{bx}^T}_{\mathbf{X}} = \mathbf{W}_x \mathbf{X} \quad (16)$$

$$\mathbf{Y}'' = \mathbf{W}_{cy}^T \mathbf{Y}' = \underbrace{\mathbf{W}_{cy}^T \mathbf{W}_{by}^T}_{\mathbf{Y}} = \mathbf{W}_y \mathbf{Y} \quad (17)$$

Based on the  $\mathbf{X}''$  and  $\mathbf{Y}''$  matrices, a new cross-modal feature vector  $\mathbf{F}''$  can be generated by the concatenation (or summation) operation.

Because there are too many combinations, many cross-modal correlations are obtained in the second-layer feature selection procedure. To improve the real-time efficiency, an attempt is made to use the implicit feature selection characteristic of the GBDT algorithm to compress the number of image features. Hence, the decision tree in the GBDT algorithm guides the RCA model to complete the last-layer feature selection adaptively. Figure 2 briefly illustrates this idea. In Figure 2, features enclosed within the red rectangles are chosen by the decision tree in the GBDT algorithm. They are concatenated automatically for breast cancer recognition. Feature dimensions are reduced to a certain range simultaneously.

### C. PROPOSED RCA MODEL

Based on the above analysis, the RCA model is as follows.

**Algorithm 1** RCA Model**Input:** the original image feature called  $F \in \{S, G, H, L, R, D, V\}$ **Output:** the novel feature called  $F'''$ 

- (1) **Repeat**
- (2) Calculate the weight  $E_w$  of each  $F$  based on Equation (4)
- (3) Calculate the weight  $K_w$  of each  $F$  based on Equation (7)
- (4) Calculate the weight  $N-W_i$  of each  $F$  based on Equation (10)
- (5) **Until** each original image feature has been processed
- (6) Obtain refined image feature called  $F' \in \{\tilde{S}, \tilde{G}, \tilde{H}, \tilde{L}, \tilde{R}, \tilde{D}, \tilde{V}\}$
- (7) **Repeat**
- (8) Choose a refined feature  $X$  from  $F'$
- (9) **Repeat**
- (10) Choose another different refined feature  $Y$  from  $F'$
- (11) Map  $X$  and  $Y$  into a new space based on Equations (16) and (17)
- (12) Obtain  $X''$  and  $Y''$ , use the concatenation mode to build the feature  $F''$
- (13) **Until** each refined feature has been processed
- (14) **Until** each refined feature has been processed
- (15) Use  $F''$  and the decision-tree-guided method (Figure 2) to adaptively create the novel feature  $F'''$ .

**IV. EXPERIMENTAL RESULTS AND ANALYSIS****A. DATASETS AND BASELINES**

## 1) DATASETS

Two well-known mammographic datasets, CBIS-DDSM and INbreast, were used for the experiments. The details of these two datasets are given in Table 1.

**TABLE 1.** Details of the CBIS-DDSM and INbreast datasets.

Dataset	File Format	Size after Preprocessing	Nega-tive	Positive	Train-Test Ratio
CBIS-DDSM [39]	PNG	1152×896	1434	1347	7:3
INbreast [40]	PNG	2500×3300	287	100	

For the INbreast dataset, the same setting of reference [25] was used: all the mammograms labeled 1 and 2 were regarded as negative samples, whereas mammograms 4, 5, and 6 were regarded as positive samples. Owing to the lack of a definite category for label 3, the corresponding mammograms were ignored in the experiments. Hence, as shown in Table 1, the INbreast dataset is imbalanced, making recognition more difficult. For the CBIS-DDSM dataset, the SIFT and HOG features were reduced to 500 dimensions, respectively. For the INbreast dataset, these two features were reduced to 500 and 300 dimensions, respectively. To better extract the corresponding deep-learning-based features, the correlated research works [25] were followed, and each mammogram was resized to  $224 \times 224$ , which fits the input size of the CNN model well. The last fully connected layer of the VGG 16 model was regarded as the VGG feature (4096 dimensions). The last average pooling layer of

the DenseNet 161 model was regarded as the DenseNet feature (2208 dimensions). Similar to the DenseNet feature, the corresponding ResNet feature (2048 dimensions) was extracted from the last average pooling layer of the ResNet 50 model.

## 2) BENCHMARK MODELS

The proposed RCA model was compared with five kinds of benchmark model.

(1) Two variants of the RCA model were used for direct comparisons, wherein each feature selection layer is added gradually into the RCA model. Thus, the RCA and RCA models were obtained. This is another ablation analysis mode of the proposed RCA model.

(2) Prevailing deep-learning models for direct comparisons were used: VGG16 [13], ResNet152 [14], and DenseNet121 [41].

(3) State-of-the-art feature selection models were used for direct comparisons: Fisher score [42], ERGS [27], PSO [19], hypergraph-based sparse canonical correlation analysis (HGSCCA) [43], and GS-XGBoost [44].

(4) State-of-the-art breast cancer recognition models were used for direct comparisons: Shen's model [25], Dhungel's model [27], and Zhang's DE-Ada\* model [45].

(5) Recent region of interest (ROI)-based breast cancer recognition models were used for indirect comparisons: Tsochatzidis's model [46], Rampun's model [47], AlexNet + sparse multiple instance learning [48], and Carneiro's model [49].

The reproduction of each benchmark model follows the experimental settings in a previous report as much as possible; however, considering the difference in the operating computer environment, there may be a certain deviation.

## 3) EVALUATION METRICS

A CAD system should be evaluated comprehensively from diverse perspectives. Hence, similar to the state-of-the-art studies [45] - [49], several specific methods were used to evaluate the model. The two most important evaluation metrics include accuracy (Acc) and area under the receiver operating characteristic curve (AUC). Higher Acc (AUC) means that better recognition performance is obtained.

$$Acc = \frac{TP + TN}{TP + FP + TN + FN} \quad (18)$$

where, TP is the number of correctly classified positive samples, TN is the number of correctly classified negative samples, FP is the number of misclassified positive samples, FN is the number of misclassified negative samples.

At the same time, the AUC is used as well. Similar to Acc, a larger AUC value indicates satisfactory performance, which also means that the corresponding ROC curve is close to the (0, 1) point and far from the  $45^\circ$  diagonal of the coordinate axis. The AUC metric can provide objective evaluations in response to the data imbalance problem — specifically, in the INbreast dataset. Meanwhile, other metrics

**TABLE 2.** Performance comparisons between the RCA model and all baselines in CBIS-DDSM. (Note: the best result of each metric is shown as **93.30**. The unit is %. “/” indicates that the corresponding work did not provide the result.)

(Note: the best result of each metric is shown as **93.30**. The unit is %. “/” indicates that the corresponding work did not provide the result.)

Setting	Model	Acc	AUC	Setting	Model	Acc	AUC
85-15 Whole	RESNET-RESNET [25]	/	87.00	70-30 Whole	GS-XGBoost [46]	78.71	85.94
	RESNET-VGG [25]	/	88.00		VGG16 [13]	50.46	51.21
	VGG-VGG [25]	/	86.00		ResNet152 [14]	50.46	53.22
	VGG-RESNET [25]	/	88.00		DenseNet121 [41]	50.69	54.69
	Model Averaging [25]	/	91.00		Fisher Score [42]	80.50	90.65
	GS-XGBoost [44]	76.99	72.81		ERGS [38]	79.31	93.73
	DE-Ada* [45]	87.05	92.19		PSO [30]	74.40	83.36
	Original (S)	77.99	86.32		HGSCCA [43]	53.23	50.00
	RCA	76.31	86.59		DE-Ada* [45]	90.91	<b>98.36</b>
	RCA	76.55	84.42		Original (S)	80.62	95.00
ROI	RCA	76.07	85.44	RCA	80.86	97.00	
	Tsochat [46]	74.90	80.40	RCA	91.39	97.15	
	Rampun [47]	80.40	84.00	RCA	<b>93.30</b>	97.22	

are needed, including sensitivity (Sen), specificity (Spe), and precision (Pre), to evaluate the RCA model from the perspective of practicality.

$$Sen = \frac{TP}{TP + FN} \quad (19)$$

$$Spe = \frac{TN}{TN + FP} \quad (20)$$

$$Precision = \frac{TP}{TP + FP} \quad (21)$$

A higher Sen means that the corresponding FNR is low and the number of misdiagnoses is reduced. A higher Spe means that the corresponding FPR is low, and the probability of diagnosis is higher. Moreover, a Kiviat-diagram-based evaluation metric [45] was used to evaluate the overall performance of each recognition model more comprehensively. The Acc, Sen, Spe, and AUC values of each model are plotted in the Kiviat diagram. Each of the metrics belongs to the set  $I = \{\text{Acc}, \text{Sen}, \text{Spe}, \text{AUC}\}$ . The area of the Kiviat diagram (AKD) of each model is calculated according to Equation (22). The highest AKD demonstrates the best overall recognition performance.

$$AKD = \frac{1}{2} \sin\left(\frac{2\pi}{len}\right) \times \left(\sum_{i=1}^{len-1} a_i \times a_{i+1} + a_1 \times a_{len}\right) \quad (22)$$

where  $a_i \in \{\text{Acc}, \text{Sen}, \text{Spe}, \text{AUC}\}$ ,  $len = 4$ .

## B. PERFORMANCE COMPARISONS WITH STATE-OF-THE-ART MODELS

As described in Section IV(A) 2), it is necessary to make direct and indirect comparisons to demonstrate the effectiveness of the RCA model. The corresponding experimental results of different datasets are exhibited in Tables 2 and 3, respectively. To make fair comparisons with the references [25], their train-test settings were used. For example,

in the CBIS-DDSM dataset, the corresponding experiments of the 85-15 train-test setting were added. The related experimental results are exhibited in the upper left (UL) part of Table 2. Meanwhile, the corresponding comparisons of the 70-30 train-test setting are shown in the upper right (UR) part of each table. The bottom left (BL) part of Table 2 exhibits the ROI-based models for indirect comparisons in the CBIS-DDSM dataset. Similarly, the bottom right (BR) part of Table 3 exhibits ROI-based models for indirect comparisons in the INbreast dataset. All the deep-learning models [13], [14], [41] obeyed the 70-30 train-test setting. Parameters  $\alpha$  and  $\theta$  in the MvERGS were set as 0.5 and 0.7 respectively. “Original (S)” refers to the original SIFT feature that obtains the best performance among all the original image features.

As shown in Table 2, first, the RCA model outperforms all the deep-learning models with a large performance margin. Obviously, the deep-learning models overfit to a certain degree owing to the sample scarcity problem. Second, the RCA (MvERGS) model outperforms most baselines, especially the traditional ERGS algorithm [38]. As a variant of the ERGS algorithm, the novel MvERGS algorithm is more scalable. It employs two complementary views to refine the original features and improves their discriminant abilities. Thus, the MvERGS algorithm can be used in some downstream research fields that need elaborate feature selection. It can also be regarded as a valuable by-product of the RCA model. Third, the RCA model outperforms some state-of-the-art feature selection algorithms, such as GS-XGBoost [44], Fisher score [42], HGSCCA [43], and PSO [30]. The model ensures the integrity of the effective information in the original feature space, indicating that it can effectively process any feature in any research field. Meanwhile, steady performance improvements can be observed after each feature selection layer (R, C, and A) is gradually



**TABLE 3. Performance comparisons between the RCA model and all baselines in INbreast. (Note: the best result of each metric is shown as **98.00**. The unit is %. “/” indicates that the corresponding work did not provide the result.)**

(Note: the best result of each metric is shown as **98.00**. The unit is %. “/” indicates that the corresponding work did not provide the result.)

Setting	Model	Acc	AUC	Setting	Model	Acc	AUC
	RESNET-RESNET [25]	/	95.00		ERGS [38]	81.89	78.41
	RESNET-VGG [25]	/	95.00		PSO [30]	79.31	80.43
	VGG-VGG [25]	/	95.00		HGSCCA [43]	78.56	50.00
	VGG-RESNET [25]	/	95.00	70-30	Original (D)	78.45	72.56
70-30	Model Averaging [25]	/	<b>98.00</b>	Whole	DE-Ada* [45]	<b>87.93</b>	92.65
Whole	GS-XGBoost [44]	85.35	82.84		RCA	85.34	83.08
	ResNet [14]	75.86	54.14		RCA	87.07	93.00
	VGG [13]	75.86	52.52		RCA	87.07	93.00
	DenseNet [41]	75.86	62.01				
	Fisher Score [42]	81.03	88.51	ROI	SMIL [48]	90.00±2	89.00±3
					Carneiro [49]	/	86.00

absorbed into the RCA model, especially in the UR part. Clearly, any feature selection layer contributes to boosting the final recognition performance. As another important finding, new features, including SD, SR, and SH (please refer to Fig. 1), play more-important roles in the recognition procedure. The implicit complementarity among the state-of-the-art traditional and deep-learning-based features is fully mined by the RCA model to promote the final recognition performance. Fourth, although Shen *et al.* [25] used the ensemble-learning method (model averaging) to obtain the best AUC (91%), the proposed model (86.59%) almost outperforms a single model of Shen (86%) in the UL part. The RCA model is competitive and robust in the 85-15 train-test setting. The RCA model has the best Acc (93.30%) and suboptimal AUC (97.22%) in the 70-30 train-test setting, whereas the corresponding Acc improvement (2.39%) is more evident in the UR part. Fewer data are necessary to train the model effectively. Hence, the RCA model handles the sample scarcity problem well. Finally, the model completely outperforms the ROI-based recognition models. As previously mentioned, there are several obvious advantages of breast cancer recognition in a whole mammogram, including improving the utilization of context information in whole mammograms, saving annotation cost, and representing the real clinical process.

Because of the multistage and layer-by-layer refinement characteristics of the RCA model, it can screen out the most discriminative features to handle the sample scarcity problem well and achieve satisfactory recognition performance in the CBIS-DDSM dataset.

As Table 3 shows, first, the RCA model outperforms all the deep-learning models in the INbreast dataset. Second, the RCA (MvERGS) model is quite competitive compared with most baselines. It can maximumly retain the key discriminant information in the original feature space, and each view of the MvERGS algorithm is useful for breast cancer recognition. Hence, the innovative algorithm is a valuable by-product of the RCA model. Third, the RCA model is superior to all the feature selection algorithms

in both metrics. Obvious recognition performance improvements can be observed after the gradual addition of diverse feature selection methods. The proposed fine-grained feature selection idea is effective. Fourth, the RCA model obtains a suboptimal Acc (87.07%), which is close to the DE-Ada\* model [45] (87.93%). From the AUC perspective, the RCA model (93%) beats the DE-Ada\* model and approaches Shen’s single model [25] (95%). Meanwhile, new features generated by the proposed model, including SD and SR, play significant roles in the RCA model. The implicit pathological semantic information among heterogeneous features is fully mined and used to improve the final performance. Objectively, the AUC metric of the RCA model should be improved further. A dramatic decline in dimensions may be the key factor that affects the final performance. Compared with the CBIS-DDSM dataset, the INbreast dataset contains mammograms with high quality. Dimensions that are too low destroy the discriminant abilities of the generated features. How to make a trade-off between feature dimension and recognition performance is a planned focus point of future work. Finally, similar to Table 2, the RCA model outperforms all the ROI-based recognition models. The RCA model achieves satisfactory recognition performance in the INbreast dataset.

In summary, the proposed RCA model is effective for breast cancer recognition on both benchmark datasets. It is a fine-grained feature selection model that focuses on mining the deep-level pathological information contained in the limited labeled samples from the shallower to the deeper. The deep-level pathological information is more discriminant but lower-dimensional than before. One can use it to train a more powerful classification model. Most importantly, this helps to address the data scarcity problem well. As another important finding, the RCA model does well on low-quality mammograms (CBIS-DDSM). Much more noisy information in these mammograms can be suppressed to a certain degree by the RCA model. Moreover, the MvERGS algorithm is scalable and robust, and it can be regarded as a valuable by-product of the RCA model.

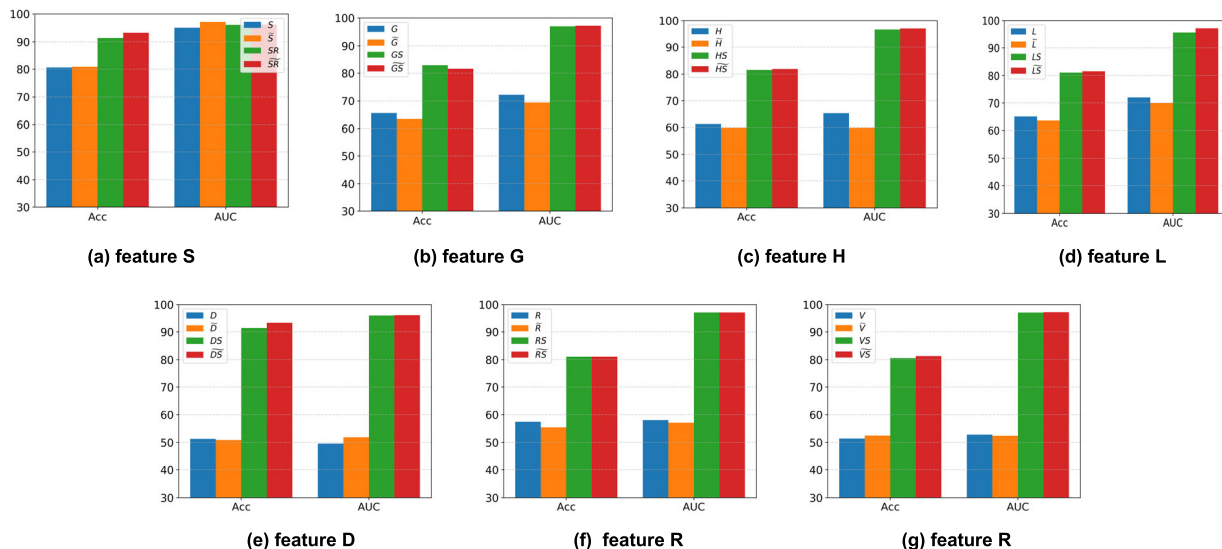


FIGURE 3. Performance comparisons of feature selection layers (CBIS-DDSM).

Finally, owing to the simplicity of the model, a normal workstation can be employed to deploy it, demonstrating its high practicality.

C. ROBUST VALIDATION OF THE NOVEL FEATURES

1) RECOGNITION PERFORMANCE

As illustrated in Figure 1, the MvERGS algorithm was first designed for the first-layer feature selection. Then, the cross-modal correlations among image features were mined by the second-layer feature selection. Finally, the adaptive feature selection strategy of the GBDT algorithm was used to obtain the final features. The final features were used to implement breast cancer recognition. To demonstrate the effectiveness of the novel features, the Acc and AUC values of each feature in each feature selection layer were computed, resulting in Figures 3 (CBIS-DDSM) and 4 (INbreast), respectively. For example, in Figure 3(a), S represents the original image feature SIFT.  $\tilde{S}$  represents the refined image feature of SIFT generated by the MvERGS algorithm. The best cross-modal correlation that includes the corresponding single feature was chosen — for example, SR represents the best cross-modal correlation that includes S and R. DS represents the best cross-modal correlation that includes the corresponding refined feature of D and S, respectively.  $\tilde{SR}$  represents the final feature generated by the adaptive decision tree embedded in the GBDT algorithm.

As shown in Figure 3, for these two metrics, with the deep-level feature selection, a gradually improved trend can be observed easily. Meanwhile, the new features generated by the RCA model (i.e.,  $\tilde{SR}$ ) almost outperform all the other features. For the AUC metric in particular, a very large performance gap between the first and last layers can be observed in most subfigures. These results demonstrate the effectiveness of the proposed fine-grained feature selection idea. It can gradually mine much more discriminant

information from heterogeneous image features to improve recognition performance.

Another important phenomenon is that the original high-dimensional features lead to the overfitting problem (it is more evident on the deep learning-based features, such as V and D). This phenomenon should be considered carefully. However, lower-dimensional but more-discriminant features generated by the RCA model can cope with the overfitting problem well (for example, the  $\tilde{SR}$  feature obtains the best AUC value). The overfitting problem can be suppressed to a certain degree when the feature dimension fits the sample size well. This finding also shows that one should take full advantage of the traditional and state-of-the-art deep-learning-based features to boost the final recognition performance. More-discriminant information can be mined among these features.

The SIFT feature obtains the best recognition performance among all the single features, meaning that shape variation is the most important factor to characterize the key visual appearance of the lesion areas in whole mammograms. Compared with a benign breast mass, a malignant breast mass usually has some abnormal shape variation. Most importantly, the S (SIFT) feature plays a significant role in the subsequent feature selection procedure. For example, the  $\tilde{SR}$  feature obtains the best overall recognition performance among all the features. Moreover, S (or R) is the best choice among all the traditional (or deep-learning-based) features. The implicit valuable complementary information hides between the S and R features. The proposed second-layer feature selection algorithm is better qualified for the cross-modal correlation mining task. It not only maximizes the correlations of the corresponding feature components between two heterogeneous features but also weakens the correlations of the feature components belonging to different categories within the homogeneous features. Hence, the mapped

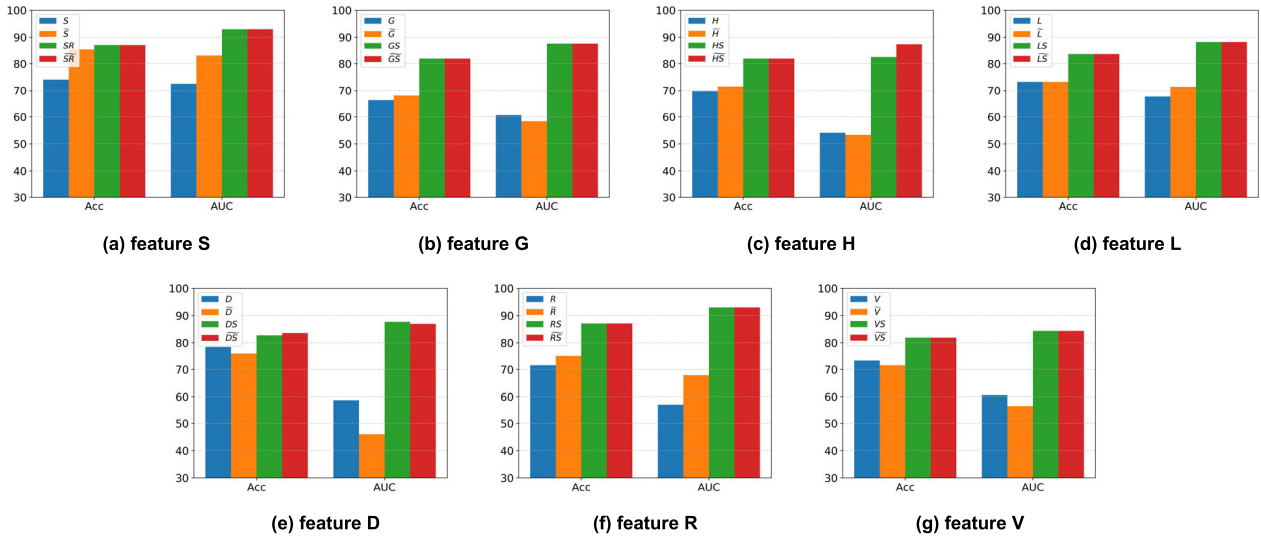


FIGURE 4. Performance comparisons of feature selection layers (INbreast).

features are more discriminative and compact, serving as important foundations for breast mass classification. Because of the power discriminant ability of the cross-modal correlations, only a slight performance improvement can be obtained in the last-layer feature selection.

Similar to Figure 3, with the deep-level feature selection, a gradually improved trend can be observed in Figure 4. The trend is more evident on the S, L, and R features, which can demonstrate the robustness of the proposed RCA idea. The MvERGS algorithm produces a marked effect on some original image features, including the S, L, and R features. As analyzed above, the MvERGS algorithm can maximally retain the discriminant information in the original feature space because of its multiview characteristics. Moreover, not only the sparse features, such as S and L, but also the dense features, such as R and D, can be processed well by the MvERGS algorithm, building a strong foundation for the subsequent feature selection stage. Similar to Figure 3, large performance improvements can be observed when the cross-modal correlations are used, especially for the AUC metric. Clearly, the implicit complementary information among the refined image features is valuable for breast cancer recognition. The proposed second-layer feature selection strategy can mine the maximum correlations among two heterogeneous features.

The overfitting problem also occurs in the INbreast dataset when the deep-learning-based features are used. (Please refer to the R and V features. Surprisingly, as shown in Table 1, although training samples are scarce in the INbreast dataset, the overfitting problem is not serious compared with that in the CBIS-DDSM dataset. This may mainly result from the high quality of the mammograms in the INbreast dataset. Experiments are planned to test this supposition.) Hence, fine-grained feature selection is necessary to handle the overfitting problem. Among all the features, the  $\tilde{SR}$  feature obtains the best overall recognition performance. Both the

shape variation and deep-level pathological semantics can provide more-effective breast cancer recognition.

## 2) VISUALIZATION RESULTS OF THE FEATURES

To demonstrate intuitively the effectiveness of the novel features generated by the RCA model, the well-known t-SNE tool [50] was used for feature visualization. Some representative results are illustrated in Figures 5 and 6. As previously established, S, G, R, and D denote the corresponding original image features, whereas  $\tilde{S}$ ,  $\tilde{G}$ ,  $\tilde{R}$ , and  $\tilde{D}$  denote the corresponding refined features generated by the MvERGS algorithm. SG, SD, SR, and GD denote the new features generated by cross-modal correlation mining.

First, the results of three representative original image features are shown. Then, the results of the corresponding refined feature of each original feature are presented. Finally, the results of the three best cross-modal correlations, including the representative original features, are shown. The visualization results of the final features output by the last layer are not exhibited because this layer only compresses the number of features rather than creating novel features. Hence, nine subfigures of each dataset were obtained.

As shown in Figures 5(a)–(c), the aggregation degree of the samples with the same labels is not high, whereas there are serious confusions between different kinds of sample when the original image features are used. This means that the original image features have too much noisy information, which adversely affects the final recognition performance (see Figure 3). In other words, owing to sample scarcity and high dimensions, a very complex rather than robust decision surface (boundary) is needed for breast cancer recognition, which causes the overfitting problem. Hence, one should reduce the feature dimensions but retain the discriminant information in the original feature space. However, the proposed fine-grained feature selection idea can address this issue.

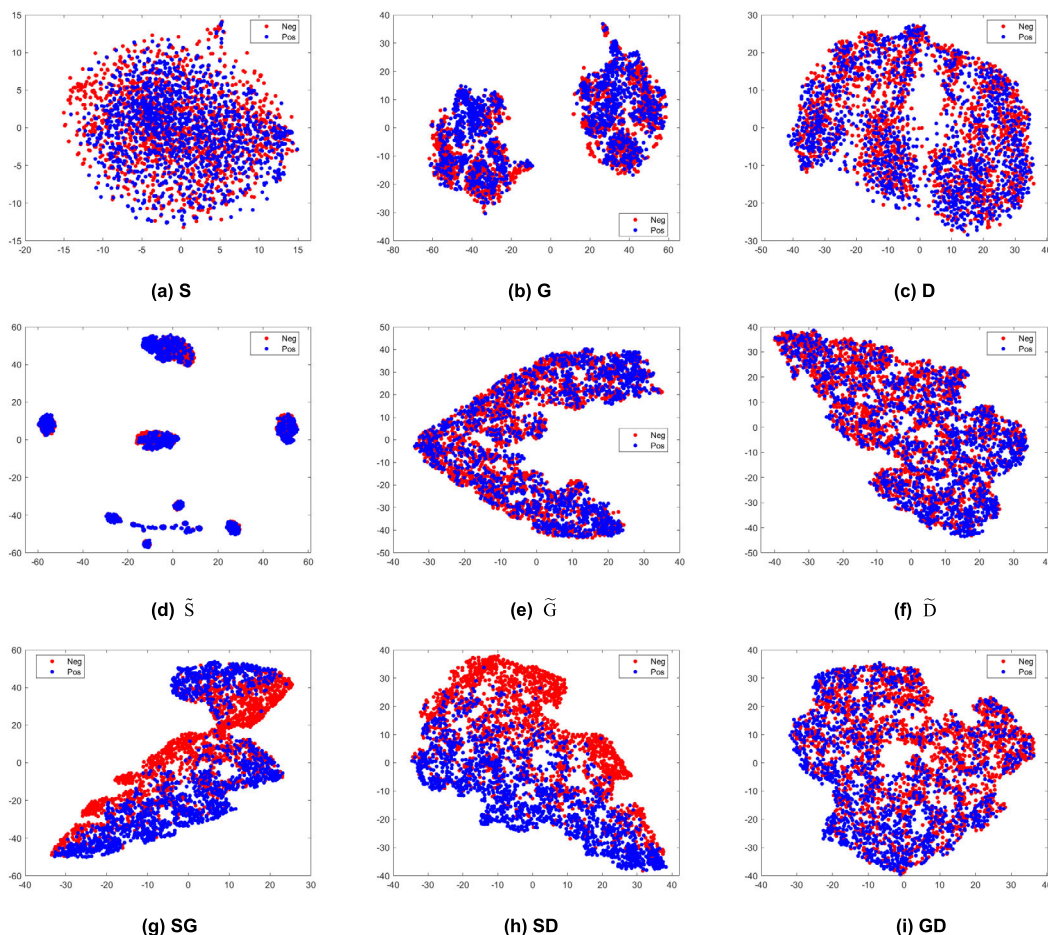


FIGURE 5. t-SNE results in the DDSM dataset.

Second, with the application of the MvERGS algorithm, as in Figure 5(e), although confusion still occurs, a denser aggregation degree of the samples with the same labels can be observed, whereas the confusion degree among different kinds of sample gradually decreased. This implies that the proposed MvERGS algorithm can refine the original image features, and the refined features are more discriminant for effective and robust recognition. Unlike the traditional ERGS algorithm, the MvERGS algorithm employs two complementary views to complete feature selection effectively.

Third, with the application of cross-modal correlation mining, as in Figure 5(g), the positive samples are clustered into a denser and more regular cluster, and the negative samples also form a large cluster. Most importantly, there is an obvious “decision surface (boundary)” between the two types of sample. The second-layer feature selection is quite powerful. It not only maximizes the correlations of the corresponding feature components between two heterogeneous features but also weakens the correlations of the feature components belonging to different categories within the homogeneous features. Hence, more-notable and more-robust decision boundaries can be easily obtained for effective breast cancer recognition.

As shown in Figure 6, similar experimental phenomena can be observed in the INbreast dataset. With the deep-level feature selection, the samples with the same label start to form a denser cluster, which helps to train more-robust decision boundary and improve the final recognition performance. Compared with the CBIS-DDSM dataset, clearer clusters can be observed in Figure 6. Fewer samples may be the primary reason.

In summary, the proposed RCA model is effective and robust for breast cancer recognition. Owing to lower dimensions, the RCA model is also efficient. It reduces the computing resources needed and can be deployed on a normal workstation.

### 3) STATISTICAL SIGNIFICANCE TEST

The statistical significance test is an effective method that can be used to analyze the real classification performance carefully. It was used to evaluate the proposed RCA model from the statistical perspective and provide more-objective conclusions. Hence, the t-test was used to demonstrate whether the performance improvement of the proposed RCA model is an essential improvement. The hypothesis is that classification performance improvement is an essential

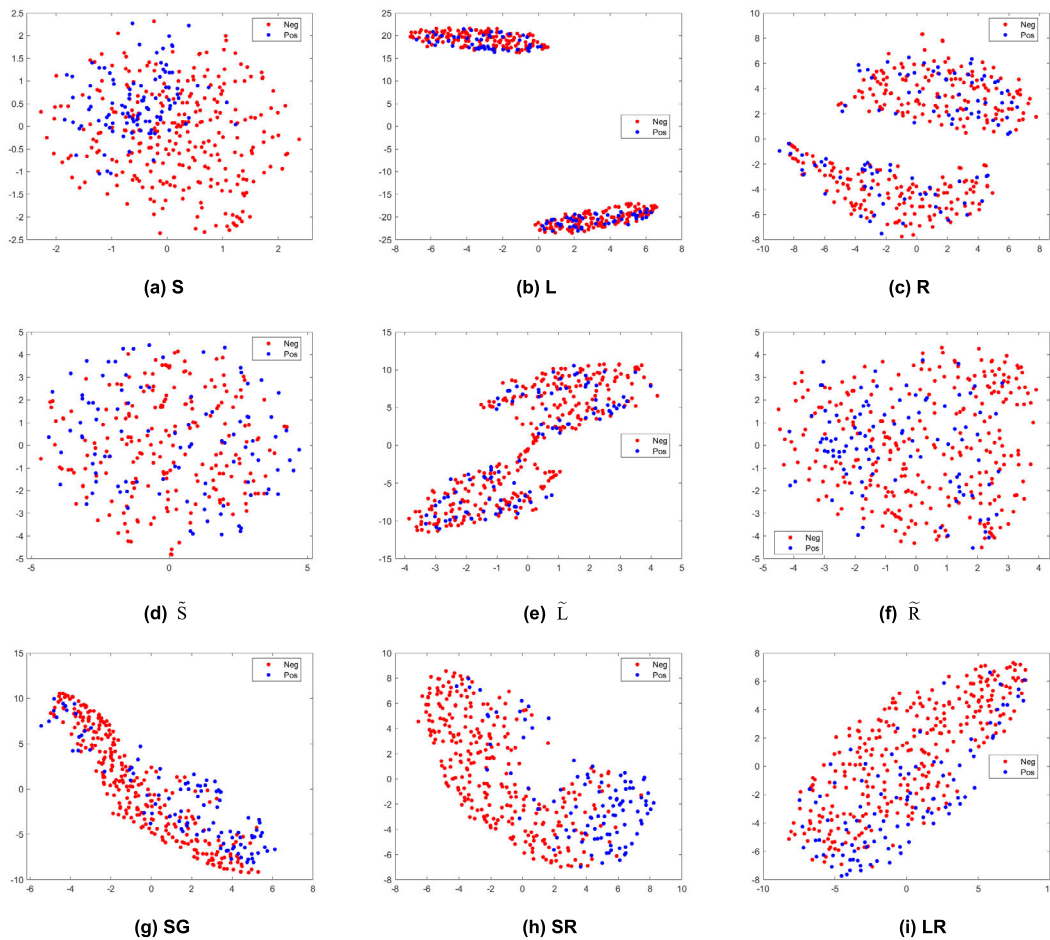


FIGURE 6. t-SNE results in the INbreast dataset.

TABLE 4. t-test results of the RCA model.

Dataset	Acc	AUC
CBIS-DDSM	0.000831	0.001806
INbreast	0.000654	0.000085

improvement. All the results of the t-test are shown in Table 4. The t-test experiment was implemented using the two most important metrics: Acc and AUC. Similar experiments can be implemented using other metrics.

As shown in Table 4, on the two well-known mammographic datasets, the observed performance improvements are essential compared with the initial performance on the original image features because each t-test value in Table 4  $\ll$  0.05. (According to the statistical significance test, if the t-test value is  $< 0.5$ , the hypothesis introduced above should be accepted.) The phenomenon is more evident on the INbreast dataset. This means that the RCA model is effective for breast cancer recognition. Hence, the proposed fine-grained feature selection idea makes sense. The novel image features generated by the RCA model have more discriminant ability.

In summary, based on Figures 3–6, Table 4, and scientific reasoning, the effectiveness, efficiency, and robustness of the RCA model has been comprehensively demonstrated from diverse perspectives.

#### 4) EVALUATIONS OF OTHER IMPORTANT METRICS

In addition to the above-mentioned Acc and AUC metrics, sensitivity (Sen) and specificity (Spe) are two other important metrics that are usually used to evaluate a CAD system. On the one hand, higher sensitivity means that a lower FNR or fewer missed diagnoses can be obtained. Thus, real patients can be treated in a timely manner. On the other hand, higher specificity means that a lower FPR or higher probability of definite diagnosis can be obtained. Hence, these two metrics evaluate the real practicality of a breast cancer recognition model. In most cases, the former metric (Sen) is more important than the latter one (Spe). These two metrics are evaluated in Figures 7(a) and (b). Several state-of-the-art baselines were chosen to provide the corresponding values. Moreover, to measure the overall recognition performance, the Kiviat-diagram-based evaluation metric of Equation (22) [44], which considers the four most important metrics (Acc, AUC, Spe, and Sen), was used in Figures 7(c) and (d).

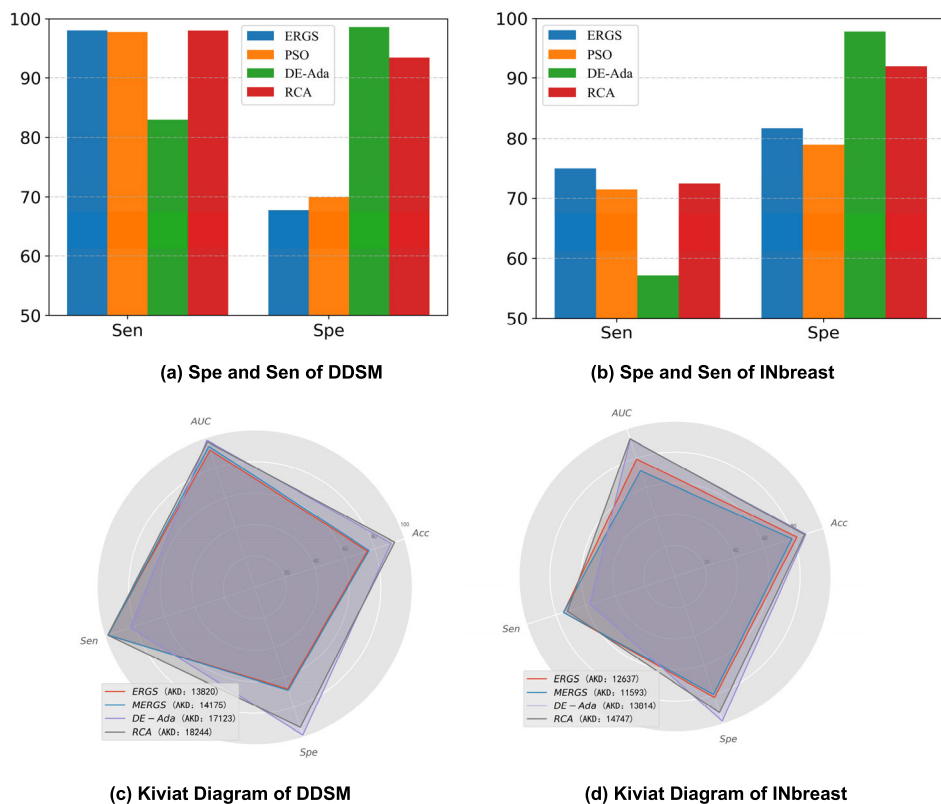


FIGURE 7. Sensitivity, specificity, and AKD values of each model.

As shown in Figures 7(a) and (b), the RCA model can achieve competitive sensitivity compared with state-of-the-art baselines, including the ERGS, PSO, and DE-Ada\* models, indicating that a lower FNR or fewer missed diagnoses can be obtained with the proposed model. More importantly, real patients can be traced in a timely manner. Meanwhile, the specificity of the RCA model is also satisfactory. Thus, a lower FPR can be obtained in the real diagnosis process. In most cases, the former metric (FNR) is more significant than the latter one (FPR). For example, although the DE-Ada\* model has better specificity, the RCA model obtains better sensitivity on both datasets. Hence, this important finding proves the practicality of the RCA model. Most importantly, the RCA model obtains the best overall recognition performance if all the four metrics in a Kiviati diagram are considered. As shown in Figures 7(c) and (d), the RCA model obtains the largest AKD value. In particular, it outperforms the state-of-the-art DE-Ada\* model with a large performance gap of approximately 1121. In general, the RCA model can obtain satisfactory sensitivity and AKD, in addition to the Acc and AUC, demonstrating its higher practicality.

D. ABLATION ANALYSIS

Ablation analysis is necessary to evaluate the real contribution of each feature selection layer of the RCA model. Hence, several model variations were used, and the

corresponding experimental results are exhibited in Table 5. As previously mentioned, the RCA model consists of three feature selection layers: the first-layer feature refinement, the second-layer cross-modal correlation mining, and the last-layer decision-tree-guided adaptive feature selection. Therefore, the results from four variations are used to evaluate the importance of each feature selection layer. First, the RCA model was obtained by only removing the MvERGS algorithm from the RCA model (R means that the feature refinement layer was removed, and E and A have similar meanings). For example, a performance degradation of Acc (77.75% – 93.30% = –15.55%) can be observed in the fourth column. Second, the RCA model was obtained by only removing the cross-modal correlation mining module from the RCA model. For example, a performance degradation of Acc (81.34% – 93.30% = –11.96%) can be observed in the fifth column. Third, the RCA model was obtained by only removing the adaptive feature selection layer from the RCA model. For example, a performance degradation of Acc (91.39% – 93.30% = –1.91%) can be observed in the sixth column. Moreover, the RCA model was obtained by removing all the layers from the proposed RCA model. Thus, the RCA model reduces into a traditional classification model that only uses the best original single feature (S and D) and the GBDT classifier to complete the recognition. The corresponding ablation analysis was completed on the two most important metrics (Acc and AUC). Finally, the average

**TABLE 5. Ablation analysis results (unit: %). (Note: A red strikethrough indicates that the feature selection layer has been removed.)**

(Note: A red strikethrough indicates that the feature selection layer has been removed.)

Dataset	Metric	RCA	<del>RCA</del>	<del>RCA</del>	<del>RCA</del>	<del>RCA</del>	Mean
CBIS-DDSM	Acc	93.30	-15.55	-11.96	-1.91	-12.68	-10.53
	AUC	97.22	-8.08	-0.10	-0.07	-2.22	-2.62
INbreast	Acc	87.07	-8.62	-1.72	0.00	-8.62	-4.74
	AUC	93.00	-15.75	-9.84	0.00	-20.44	-11.51

value of each row, which can be used to evaluate the overall effect of the corresponding ablation analysis, was calculated, and the contribution of each component was determined.

In Table 5, the effect of removing the corresponding feature selection layer is apparent, demonstrating that the ablation analysis is useful. On the one hand, removing the MvERGS algorithm leads to a large performance degradation on both metrics. On the other hand, removing the last-layer feature selection leads to a small performance degradation. This trend is more significant on the Acc metric in the CBIS-DDSM dataset, whereas it is more significant on the AUC metric in the INbreast dataset. Compared with the last layer, the MvERGS algorithm mainly refines the original feature space rather than compressing the number of features, which can maximally retain the key discriminant information. Clearly, the RCA model obtains a larger Acc improvement on a relatively coarse-grained mammographic dataset (CBIS-DDSM). However, the RCA model obtains a larger AUC improvement on a relatively fine-grained mammographic dataset (INbreast). This is an interesting conclusion. Because of the cross-modal correlation mining, the new features generated by the second layer have similar discriminant abilities. Removal of the second layer also leads to a large performance degradation. However, the last-layer feature selection contributes little to the RCA model because it focuses on compressing the number of features. Thus, the importance of each feature selection layer of the RCA model in a descending order is  $R$  (MvERGS)  $>$   $C$  (DCA)  $>$   $A$  (adaptive). Moreover, the order explains the name of the new model: the first layer is the most important of the three, whereas the last layer is the least important.

Hence, the MvERGS algorithm is a valuable by-product of the RCA model. It can be used in some downstream research fields that need elaborate or fine-grained feature selection. Meanwhile, cross-modal correlation mining is another valuable method for boosting the final recognition performance. More-discriminant but low-dimensional features are needed to fit the sample size well. Most importantly, removing all the layers results in a relatively large performance degradation. For example, in the CBIS-DDSM dataset, it was found that approximately 12.68% of the accuracy degradation occurred when the RCA model reduced to the traditional classification model. This trend is more evident in the AUC metric. All these results indicate that the real practicality of the traditional classification model degrades dramatically.

In summary, these results help to understand better the real contribution of each feature selection layer. It is planned to modify the corresponding layer purposely in future work and to improve the recognition performance further.

### E. ONLINE BREAST CANCER RECOGNITION SYSTEM

To verify further the practicality of the RCA model in a real application, an end-to-end online breast cancer recognition system was designed based on the proposed model. The system can provide rapid and accurate online breast cancer recognition for radiologists or pathologists, providing more convenience in clinical diagnoses. This system has been on the authors' web server, and the corresponding internal testing has been completed. In the future, it is planned to transplant the system to the web server of a cooperating hospital in Nanchang and to complete the corresponding  $\beta$  testing within the practical pathological diagnosis. The recognition results of the system are illustrated in Figure 8. Channel 1 is the first testing channel based on the CBIS-DDSM dataset, whereas Channel 2 is the second testing channel based on the INbreast dataset. In Figure 8 (a), radiologists and pathologists should click the button "Submit Testing" to upload their local untested mammograms to the web server and perform online breast cancer recognition. The middle part of Figure 8 (b) exhibits the real-time recognition results. As shown in Figure 8, the second testing channel was chosen to implement real-time breast cancer recognition ((a)), and the current mammogram is predicted as a negative one in approximately 1 ms ((b)).

The proposed online breast cancer recognition system can obtain satisfactory real-time recognition performance. More importantly, it helps to narrow the gap between theoretical research and practical application. Owing to its general framework, the system can be directly absorbed into practical clinical diagnoses. It can assist radiologists and pathologists in formulating their clinical diagnoses and improving real-time efficiency.

### F. PARAMETER TUNING AND REAL-TIME EFFICIENCY

#### 1) PARAMETER TUNING

Parameters  $\alpha$  and  $\theta$  in the MvERGS algorithm must be tuned carefully. The parameter  $\theta$  was used to select the most-correlated feature component. During the experiment, the parameter  $\theta$  was set to the same value in each perspective of the MvERGS algorithm. Because the features were normalized,  $\theta$  was set from 0.1 to 0.9. When  $\theta$  was 0.5, the best Acc and AUC values were obtained. The parameter  $\alpha$  in Equation (9) was used to adjust the corresponding weights of  $K_w$  and  $E_w$ . The parameter  $\alpha$  was set from 0.1 to 0.9 as well.

To determine the value of  $\alpha$  better, the AKD metric was used to evaluate the corresponding performance of each  $\alpha$  value more comprehensively. It was found that, when  $\alpha$  was 0.7, the best AKD value, 27,544, was obtained. Please see Table 6.

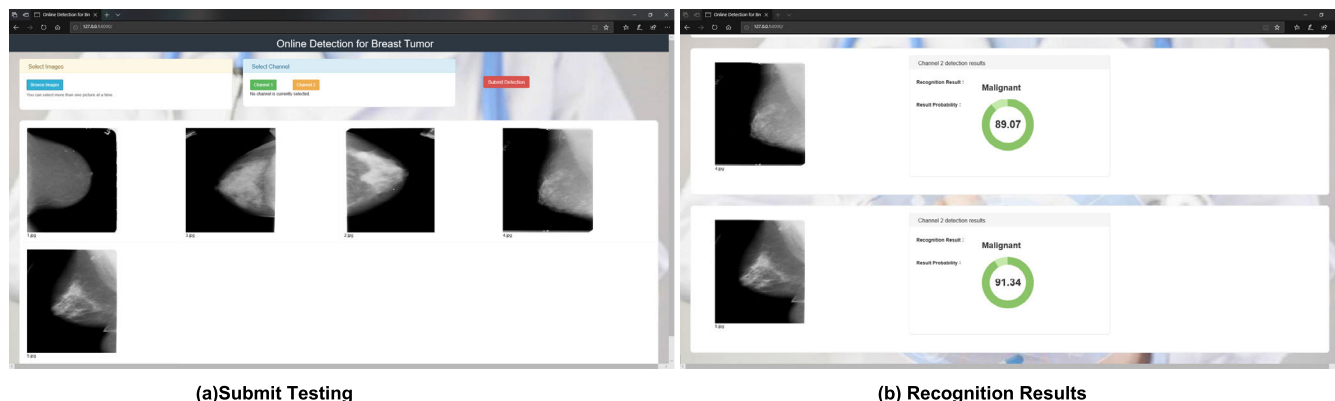


FIGURE 8. Real-time online detection (note: the test was implemented on an internal web server (<http://127.0.0.1:8000>)).

TABLE 6. AKD of each  $\alpha$ .

$\alpha$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
AKD	27370	27355	27356	27521	27390	27375	27544	27387	27404

TABLE 7. Test time of each batch (unit: ms).

batch	1	2	3	4	5	6	7	8	9	10
cost time	0.999	0.999	0.999	0.999	1.000	1.000	0.99	0.997	0.979	0.999

## 2) REAL-TIME EFFICIENCY

To evaluate the efficiency of the RCA model, the test set was divided into 10 batches, each containing 86 images, and the test time of each batch was calculated.

As shown in Table 7, the test time of each batch was approximately 1 ms. Considering the short test time, the proposed model provides timely results.

## V. CONCLUSION AND FUTURE WORK

Early detection of breast cancer is crucial for improving the survival rates of patients. Pathologists and radiologists need a CAD system to assist their clinical diagnoses effectively and efficiently. However, current breast cancer recognition models face the sample scarcity problem. To alleviate this problem, the simple, effective RCA model was proposed from the perspective of fine-grained feature selection. The most-discriminant information can be mined progressively by the RCA model. Extensive experimental results demonstrate the effectiveness, efficiency, and robustness of the RCA model, proving its practicality. More importantly, the AKD metric proves that the RCA model obtains the best overall recognition performance. Other important advantages are that the model requires fewer parameters and can be deployed on a normal workstation. The RCA model, however, is not an end-to-end model.

Hence, in addition to the theoretical research, an end-to-end breast cancer diagnosis system was designed based on the proposed model, and the corresponding  $\alpha$  testing was completed. The online diagnosis system can provide rapid and effective breast cancer recognition, making clinical diagnoses more convenient and narrowing the gap between

theoretical research and practical application. Moreover, as a valuable by-product of the RCA model, the innovative MvERGS algorithm can be used in some downstream research fields that require elaborate or fine-grained feature selection. The last-layer feature selection strategy of the model must be optimized further. Other state-of-the-art boosting models, such as CatBoost [39] and LightGBM [51], can be used to achieve this goal. In planned future research, the RCA model will be used to perform COVID-19 detection [52].

## ACKNOWLEDGMENT

The authors extend their gratitude to Dr. Rebecca Sawyer Lee and The Cancer Imaging Archive Public Access (TCIA) for collecting the CBIS-DDSM dataset and annotating the lesion areas to promote the research of breast mass classification. They also give thanks to the Breast Research Group, Hospital de São João, Breast Centre, Porto and Portugal for providing, sorting and annotating the INbreast dataset and thus promoting the research of breast mass classification. They also give thanks to their other group members: Haowei Shi and Qipeng Xiong. The authors would also like to acknowledge the editor and the reviewers for their helpful suggestions.

## REFERENCES

- [1] A. I. Shahin and S. Almotairi, "An accurate and fast cardio-views classification system based on fused deep features and LSTM," *IEEE Access*, vol. 8, pp. 135184–135194, 2020.
- [2] A. Sedik, A. M. Iliyasu, B. A. El-Rahiem, M. E. A. Samea, A. Abdel-Raheem, M. Hammad, J. Peng, F. E. Abd El-Samie, and A. A. A. El-Latif, "Deploying machine and deep learning models for efficient data-augmented detection of COVID-19 infections," *Viruses*, vol. 12, no. 7, p. 769, 2020.



- [3] X. Zhang, R. Li, H. Dai, Y. Liu, B. Zhou, and Z. Wang, "Localization of myocardial infarction with multi-lead bidirectional gated recurrent unit neural network," *IEEE Access*, vol. 7, pp. 161152–161166, 2019.
- [4] X. Zhang, D. Cheng, P. Jia, Y. Dai, and X. Xu, "An efficient android-based multimodal biometric authentication system with face and voice," *IEEE Access*, vol. 8, pp. 102757–102772, 2020.
- [5] M. Hammad, Y. Liu, and K. Wang, "Multimodal biometric authentication systems using convolution neural network based on different level fusion of ECG and fingerprint," *IEEE Access*, vol. 7, pp. 26527–26542, 2019.
- [6] M. Allam and M. Nandhini, "A study on optimization techniques in feature selection for medical image analysis," *Int. J. Comput. Sci. Eng.*, vol. 9, no. 3, pp. 75–82, 2017.
- [7] R. J. S. Raj, S. J. Shobana, I. V. Pustokhina, D. A. Pustokhin, D. Gupta, and K. Shankar, "Optimal feature selection-based medical image classification using deep learning model in Internet of Medical things," *IEEE Access*, vol. 8, pp. 58006–58017, 2020.
- [8] Z. Wu, W. Lin, and Y. Ji, "An integrated ensemble learning model for imbalanced fault diagnostics and prognostics," *IEEE Access*, vol. 6, pp. 8394–8402, 2018.
- [9] W. Książek, M. Hammad, P. Plawiak, U. R. Acharya, and R. Tadeusiewicz, "Development of novel ensemble model using stacking learning and evolutionary computation techniques for automated hepatocellular carcinoma detection," *Biocybernetics Biomed. Eng.*, vol. 40, no. 4, pp. 1512–1524, 2020.
- [10] E. Wu, K. Wu, D. Cox, and W. Lotter, "Conditional infilling GANs for data augmentation in mammogram classification," in *Image Analysis for Moving Organ, Breast, and Thoracic Images*. Cham, Switzerland: Springer, 2018, pp. 98–106.
- [11] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015, *arXiv:1511.06434*. [Online]. Available: <http://arxiv.org/abs/1511.06434>
- [12] M. Pakravan, M. C. Heuzey, and A. Aji, "A fundamental study of chitosan/PEO electrospinning," *Polymer*, vol. 52, pp. 4813–4824, Sep. 2011.
- [13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2012, pp. 1097–1105.
- [14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: <http://arxiv.org/abs/1409.1556>
- [15] T. A. E. Silva, L. F. Silva, D. C. Muchaluat-Saade, and A. Conci, "A computational method to assist the diagnosis of breast disease using dynamic thermography," *Sensors*, vol. 20, p. 3866, Jan. 2020.
- [16] J. Cong, B. Wei, Y. He, Y. Yin, and Y. Zheng, "A selective ensemble classification method combining mammography images with ultrasound images for breast cancer diagnosis," *Comput. Math. Methods Med.*, vol. 2017, Jan. 2017, Art. no. 4896386.
- [17] B. Q. Huynh, H. Li, and M. L. Giger, "Digital mammographic tumor classification using transfer learning from deep convolutional neural networks," *J. Med. Imag.*, vol. 3, Aug. 2016, Art. no. 034501.
- [18] N. Dhungel, G. Carneiro, and A. P. Bradley, "The automated learning of deep features for breast mass classification from mammograms," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI*. Cham, Switzerland: Springer, 2016, pp. 106–114.
- [19] H. Chougrad, H. Zouaki, and O. Alheyane, "Multi-label transfer learning for the early diagnosis of breast cancer," *Neurocomputing*, vol. 392, pp. 168–180, Jun. 2020.
- [20] S. Mourragui, M. Loog, M. A. Van De Wiel, M. J. T. Reinders, and L. F. A. Wessels, "PRECISE: A domain adaptation approach to transfer predictors of drug response from pre-clinical models to tumors," *Bioinformatics*, vol. 35, pp. i510–i519, Jul. 2019.
- [21] A. Medela, A. Picon, C. L. Saratxaga, O. Belar, V. Cabezón, R. Cicchi, R. Bilbao, and B. Glover, "Few shot learning in histopathological images: Reducing the need of labeled data on biological datasets," in *Proc. IEEE 16th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2019, pp. 1860–1864.
- [22] J. Chen, J. Jiao, S. He, G. Han, and J. Qin, "Few-shot breast cancer metastases classification via unsupervised cell ranking," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, early access, Dec. 16, 2019, doi: [10.1109/TCBB.2019.2960019](https://doi.org/10.1109/TCBB.2019.2960019).
- [23] N. Dhungel, G. Carneiro, and A. P. Bradley, "Fully automated classification of mammograms using deep residual neural networks," in *Proc. ISBI*, Apr. 2017, pp. 310–314.
- [24] B. E. Bejnordi, G. Zuidhof, M. Balkenhol, M. Hermesen, P. Bult, B. van Ginneken, N. Karssemeijer, G. Litjens, and J. van der Laak, "Context-aware stacked convolutional neural networks for classification of breast carcinomas in whole-slide histopathology images," *J. Med. Imag.*, vol. 4, Jan. 2017, Art. no. 044504.
- [25] L. Shen, L. R. Margolies, J. H. Rothstein, E. Fluder, R. McBride, and W. Sieh, "Deep learning to improve breast cancer detection on screening mammography," *Sci. Rep.*, vol. 9, pp. 1–12, Aug. 2019.
- [26] C. L. Chowdhary, M. Mittal, P. A. Pattanaik, and Z. Marszalek, "An efficient segmentation and classification system in medical images using intuitionist possibilistic fuzzy C-mean clustering and fuzzy SVM algorithm," *Sensors*, vol. 20, p. 3903, Jan. 2020.
- [27] H. Ji, K. Guo, L. Yang, W. Jiang, Z. Zhao, X. Zhu, and A. Hou, "Research on feature selection and classification algorithm of medical optical tomography images," in *Proc. ICCV*, Aug. 2019, pp. 561–566.
- [28] A. Veeramuthu, S. Meenakshi, and A. Kameshwaran, "A plug-in feature extraction and feature subset selection algorithm for classification of medicinal brain image data," in *Proc. ICCSP*, Apr. 2014, pp. 1545–1551.
- [29] M. N. Sudha, S. Selvarajan, and M. Suganthi, "Feature selection using improved lion optimisation algorithm for breast cancer classification," *IJBIC*, vol. 14, no. 4, pp. 237–246, 2019.
- [30] S. U. Kumar and H. H. Inbarani, "PSO-based feature selection and neighborhood rough set-based classification for BCI multiclass motor imagery task," *Neural Comput. Appl.*, vol. 28, no. 11, pp. 3239–3258, 2016.
- [31] X. Zhu, H. Suk, S. Lee, and D. Shen, "Subspace regularized sparse multitask learning for multiclass neurodegenerative disease identification," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 3, pp. 607–618, Mar. 2015.
- [32] D. Zhang and D. Shen, "Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease," *NeuroImage*, vol. 59, pp. 895–907, Jan. 2012.
- [33] T. Zhou, M. Liu, K.-H. Thung, and D. Shen, "Latent representation learning for Alzheimer's disease diagnosis with incomplete multi-modality neuroimaging and genetic data," *IEEE Trans. Med. Imag.*, vol. 38, no. 10, pp. 2411–2422, Oct. 2019.
- [34] A. Kumar, M. Fulham, D. Feng, and J. Kim, "Co-learning feature fusion maps from PET-CT images of lung cancer," *IEEE Trans. Med. Imag.*, vol. 39, pp. 204–217, 2019.
- [35] X. Zheng, J. Shi, Y. Li, X. Liu, and Q. Zhang, "Multi-modality stacked deep polynomial network based feature learning for Alzheimer's disease diagnosis," in *Proc. ISBI*, Apr. 2016, pp. 851–854.
- [36] C. Zu, Y. Wang, L. Zhou, L. Wang, and D. Zhang, "Multi-modality feature selection with adaptive similarity learning for classification of Alzheimer's disease," in *Proc. ISBI*, Apr. 2018, pp. 1542–1545.
- [37] F. Zhang, Z. Li, B. Zhang, H. Du, B. Wang, and X. Zhang, "Multi-modal deep learning model for auxiliary diagnosis of Alzheimer's disease," *Neurocomputing*, vol. 361, pp. 185–195, Oct. 2019.
- [38] X. Lin, H. Song, M. Fan, W. Ren, L. Li, and L. Yao, "The feature selection algorithm based on feature overlapping and group overlapping," in *Proc. BIBM*, Dec. 2016, pp. 619–624.
- [39] R. S. Lee, F. Gimenez, A. Hoogi, K. K. Miyake, M. Gorovoy, and D. L. Rubin, "A curated mammography data set for use in computer-aided detection and diagnosis research," *Sci. Data*, vol. 4, Dec. 2017, Art. no. 170177.
- [40] I. C. Moreira, I. Amaral, I. Domingues, A. Cardoso, M. J. Cardoso, and J. S. Cardoso, "Inbreast: Toward a full-field digital mammographic database," *Acad. Radiol.*, vol. 19, no. 2, pp. 236–248, 2012.
- [41] F. Iandola, M. Moskewicz, S. Karayev, R. Girshick, T. Darrell, and K. Keutzer, "DenseNet: Implementing efficient ConvNet descriptor pyramids," 2014, *arXiv:1404.1869*. [Online]. Available: <http://arxiv.org/abs/1404.1869>
- [42] J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, and H. Liu, "Feature selection: A data perspective," *ACM Comput. Surv.*, vol. 50, no. 6, pp. 1–45, Jan. 2018.
- [43] W. Shao, S. Xiang, Z. Zhang, K. Huang, and J. Zhang, "Hyper-graph based sparse canonical correlation analysis for the diagnosis of Alzheimer's disease from multi-dimensional genomic data," *Methods*, Apr. 2020.
- [44] H. Zhang, D. Qiu, R. Wu, Y. Deng, D. Ji, and T. Li, "Novel framework for image attribute annotation with gene selection XGBoost algorithm and relative attribute model," *Appl. Soft Comput.*, vol. 80, pp. 57–79, Jul. 2019.
- [45] H. Zhang, R. Wu, T. Yuan, Z. Jiang, S. Huang, J. Wu, J. Hua, Z. Niu, and D. Ji, "DE-Ada\*: A novel model for breast mass classification using cross-modal pathological semantic mining and organic integration of multi-feature fusions," *Inf. Sci.*, vol. 539, pp. 461–486, May 2020.

- [46] L. Tsochatzidis, L. Costaridou, and I. Pratikakis, "Deep learning for breast cancer diagnosis from mammograms—A comparative study," *J. Imag.*, vol. 5, p. 37, Mar. 2019.
- [47] A. Rampun, B. W. Scotney, P. J. Morrow, and H. Wang, "Breast mass classification in mammograms using ensemble convolutional neural networks," in *Proc. Healthcom*, Sep. 2018, pp. 1–6.
- [48] W. Zhu, Q. Lou, Y. S. Vang, and X. Xie, "Deep multi-instance networks with sparse label assignment for whole mammogram classification," in *Medical Image Computing and Computer Assisted Intervention—MICCAI*, vol. 10435. Cham, Switzerland: Springer, 2017, pp. 603–611.
- [49] G. Carneiro, J. Nascimento, and A. P. Bradley, "Deep learning models for classifying mammogram exams containing unregistered multi-view images and segmentation maps of lesions," in *Deep Learning for Medical Image Analysis*. New York, NY, USA: Academic, Jan. 2017, ch. 14, pp. 321–339, doi: [10.1016/B978-0-12-810408-8.00019-5](https://doi.org/10.1016/B978-0-12-810408-8.00019-5).
- [50] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.
- [51] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 3146–3154.
- [52] W. M. Shaban, A. H. Rabie, A. I. Saleh, and M. A. Abo-Elsoud, "A new COVID-19 Patients Detection Strategy (CPDS) based on hybrid feature selection and enhanced KNN classifier," *Knowl.-Based Syst.*, vol. 205, Oct. 2020, Art. no. 106270.



**JIANWU ZHUO** was born in Guangzhou, China, in 1994. He received the B.S. degree from the School of Intelligent Manufacturing Engineering, Jiangxi University of Applied science, Nanchang, Jiangxi, China, in 2019. He is currently pursuing the M.S. degree in computer application technology with East China Jiaotong University, Nanchang. His research interests include image understanding, natural language processing, and machine learning.



**ZILIAN JIANG** was born in Jiangxi, China, in 1997. He received the B.S. degree in software engineering from East China Jiaotong University, Nanchang, Jiangxi, in 2018, where he is currently pursuing the M.S. degree in software engineering. His research interests include image captioning, material recognition, and machine learning.



**JINPENG WU** was born in Shanxi, China, in 1992. He received the B.S. degree in software engineering from the Taiyuan University of Technology, Taiyuan, Shanxi, in 2016. He is currently pursuing the M.S. degree in software engineering with East China Jiaotong University, Nanchang, Jiangxi, China. His research interests include image understanding, natural language processing, and machine learning.



**DONGHONG JI** received the B.S., M.S., and Ph.D. degrees from the Computer School, Wuhan University, China, in 1988, 1991, and 1995, respectively. He was a Postdoctoral Research Fellow with Tsinghua University, from 1995 to 1998. From 1998 to 2008, he was a Research Scientist with the Institute for Infocomm Research, Singapore. He is currently a Professor and a Ph.D. Supervisor with the School of Cyber Science and Engineering, Wuhan University. His research interests include natural language processing, machine learning, and data mining.



**HONGBIN ZHANG** was born in Jiangsu, China, in 1979. He received the B.S. degree in computer science and technology and the M.S. degree in traffic information engineering and control from East China Jiaotong University, Nanchang, Jiangxi, China, in 2001 and 2004, respectively, and the Ph.D. degree in computer software and theory from Wuhan University, Wuhan, Hubei, China, in 2016. From 2017 to 2018, he was a Visiting Scholar with the School of Computing and



**GUANGLI LI** was born in Guangxi, China, in 1977. She received the B.S. degree in computer science and technology and the M.S. degree in computer application technology from East China Jiaotong University, Nanchang, Jiangxi, China, in 2001 and 2008, respectively. From 2017 to 2018, she was a Visiting Scholar with the School of Computing and Information Sciences, Florida International University, Miami, FL, USA. She is currently an Associate Professor with the School of Information Engineering, East China Jiaotong University. Her co-supervisor is Prof. Tao Li. Her current research interests include cross-modal retrieval, recommendation systems, and machine learning.



**TIAN YUAN** was born in Hubei, China, in 1994. He received the B.S. degree in computer and information science from Hubei Engineering University, Xiaogan, Hubei, in 2018. He is currently pursuing the M.S. degree in computer application technology with East China Jiaotong University, Nanchang, Jiangxi, China. His research interests include image understanding, tumor image recognition, and machine learning.



**CHUANXIU LI** was born in Shandong, China, in 1995. He received the B.S. degree from the Software School, Qingdao University, Qingdao, Shandong, in 2019. He is currently pursuing the M.S. degree in computer application technology with East China Jiaotong University, Nanchang, Jiangxi, China. His research interests include image understanding, tumor image recognition, and machine learning.