Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

A review of deep learning-based multiple-lesion recognition from medical images: classification, detection and segmentation☆,☆☆

Huiyan Jiang [a,b,*], Zhaoshuo Diao [a], Tianyu Shi [a], Yang Zhou [a], Feiyu Wang [c], Wenrui Hu [a], Xiaolin Zhu [c], Shijie Luo [a], Guoyu Tong [a], Yu-Dong Yao [d]

[a] *Software College, Northeastern University, No. 195, Chuangxin Road, Hunnan District, Shenyang, 110169, Liaoning, China*
[b] *Key Laboratory of Intelligent Computing in Medical Image, Ministry of Education, Northeastern University, No. 195, Chuangxin Road, Hunnan District, Shenyang, 110169, Liaoning, China*
[c] *Northeastern University, No. 195, Chuangxin Road, Hunnan District, Shenyang, 110169, Liaoning, China*
[d] *Department of Electrical and Computer Engineering, Stevens Institute of Technology, 1 Castle Point Ter, Hoboken, NJ 07030, NJ, USA*

ARTICLE INFO

Keywords:
Medical image
Deep learning
Classification
Detection
Segmentation

ABSTRACT

Deep learning-based methods have become the dominant methodology in medical image processing with the advancement of deep learning in natural image classification, detection, and segmentation. Deep learning-based approaches have proven to be quite effective in single lesion recognition and segmentation. Multiple-lesion recognition is more difficult than single-lesion recognition due to the little variation between lesions or the too wide range of lesions involved. Several studies have recently explored deep learning-based algorithms to solve the multiple-lesion recognition challenge. This paper includes an in-depth overview and analysis of deep learning-based methods for multiple-lesion recognition developed in recent years, including multiple-lesion recognition in diverse body areas and recognition of whole-body multiple diseases. We discuss the challenges that still persist in the multiple-lesion recognition tasks by critically assessing these efforts. Finally, we outline existing problems and potential future research areas, with the hope that this review will help researchers in developing future approaches that will drive additional advances.

1. Introduction

Computer-aided diagnosis (CAD) is an important component of medical informatization. Classification, identification, and segmentation of lesions based on medical imaging are critical for disease follow-up diagnosis and treatment plan formulation. Deep learning techniques, particularly convolutional neural networks (CNN), are quickly becoming the ideal solution for automated medical lesion recognition, thanks to the outstanding advances of artificial intelligence technologies represented by deep learning in natural picture processing [1–3]. Convolutional neural networks [4–6] have currently achieved excellent results in the classification, detection, and segmentation of single lesions such as liver tumors, kidney tumors, nasopharyngeal lesions, and other single lesions.

In most clinical applications, it is not evident whether a patient has a specific disease until the patient has the relevant radiographic examination, and the previously described single lesion model can only output whether the patient has this disease or not. Patients, on the other hand, will suffer from a variety of diseases, and there are numerous forms of lesions in a single organ. As a result, the model must be able to distinguish several lesions, output the type of lesion the patient may have, and segment the lesion region. The multiple lesions discussed in this paper have two dimensions: the recognition of multiple lesions within a single organ, such as lung nodules, lung cancer, and COVID-19 based on lung CT, and the recognition of multiple lesions with multiple locations, such as lymphoma, which occurs in all lymphatic systems throughout the body. The major challenge of multiple-lesion classification, detection, and segmentation is the increased complexity of the problem with limited data. The number of recognition classes for multiple-lesion recognition within a single organ is increased from two (background + foreground) to N for the model. The model requires more complex classifiers and must learn more disease features and distinguish them within a limited data set. The range of model input

☆ This work was supported by the Natural Science Foundation of Liaoning Province, China (No. 2021-YGJC-07) and National Natural Science Foundation of China (No. 61872075)

☆☆ Huiyan Jiang, Zhaoshuo Diao and Tianyu Shi are co-first authors of this article.

* Corresponding author at: Software College, Northeastern University, No. 195, Chuangxin Road, Hunnan District, Shenyang, 110169, Liaoning, China.
E-mail address: hyjiang@mail.neu.edu.cn (H. Jiang).

https://doi.org/10.1016/j.compbiomed.2023.106726
Received 15 November 2022; Received in revised form 7 February 2023; Accepted 27 February 2023
Available online 1 March 2023

Table 1
The number of papers related to the recognition method based on deep learning of human main organs and tissues in the last two years. Its data was collected using title and abstract keywords from the Scopus website.

Organ/Tissue	Paper Number
Brain	321
Ocular	91
Nasal	4
Oral	32
Breast	227
Cardiac	53
Lung	193
Abdomen (Liver+Kidney+Stomach)	83
Prostate	51
Skin	132
Bone	43

is greatly expanded for the recognition of whole-body multiple-lesion, and the model must detect lesion regions in a wider range.

Researchers have increased their focus on recognizing multiple diseases in recent years. In terms of classification, Tushar et al. [7] classified 6 organs and 15 diseases using the weakly supervised method based on body CT scans and Ouda et al. [8] classified 45 diseases based on fundus images. As for detection, Li et al. [9] detected 12 diseases based on color fundus photography and LaRose et al. [10] detected multiple sclerosis cortical lesion based on ultra-high-field MRI. In terms of segmentation, Liang et al. [11] segmented multiple testicular cell types using interactive learning method and Kamraoui et al. [12] segmented multiple sclerosis lesion using broader deep-learning generalization method. Although some recent review articles on deep learning algorithms for disease recognition have been published [13–16], they are either focused on a single organ or a single disease, and there is a dearth of review on multiple diseases. Consequently, the focus of this review will be on the use of deep learning to recognize multiple diseases in various organs and tissues. We have retrieved the number of papers related to the recognition method based on deep learning of human main organs and tissues in the last two years and the statistical results are shown in Table 1. As the primary organs of interest in our review, we chose the six tissues and organs from Table 1 that have received the most recent attention. They are brain, ocular, breast, skin, lung and abdomen. We adhere to the following guidelines and procedures when reviewing each organ-related works. First, we will go over several common diseases, imaging examination techniques for this organ, and the imaging symptoms of various diseases in line with them. Then, we will discuss the pertinent multi-lesion recognition techniques that have recently been published. Finally, we will compare these techniques and display them as tables.

In this paper, we review the recent years of literature on deep learning-based multiple-lesion recognition. In Section 2, we give the general technical route for classification, detection and segmentation of multiple-lesion. After that, in Section 3, the paper will review the recognition of multiple-lesion in six organ and tissue areas, including brain, eye, skin, breast, lung, and abdomen. In Section 4 the paper reviews whole-body multiple-lesion recognition. Based on this, we summarize and proposes the current problems and future research directions in multiple-lesion recognition in Section 5.

2. Generalized paradigm for multiple-lesion recognition

Multiple-lesion recognition entails classification, detection, and segmentation of lesions [17]. Deep learning-based lesion detection, classification, and segmentation are all made up of deep learning units, albeit the specific implementation varies slightly. A standard deep learning unit is shown in Fig. 1(a). Data pre-processing, network structure, loss function, and data post-processing are the four components of a deep learning unit. Different pre- and post-processing, network structures, and loss functions are used in lesion classification, detection, and segmentation. The three tasks are linked in a descending order. The classification task, for example, is the classification of the entire image, the detection task is the classification of specific regions in the image (generally referred to as ROI areas), and the segmentation task is the classification of all pixel points within the image [18,19]. Although each task has its own network structure and loss function, some works use the network of high-level tasks for low-level tasks, such as segmentation networks for detection.

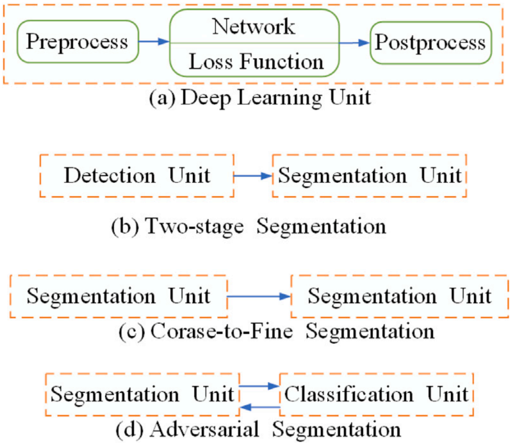


Fig. 1. Generalized paradigm for deep learning based segmentation. (a) deep learning unit. (b) two-stage segmentation (c) coarse-to-fine segmentation (d) adversarial segmentation.

Because of the intricacy of the segmentation problem, it is usual to connect numerous units in series or parallel, as shown in Fig. 1(b)(c)(d) for three segmentation procedures. In the two-stage segmentation technique depicted in Fig. 1(b), a detection unit is first conducted to get the ROI region, followed by another segmentation unit based on the ROI region. The Coarse-to-Fine strategy shown in Fig. 1(c) is also a two-stage model that will perform two segmentation units, the first as a result of coarse segmentation and the second to do the fine segmentation. In adversarial segmentation [20], it can alternatively be thought of as two deep learning units, one for segmentation and the other for classification. In adversarial learning, the segmentation unit corresponds to the generator, and the classifier corresponds to the discriminator [21]. The deep learning models will be concluded in three parts: classification model, detection model, and segmentation model.

2.1. Classification model

Convolutional neural networks, as represented by AlexNet [4], have become the most often used deep learning model in medical image classification due to its outstanding performance in natural image classification. The VGG families [22], represented by VGG-16 or VGG-19, are currently the most often employed network structure in lesion classification. Fig. 2(a) depicts the VGG network, which is made up of several convolutional blocks and a final fully connected layer. Compared to AlexNet, VGG replaces the larger 7 × 7 convolutions in AlexNet with 3 × 3 convolutional combinations. The composition of the convolutional blocks in the VGG network is shown in Fig. 2(b). There are basically N 3 × 3 convolutions, and the convolution blocks are connected by a max pooling operation. The goal of max pooling is to lower the size of the feature map on the one hand while abstracting the image features on the other. The network finally connects a fully connected layer, and then outputs the predicted probabilities for each class. To deal with the situation of gradient vanishment due to the deep network

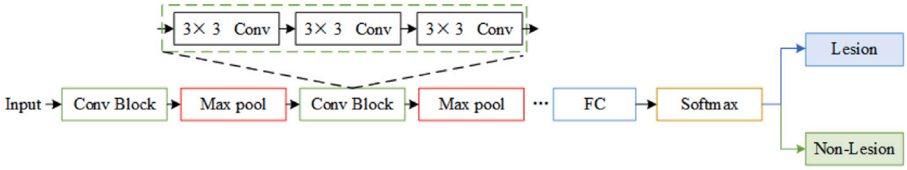


Fig. 2. Diagram of the classification network.

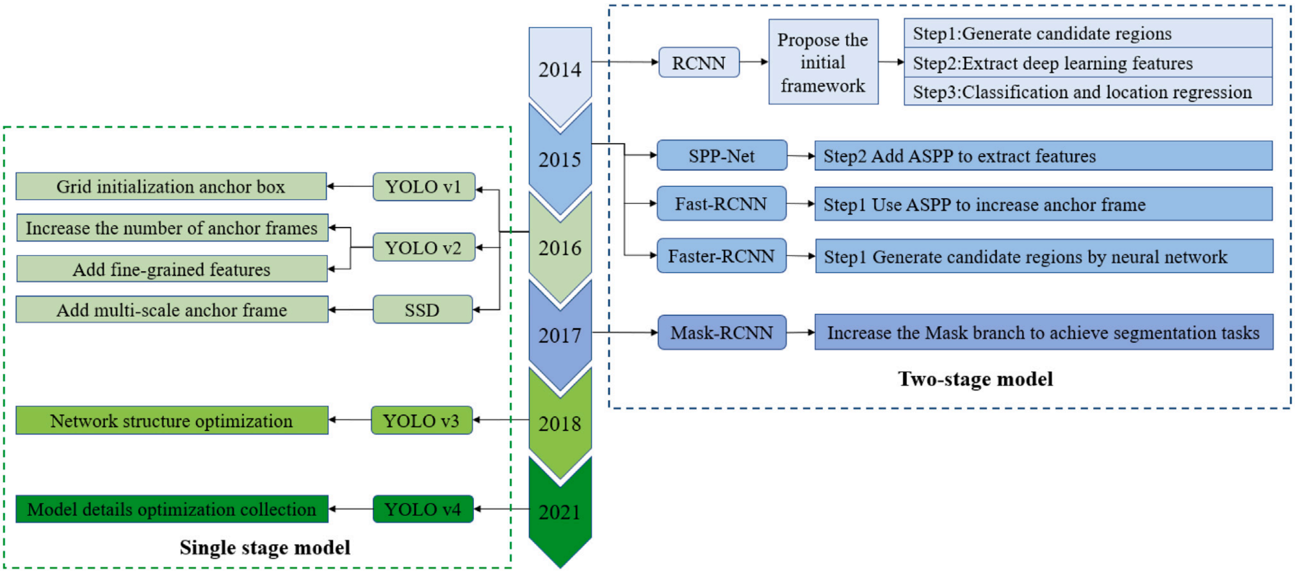

2014
2015
2016
2017
2018
2021
RCNN
Propose the initial framework
Step1:Generate candidate regions
Step2:Extract deep learning features
Step3:Classification and location regression
SPP-Net
Step2 Add ASPP to extract features
Fast-RCNN
Step1 Use ASPP to increase anchor frame
Faster-RCNN
Step1 Generate candidate regions by neural network
Mask-RCNN
Increase the Mask branch to achieve segmentation tasks
Two-stage model
Grid initialization anchor box
YOLO v1
Increase the number of anchor frames
YOLO v2
Add fine-grained features
Add multi-scale anchor frame
SSD
Network structure optimization
YOLO v3
Model details optimization collection
YOLO v4
Single stage model

Fig. 3. Diagram of the evolution of the lesion detection network.

structure in VGG networks, ResNet [23] was proposed. The fundamental convolutional block in ResNet is shown in Fig. 2(c). In comparison to the VGG network's convolutional block, the residual mechanism is introduced to the convolutional block in ResNet, and the convolved feature map is passed to the next convolution operation alongside the initial feature map. DenseNet [24] extends the residual mechanism to all convolution levels based on ResNet, where each feature map not only performs the residual mechanism at the next convolution, but also needs to perform the residual mechanism for all previous feature maps. In addition to residual mechanism, to compensate for small receptive fields of specific convolutional kernel, dilated convolution and pyramidal pooling are proposed. Also, GoogleNet [25] proposed to use several 1 × 1 convolutions to replace other convolutions to reduce the parametric number of the model.

The multiple cross-entropy loss is the most commonly used loss function in classification networks for multi-classification issues. The formula for multiple cross-entropy loss is as follows,

$$L = -\sum_{i=1}^{K} y_i \log p_i \tag{1}$$

where K is the number of classes classified and y_i is the label. $y_i = 1$ when and only when $y = i$, otherwise $y_i = 0$. p_i is the predicted probability value of the model for the current image belonging to class i. In addition to multiple cross-entropy, Barz et al. [26] introduced cosine loss to solve the problem of limited training data.

2.2. Detection model

Object detection models in medical images are often classified into two types: two-stage models and single-stage models. The evolution of common detection models is shown in Fig. 3.

2.2.1. Single-stage detection models

Because the detection model has high requirements for timely performance and the algorithm's running rate differs from other problems, it requires quick performance efficiency, so more study is focused on single-stage object detection. The YOLO v1 [27] model meshy the input images instead of the process of initializing the detection frame, but it cannot cope with the situation where multiple objects fall on a grid at the same time. YOLO v2 [28] added the number of boxes to be selected to the grid, and added fine-grained features through the passthrough layer. SSD [29] added the number of selected frames to the features extracted at multiple scales, and the scale of the selected frames also increased from small to large, thereby significantly improving the detection accuracy of small objects. YOLO v3 [30] added residual structure and feature fusion strategy, using convolution operation instead of pooling operation, which not only improves generalization performance but also reduced parameters. YOLO v4 [31] integrated optimization strategies such as data processing, backbone network, activation function, loss function, etc., to achieve an efficient and powerful detection model, but there is still optimization and improvement in data enhancement and the setting selected boxes.

2.2.2. Two-stage detection models

The RCNN model, first proposed by Girshick et al. [32], is a deep learning method for object detection that consists of three steps: creating candidate regions, extracting deep learning features, classification, and positioning regression. The feature pyramid structure is utilized on this basis to cascade it in front of the completely connected layer or to unify the size of the candidate region. To increase detection accuracy and algorithm speed, He et al. [33] and Girshick et al. [34] introduced SPP-Net and Fast-RCNN, respectively. Girshick et al. [35] modified the algorithm idea based on the original RCNN and Fast-RCNN and introduced Faster-RCNN, employing neural networks to construct the

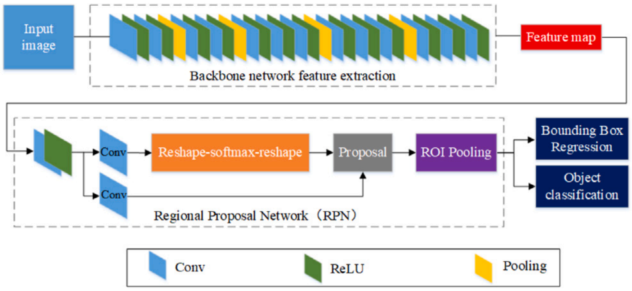


Fig. 4. Diagram of the two-stage detection network (take Fast-RCNN as an example).

selected boxes, to accomplish a two-stage end-to-end technique, considerably boosting speed while maintaining comprehensive performance. Faster-RCNN is also one of the more traditional algorithms in the two-stage technique [36]. On the basis of Faster-RCNN, it is proposed that Mask R-CNN adds the Mask branch composed of ROIAlign and FCN, which realizes simultaneous detection and segmentation tasks while also improving model performance. (See Fig. 4.)

2.2.3. Common loss functions

The loss of the detection model includes classification loss and location loss. The classification loss function is the multivariate cross entropy loss, and the formula for the multivariate cross entropy loss is as follows:

$$L = -\sum_{i=1}^{K} y_i \log p_i \tag{2}$$

MSE loss and IoU loss are commonly used for location loss. The MSE loss intuitively considers the difference in position coordinates, and is defined as follows:

$$L_{MSE} = \frac{1}{n \sum (x_i - y_i)^2} \tag{3}$$

The IoU loss constrains the degree of overlap between the prediction area and the gold standard area, which is defined as follows:

$$L_{IoU} = (X \cap Y)/(X \cup Y) \tag{4}$$

2.3. Segmentation model

Segmentation models, as previously stated, are special classification models that are identical to classification models extended to execute classification tasks for each pixel point in an image. Current deep learning-based segmentation models follow the architecture of encoder–decoder of FCN [5] and SegNet [37]. The structure diagram of the detailed encoder–decoder is shown in Fig. 5. In the encoder stage, features at high level abstraction are obtained by convolution and downsampling. In the decoder stage, the feature map is recovered to the original image size by convolution and upsampling, and the segmentation result is output. U-Net [6] adds a skip-connection mechanism to encoder–decoder, and in the decoder stage, the feature map corresponding to the encoder stage is passed over for concatenation. The U-Net network structure diagram is shown in Fig. 5. U-Net has now become the standard network for lesion segmentation. Based on the improvement of the skip-connection mechanism in U-Net, Zhou and Huang et al. [38,39] proposed U-Net++ and U-Net3+. In addition, for 3D medical image segmentation, Milletari et al. [40] proposed V-Net. Isensee et al. [41] proposed nnU-Net, which has achieved the best results in several medical image segmentation challenges by adding a multi-level supervision mechanism based on U-Net and standardizing both pre-processing and post-processing.

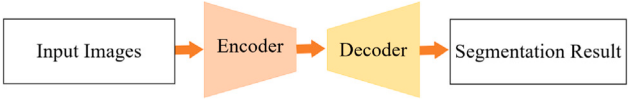


Fig. 5. Diagram of encoder–decoder architecture segmentation network.

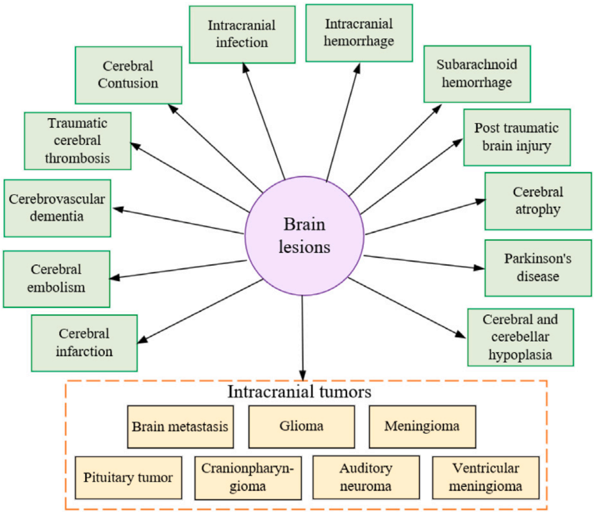


Fig. 6. Major lesions in the brain.

The softmax cross-entropy loss, which is commonly utilized in classification, is still used as the loss function for multiple-lesion segmentation. Furthermore, because it allows for direct optimization of the most significant assessment measure in medical image segmentation, Dice loss is commonly employed. Dice loss is defined as follows,

$$L_{Dice} = 1 - \frac{2|X+Y|}{|X|+|Y|} \tag{5}$$

where X is the predicted result and Y is the ground-truth label. But Dice loss generally faces a binary classification problem, that is, the image is only composed of background and foreground. For multi-class problems, the generalized Dice loss is used. The definitions are as follows,

$$L_{General\_Dice} = 1 - \frac{2\sum_{l=1}^{k}\sum_{n} r_{ln} p_{ln}}{\sum_{l=1}^{k}\sum_{n} r_{ln} + p_{ln}} \tag{6}$$

where k is the number of classes, r_{ln} denotes the true value of class l at the nth pixel point, and takes the value of 1 when the class of the nth pixel point is l, otherwise it is 0. p_{ln} indicates the corresponding predicted value.

3. Multiple-lesion recognition in different body regions

3.1. Brain multiple-lesion recognition

Lesions that originate within the limited area of the skull in the brain can disrupt the function of the relevant brain parts and can potentially extend to other organs. Fig. 6 summarizes the major brain lesions, with brain tumors being one of the most deadly and investigated brain illnesses in clinical medicine.

Gliomas, meningiomas, and pituitary tumors are the three most common types of brain lesions [42]. The World Health Organization (WHO) categorizes gliomas into four risk levels, with grade I being the least dangerous and grade IV being the most dangerous. Meningiomas are benign tumors that form on the membranes of the brain and spinal cord [43]. Pituitary tumors are benign or malignant tumors that develop in the pituitary gland. Patients with this sort of tumor may experience irreversible hormone insufficiency and blindness. Fig. 7 depicts images of three common brain tumors. Because brain tumors vary greatly in shape, size, and aggressiveness, different types of tumors

Table 2
Summary of all publicly available datasets used by the networks on brain multiple-lesion recognition.

Dataset	Data description	Literatures
CE-MRI	Contains three types of tumors (glioma, meningioma and pituitary tumor), and has 3064 axial, coronal and sagittal images of 233 patients [46]	[42,44,45,47–50]
REMBRANDT	MRI multisequence images from 130 patients of various diseases, grades, races and ages [51]	[42]
Harvard Medical School website	Consists of 66 real human brain MRIs with 22 normal and 44 abnormal images (glioblastoma, sarcoma and metastatic bronchopulmonary carcinoma) [52]	[53,54]
BRATS	Normal brain images (340 brain images) and abnormal brain images (260 brain images) were used in the literature.	[55]
Radiopaedia	Contains 121 MR images divided into four different grade [56]	[48]
IXI	Contains nearly 600 MR images of normal subjects (without any lesions) [57]	[49]
cancer imaging archive	MR images of glioma [58]	[49]
REMBRANDT	Includes preoperative MRI multi-sequence image analysis of 130 cases of low-grade or high-grade gliomas [59]	[49]
TCGA-GBM	Magnetic resonance image analysis of brain containing 199 cases of glioblastoma multiforme [60]	[49]
TCGA-LGG	Data collected on 299 patients with low-grade gliomas are included [61]	[49]

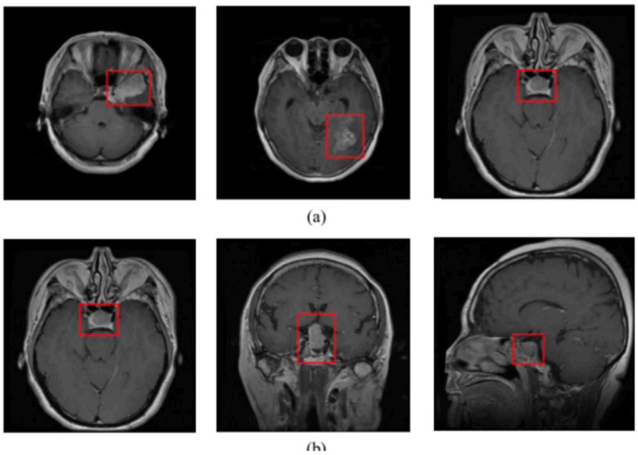


Fig. 7. Three common brain tumors shown in [42], where the tumor regions are marked by red rectangular boxes. (a) From left to right: meningioma, glioma and pituitary tumor, (b) three views of pituitary tumor with different acquisition orientations, from left to right, axial, coronal and sagittal.

may present similarly [44]. As a result, classifying brain tumors remains a difficult task.

Brain magnetic resonance imaging (MRI) is a technique commonly used in clinical practice for the diagnosis of brain tumors since it provides high-resolution images of the brain [45]. Because of the time-consuming and error-prone involvement of physicians in manual tumor margin mapping, computer-aided lesion identification has become a hot topic. There are many MRI public datasets on brain tumor classification, as listed in Table 2. The majority of the datasets cover tumor forms such as glioma, meningioma, and pituitary tumor, but some also include normal human brain MRI images or other diseases such as metastatic bronchopulmonary carcinoma.

With deep learning's success in the image field, this learning method was applied to develop classification models for the aforementioned lesions, and the overall classification results were promising. Similar to traditional methods, deep learning strategies need to be based on large-scale datasets, but it does not require many complex steps, and it enables the processing of automatic network classification while reducing the dependence on expertise [62].

Pre-processing is frequently required before sending brain MRI images into a deep learning network. The majority of deep learning algorithms use a similar data pre-processing approach, which includes downsampling, resampling, image flipping, and normalizing of the original images. It should be noted that during the downsampling process, because the size of the images will be compressed, some of the image feature information may be lost, resulting in a slight decrease in the sensitivity of the network output.

As a common deep learning network, convolutional neural networks can interpret medical images using convolution, pooling, and other operations. In the study of classification and recognition of diseases, both [42,47] used CNN networks for classification operations on MRI images of the brain. Sultan et al. [42] utilized a three-layer convolution, Relu activation, pooling operation and obtained the final classification results through Softmax. It is noteworthy that in their study, the Relu layer in the first layer was followed by a structure called cross-channel normalization layer, by which the relevant activation function can be adjusted to normalize the input layer, and in both the second and third layers, dropout operations are used to reduce the risk of overfitting. In the classification study of tumors by Kaldera et al. [47], only glioma and meningioma were considered, and a strategy for lightweight CNN network using only two convolutional layers and fewer cores was proposed, which is scalable with improved performance. It is also mentioned that the initial model combined with the Faster R-CNN can be used for the meningioma boundary recognition.

Mohsen et al. [53] used the fuzzy C-means clustering approach to segment the image before extracting the image features using the discrete wavelet transform (DWT) method in their investigation of disease process staging detection. For each image only 1024 features were extracted, which is fewer than the number of features obtained using CNN networks. By using principal component analysis (PCA), it was possible to approximate the extracted original features with low-dimensional feature vectors, and a DNN network with seven hidden layers was used for training in performing classification. Experimental results also show that the classification model combining DWT and DNN can produce higher accuracy in the problem of quadruple classification of normal tissue, glioma, sarcoma, and metastatic bronchial lung cancer. Sajjad et al. [48] used CNN networks to classify the grade of tumors, which were classified into four grades, from grade I to grade IV, in ascending order of malignancy. The overall idea was to segment the tumors using a pre-trained InputCascadeCNN network with data augmentation, then extract the deep features and finally classify the tumors. The classification process was performed using a fine-tuned VGG-19 network. Dutta et al. [54] expanded the tumor classification to five categories, namely, glioma, meningioma, metastatic adenocarcinoma, metastatic bronchial carcinoma, and sarcoma, and successfully classified the five tumors using the adaptive neuro-fuzzy inference system (ANFIS), a classifier based on fuzzy rules and fuzzy inference from fuzzy set theory. In the model, if a specific tuple does not match the fuzzy rules defined in the model, the classifier will be biased to produce results outside the solution set, thus, the model can help to identify a new tumor type. Gurunathan et al. [55] proposed two networks for classifying the malignancy degree of meningioma. The first step is to use CNN Deep Net to distinguish between normal images and case images with tumors from the input, and then perform image segmentation and feature extraction in the cases with tumors. The texture features of the segmented tumor regions were computed using a gray-level covariance matrix (GLCM), and the input images of brain tissue were classified for the second time using a GLCM-CNN classifier with a hybrid classification method to distinguish between benign and malignant categories.

In terms of model optimization and evolution, the model will reduce the robustness of the classification due to the absence of the global spatial relationship of images during the classification process, so improvements can be made in the CNN network part to improve the overall classification accuracy of the model. In [45], Capsule Networks (CapsNets) were used to replace the pooling layer in the CNN with a “protocol routing” approach that uses the degree of prediction of the capsule on subsequent capsules as its own contribution. In practice, the network was fed not only images of the brain but also pathological boundaries of the tumor, to make the network focus more on the global tumor target rather than on other minutiae locations. Genetic algorithms select optimal solution according to the survival of the fittest in nature, and Anaraki et al. [49] combined CNN with genetic algorithms to evolve classification models, which can select the best CNN structure according to natural selection mechanisms, and use bagging model averaging methods on the optimal model to reduce the final diagnostic variance. The final model has experimented for both tumor grade and tumor class classification problems, and the results were improved accordingly.

Because medical images have small sample sizes and CNN networks operate best with massive amounts of data, a mix of CNNs and other learning algorithms can be used. In [44,50], both used the strategy of combining CNN networks with transfer learning. Deepak et al. [44] first applied the concept of transfer learning to a specific brain tumor triple classification problem with a convolutional neural network using a modified GoogLeNet, which achieves the highest overall segmentation accuracy when the amount of data is small and can effectively alleviate the overfitting problem of the network compared to a CNN-only model. Swati et al. [50] used a fine-tuning operation of the VGG-19 network pre-trained on the natural image dataset ImageNet to achieve the classification of medical images, considering that for a large number of layers in the VGG-19 network. It would consume a lot of time if each layer is fine-tuned individually, so a chunked fine-tuning strategy is used in this model that reduces the complexity of training and also the training time.

3.2. Ocular multiple-lesion recognition

The detection and classification of ocular disorders is a topic of intense study. The study of ocular multiple lesions focuses mostly on retinal maculopathy. The human retina receives focused light from the lens and turns it into nerve messages. The main sensory area employed for this purpose is the macula, which is located in the center of the retina. A unique layer of photoreceptor nerve cells in the macula processes light. Light intensity, color, and minute visual characteristics are detected by these cells. The retina processes the macula’s information and delivers it to the brain via the optic nerve for visual recognition. In reality, feature perception in vision may be traced back to the retina’s encoding of light into neural impulses in the macula. Pathology such as age-related macular degeneration (AMD) and diabetic macular edema can have an impact on ocular health (DME). In 2013, AMD was the fourth most prevalent cause of blindness, causing blurring of the center of vision, blind spots, and even visual loss [63]. AMD affects about 0.4 percent of adults aged 50 to 60 and 15% of people beyond 60 [63]. Furthermore, DME is the most prevalent diabetes condition that results in vision loss [64]. DME may impair central vision in the early stages of retinopathy. DME is the most common condition that compromises eyesight in diabetic individuals, particularly those with type 2 diabetes [65]. As a result, the diagnosis of AMD and DME is a process that requires attention in computer-aided diagnosis.

Aside from pathological classification and detection, detecting different phases of AMD lesions is a significant task. Early AMD is distinguished by the accumulation of extracellular fluid beneath the retinal pigment epithelium cells. Choroidal warts are deposits that build over time, causing injury to the retinal pigment epithelium and consequent loss of photoreceptor cells [66]. The late stage of AMD is characterized by abnormal and leaking vascular growth (choroidal neovascularization (CNV)) and atrophy of retinal pigment epithelium and photoreceptor cells (topographic atrophy) [67]. These diseases are progressive diseases that lead to visual impairment and blurred vision. Therefore, the purpose of the treatment of these diseases is to retain the residual vision. The effect of early treatment is the best. Furthermore, the treatment methods of advanced AMD and DME are complex and expensive. Early diagnosis can not only protect more vision but also reduce the psychological and economic stress of patients.

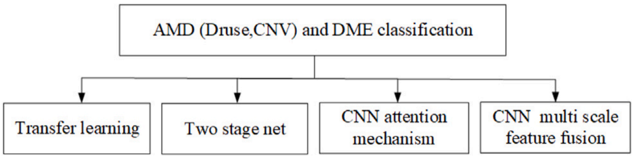

AMD (Druse,CNV) and DME classification
Transfer learning
Two stage net
CNN attention mechanism
CNN multi scale feature fusion

Fig. 8. The figure is a summary of the classification methods of multiple ocular diseases. The classification methods of age-related macular degeneration (AMD) in early, late and diabetic macular edema (DME) in eye diseases are roughly divided into two-step network through migration learning. The combination of CNN and attention mechanism, CNN and multi-scale feature fusion four types of methods.

Rasti et al. [68] advocated using various CNN training methodologies and multi-scale branch CNN to jointly model features for classification. Karri et al. [69] used transfer learning to fine-tune the present network so that it could classify multiple lesions. Li et al. [70] also used transfer learning and developed a VGG-16-based fine-tuning procedure. Lu et al. [71] proposed a residual network with a depth of 100 layers to realize classification. Rong et al. [72] proposed to take image denoising and threshold processing to get denoised images and masks, and then use these data to improve the performance. Mishra et al. [73] proposed a CNN framework with dual attention mechanism, which can automatically pay attention to the relevant regions of the input image on different levels of feature subspaces and enhance local feature perception and classification. Das et al. [74] proposed a CNN network with multi-scale feature fusion, which can capture multi-scale spatial information. Fang et al. [75] proposed an iterative fusion strategy, which can iteratively combine the features of the current layer with those of the previous convolution layer. Fang et al. [76] proposed a two-step classification network. The first step is to establish a detection network to get the attention map. The second step is to combine with the information of the attention map to accelerate the training of the classification network. Huang et al. [77] also proposed a two-step network. The first step is to segment different retina layers, and the second step is to use the retinal layer information corresponding to different lesions to pay attention to the local lesion information. Das et al. [78] proposed to use CNN to extract the features of B-scanned OCT images and input these features into the self-attention module to aggregate features. He et al. [75] proposed two multi-scale feature attention modules based on multi-mode network to obtain features from fundus and OCT images. Sunija et al. [79] proposed to add gradient-based classification activation maps to fully connected layers to analyze classification. Unlike the above classification networks, Frang et al. [80] proposed a lesion area detection network, which uses the ROI area to cut different lesions, establishes the features of different lesion areas, and classifies them through SVM. We summarize the above methods as shown in Fig. 8. The actual classification of the network structure can be seen in Table 3, and the content of the evaluation indicators can be seen in Table 4.

3.3. Breast multiple-lesion recognition

Breast cancer is one of the leading causes of cancer mortality in women worldwide, accounting for one-third of all cancer deaths. Early

Table 3
Comparison of multiple ocular multifocal diagnosis models in the papers..

Model class	DL model	Net	Disease	Refer
Modify CNN	CNN	Multi path CNN	dry age-related macular degeneration, and diabetic macular edema	[68]
	ResNet	100-layer ResNet	macular hole, cystoid macular edema, epiretinal membrane, and serous macular detachment	[71]
	CNN	CNN+Preprocess	dry age-related macular degeneration, and diabetic macular edema	[72]
	CNN	A Lightweight CNN	drusen, CNV, DME	[75]
Two stage segnet	CNN	lesion-aware convolutional neural network (LACNN)	drusen, CNV, DME	[76]
	ReLayNet	layer guided convolutional neural network (LGCNN)	diabetic macular edema,(DME) drusen, and choroidal neo vascularization(CNV)	[77]
Attention + CNN	CNN	Attention-CNN	two common macular diseases, age-related macular degeneration (AMD) and diabetic macular edema (DME)	[73]
	U-Net	B-scan attentive convolutional neural network (BACNN)	AMD, DME	[78]
CNN + Multi scale feature fusion	CNN	Multi scale feature fusion+CNN	drusen, CNV	[74]
	CNN	iterative fusion convolutional neural network (IFCNN)	CNV, DME, Drusen	[79]
	U-Net	MSAN	diabetes, hypertension, nephritis, stroke, AMD, pathologic myopia, retinitis pigmentosa, optic atrophy, and other diseases/abnormalities.	[75]
Transfer Learning	CNN	GoogLeNet	dry age-related macular degeneration, and diabetic macular edema	[69]
	CNN	VGG16	DME and AMD	[70]

Table 4
Detection performance for multiple ocular lesions.

Paper	AUC	ACC	SEN	SPE
Rasti et al. [68]	0.9985	–	–	–
Karri et al. [69]	–	0.99	–	–
Lu et al. [71]	0.984	0.959	0.996	0.9987
Rong et al. [72]	0.9783	–	–	–
Mishra et al. [73]	0.9973	–	–	–
Das et al. [74]	0.996	–	0.996	0.9987
Li et al. [70]	–	0.986	0.978	0.994
Fang et al. [75]	–	0.873	–	–
Fang et al. [76]	–	0.901	–	–
Huang et al. [77]	–	0.899	–	–
Das et al. [78]	0.95	–	–	–
He et al. [79]	0.8552	–	–	–
Sunija et al. [80]	0.9969	–	–	–
Fang et al. [81]	0.8872	–	–	–

detection and treatment can greatly reduce the spread of this dangerous disease. Early detection of breast cancer is usually accomplished using ultrasound imaging or mammography, followed by a breast tissue biopsy [82], which classifies the tissue growth as malignant, benign (non-malignant), or normal. Pathologists will typically study biopsy tissue samples by staining them with hemerocallis and eosin to heighten the contrast between the nucleus and the cytoplasm [83], combining the variability, density, and organization of the nucleus to analyze the overall tissue structure. Invasive cancer tissue, for example, has considerable nuclear diversity and density, as well as structural deformation. Manual analysis led by pathologists, on the other hand, requires the observer to have sufficient professional knowledge, is time-consuming and error-prone, and the morphological changes of the tissue structure, as well as the distinction between benign and malignant, will lead to differences within the observer [84]. The computer-aided diagnosis system can improve diagnostic efficiency by increasing consistency among observers. However, because the cells vary in shape and color, the intra-class images contain a high number of morphological and texture variations, making the classification of breast cancer multiple lesions difficult. Convolutional neural networks are a type of feedforward neural network with a deep structure that involves convolution calculations. It is frequently used for breast lesion recognition and detection. CNN's superior performance enhances the impact of breast cancer tissue image analysis. Table 5 highlights the most recent literature on the recognition of multiple breast lesions.

The investigation of the multi-classification challenge of breast histopathology images focuses primarily on two publicly available data sets: Table 6 provides the detailed information of the two public data sets of breast pathology images: (a) the International Conference on Image Analysis and Recognition (ICIAR 2018) Grand Challenge dataset [94] and (b) the BreakHis dataset [95]. In the ICIAR 2018 BACH challenge data set, the classification of breast cancer lesion types is divided into 2 categories: cancer, non-cancer, and 4 categories: normal, benign, carcinoma in situ, and invasive carcinoma. Based on this data set, Kausar et al. [86] designed a cancer tissue automatic detection technology called HWDCNN, which uses wavelet transform and deep neural network models to obtain more accurate results with less calculation time. Since the histological images of the breast come from different laboratories, the appearance changes during the staining process usually reduce the generalization ability of the deep CNN network. Kausar et al. standardized the coloring of the image, and used the color deconvolution method to transform the input RGB image into the H&E color space. The color was enhanced through rotation, scaling and elastic deformation (brightness, contrast, and tone disturbance [96]) method, to reduce data generalization errors [97]. Then Haar wavelet is applied to transform to decompose the image into a set of sub-frames to reduce the convolution time without any performance degradation, which helps to analyze the local recognition features centered on the cell nucleus in the image, as shown in Fig. 9. The HWDCNN model adopts a feature splicing strategy and combines multi-scale convolution features from different layers to achieve 0.25 seconds in the feature extraction stage and 0.045 seconds in the classification stage, which reduces the time cost. On the ICIAR 2018 verification data, the recognition of Type 4 and Type 2 has achieved an accuracy of 98.2%. Similarly, to overcome the large size of most breast histological images. Alzubaidi

Table 5
Literature summary on the recognition of multiple breast lesions.

Modality	DL mode	Dataset	Preprocess	Categories	Performance metrics	Paper
X-ray, OCT, US, CT	ResNet, auto-sklearn, AutoKeras, Google AutoML Vision	MedMNIST	downsizing operation	Chest, Derma, OCT, Pneumonia, Retina, Breast, Organ	AUC Accuracy	[85]
X-ray	CNN	National Institutes of Health 100,000 chest X-ray	data augmentation	Pneumothorax, Pneumonia, Effusion, Atelectasis, Nodule, Mass, Cardiomegaly, Edema, Lung Consolidation, Pleural Thickening, Infiltration, Fibrosis, Emphysema	ROC curve accuracy = 96%	[86]
microscope images	Deep CNN	ICIAR 2018 BACH grand challenge	stain normalization, elastic deformation,H&E space transformation, wavelet transform	normal, benign, in-situ carcinoma, invasive carcinoma(4-class) non-carcinoma and carcinoma(2-class)	accuracy = 98.2% (4-classes and 2-class)	[87]
	2 branch DCNN	ICIAR 2018 BACH grand challenge	image partition image transformations	normal, benign, in-situ carcinoma, invasive carcinoma(4-class) non-carcinoma and carcinoma(2-class)	accuracy = 89.4%	[88]
	ResNet50	BreakHis	data augmentation	multi-class: Adenosis (A), Ductal carcinoma (DC), Fibroadenoma (F), Lobular carcinoma (LC), Mucinous Carcinoma MC), Papillary Carcinoma (PC), Phyllodes Tumor (PT), and Tubular Adenoma (TA) 2-class: Malignant, Benign	accuracy = 99%(2-class) accuracy = 94-97%(multi-class)	[89]
	DCNN	BreakHis	data augmentation	multi-class: Adenosis (A), Ductal carcinoma (DC), Fibroadenoma (F), Lobular carcinoma (LC), Mucinous Carcinoma MC), Papillary Carcinoma (PC), Phyllodes Tumor (PT), and Tubular Adenoma (TA) 2-class: Malignant, Benign	accuracy = 100% (2-class) accuracy = 97% (multi-class)	[90]
	CSML, Hashing DNN	BreakHis	stain normalization RGB to grey generating overlapping & optimal patch segments	multi-class: Adenosis (A), Ductal carcinoma (DC), Fibroadenoma (F), Lobular carcinoma (LC), Mucinous Carcinoma MC), Papillary Carcinoma (PC), Phyllodes Tumor (PT), and Tubular Adenoma (TA) 2-class: Malignant, Benign	accuracy = 95.8% (multi-class) sensitivity = 100%, recognition rate = 99.3%(2-class)	[91]
–	Feature Selection, SVM, SMO, Random Forest	Breast Cancer genes	–	Luminal A (LumA), Luminal B (LumB), Triple negative/basal¬like (Basal), HER2	accuracy = 100%	[92]
	Apriori-like Feature Selection (AFS)	Breast Cancer genes	–	Luminal A (LumA), Luminal B (LumB), Triple negative/basal¬like (Basal), HER2	accuracy = 99%	[93]

et al. [90] divided the image into blocks of 512 × 512 pixels to prevent cells from being truncated in two blocks. The smaller block cannot provide identification information for the correct category of the image. Since increasing the network depth will lead to the disappearance of the image gradient, this paper proposes a two-branch deep convolutional neural network. Based on the design of increasing the width and depth, different levels of functions are integrated at each step of the network to extract features such as edges, colors, and shapes. And achieved an 89.4% image classification accuracy rate on the test images of ICIAR-2018.

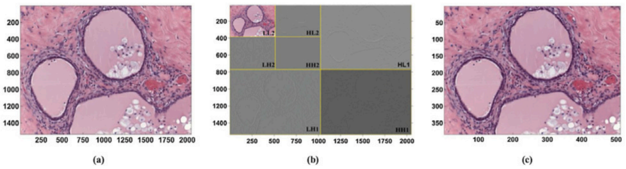


Fig. 9. Haar wavelet demo. (a) The original 2048 × 1536 pixel image (benign); (b) The second-level Haar wavelet decomposition; (c) The decomposed 512 × 384 pixel image in [86].

BreakHis comprises 7909 histological images of breast cancer from 82 patients. Each image is classified as either benign or malignant. Adenopathy (A), fibroadenoma (F), phyllodes tumor (PT), and tubular adenoma are subgroups of the benign category (TA). Malignant tumors are classified into four subtypes: ductal carcinoma (DC), lobular

Table 6
Detailed information of breast pathology image public dataset.

Dataset	Scale	Num	Categories
The International Conference on Image Analysis and Recognition (ICIAR 2018) Grand Challenge dataset	2,048 × 1,536 × 3	400 pieces of training data, 100 pieces of test data	2 classification: cancer, non-cancer; 4 categories: normal, benign, carcinoma in situ, invasive carcinoma
BreakHis dataset	700 × 460 × 3	7909 images of 82 patients: 2480 benign, 5429 malignant Magnification: 40 times, 100 times, 200 times, 400 times	2 classification: benign, malignant; 4 subcategories: benign: adenopathy (A), fibroadenoma (F), phyllodes tumor (PT), tubular adenoma (TA). malignant: ductal carcinoma (DC), lobular carcinoma (LC), mucinous carcinoma (MC), papillary carcinoma (PC)

carcinoma (LC), mucinous carcinoma (MC), and papillary carcinoma (PC). The data set contains images magnified 40 times, 100 times, 200 times, and 400 times. However, the volume of the BreakHis data set is not large enough, and the images are not evenly distributed among different classes.

In this regard, Yari et al. [88,89] developed transfer learning models such as ResNet50 and DCNN, as well as pre-trained networks, to lessen the impact of a little amount of data on accuracy. Data augmentation is accomplished through random horizontal flipping, color dithering, and random rotation, and training encompasses all image formats and magnifications. As a result of the complicated cell structure and the small number of images in the data set, the medium and lobular carcinoma (malignant) categories resulted in some recognition mistakes, which were insufficient to develop a robust enough classifier. Pratiher et al. [91] proposed a fully automatic, near real-time, and low-computing robust multi-classification depth framework, which realized the deep context classification of hybrid overall level CSML representation and local hash signature, thereby effectively distinguishing benign and Vicious subcategories. The model pipeline is composed of three main stages: the preprocessing stage of staining standardization, and the generation of overlapping and optimal block segments of a specific size for subsequent tissue index contour extraction. This method shows a high degree of specificity for malignant subtypes and is applicable for clinical deployment in handheld smart devices.

Some studies, in addition to assessing pathological images of breast cancer, employ genes to predict breast cancer subtypes. Breast cancer is classified into four subtypes: luminal A (LumA), luminal B (lum), triple-negative/baselial, and HER2. Etemad et al. [92] suggested a hierarchical model for predicting breast cancer subtypes based on a breast cancer data set of 13582 genes, employing feature selection approaches to accomplish reliable classification, using SVM, SVM polynomial, and SVM radial basis. Classifiers such as function (SVM-RBF), sequential minimum optimization (SMO), and random forest. The easier to classify subtypes are located at the top of the decision tree, and the interference of the easy to classify subtypes is effectively eliminated in the initial step, so that the most difficult subtypes, LumA and lumB, are classified at the end.

Pham et al. [93] presented the Apriori-like feature selection (AFS) approach for finding the best classification model by exploring both the feature space and the parameter space at the same time. To avoid being excessively exhaustive or greedy, its search behavior is tuned toward the best results. They use a tree-based architecture to cope with multi-class problems, with each node representing a binary classification model learned by AFS. When applied to five subtypes of breast cancer, SVM or K-NN is used as a classifier. In addition, by replacing the classifier with a regression method, AFS can also be used for regression problems. In addition, some breast disease researches are based on whole-body data sets or chest multi-disease prediction research. For example, Yang et al. [85] proposed a set of 10 preprocessed medical open data sets. Among them, BreastMNIST is based on A data set of 780 breast ultrasound images [98], divided into 3 categories: normal, benign, and malignant. When low-resolution images are used, normal and benign are combined as positive, and malignant as negative, thereby simplifying to a binary classification task. This paper selects several automatic classification algorithms for benchmark testing: resnet as the baseline method, auto-sklearn [99], AutoKeras [100], Google AutoML Vision, to promote the automated research of medical image analysis. Ramadan et al. [87] obtained chest X-ray images from the National Institutes of Health data set to predict pneumothorax, pneumonia, effusion, atelectasis, nodules, masses, cardiac hypertrophy, edema, lung consolidation, pleura Chest diseases such as thickening, infiltration, fibrosis, and emphysema. After the data was enhanced, the pre-training model Inception-v3 in migration learning was used to extract features from the data set, achieving a prediction accuracy of 96%.

3.4. Skin multiple-lesion recognition

Skin cancer is one of the most common and dangerous malignancies in the world [101]. Malignant melanoma is one of the most dangerous forms of skin cancer, having a high fatality rate. Melanoma is a type of skin cancer that cannot be treated. It has become more common in recent years. In 2021, there will be an expected 106,110 new cases and 7180 deaths. Melanoma is responsible for 5.6 percent of all new cancer cases in the United States [102]. With the beginning of the artificial intelligence era, the use of computers to assist doctors in diagnosis has increasingly outperformed manual approaches. Furthermore, the resemblance between skin lesion types and the similarity between different stages of the same lesion complicates the classification process. This also makes the classification of skin lesions common for deep learning studies in medical imaging. Skin cancer photographs are generally color RGB images, with the majority of them being dermoscopic. There are numerous publicly available skin disease data sets, each of which include a particular form of skin cancer. Table 7 summarizes common data sets including the types and number of images. It is difficult to assess human skin automatically since it varies in roughness, tone, hair mass, and so on, depending on geographical and living circumstances, climate, and hereditary variables. Recently, several public data sets, such as ISIC, that include multi-center data, have pushed the progress of skin disease recognition toward automation.

The dermoscopic image has a poor contrast, and the brightness impact will also have an effect on the image quality. This has an effect on the identification of lesions of various colors, sizes, structural forms, and textures, as well as lesions of various sorts. Usually we use filtering to remove hair in the image, commonly used are median filtering, Gaussian filtering, mean filtering, etc. Because the image size varies from dataset to dataset, an image size is usually specified to crop the image, and data enhancement operations such as rotation, translation and flip are also performed.

The automatic classification of skin diseases based on images has made significant progress with the advancement of deep learning. Fig. 10 depicts the basic classification procedure for skin diseases. For feature extraction, most studies employ CNN network architectures such as ResNet [23], Google-Net [25], VGG-Net [22], and other deep learning models mixed with transfer learning. Table 8 highlights the

Table 7
Skin cancer public data set.

Dataset	Classification
PH2 [103]	The data set contains 200 RGB images, including three categories: atypical moles (80 images), melanoma (40 images) and common mole images (80 images)
HAM1000 [104]	The data set contains seven different types of skin diseases: melanocyte nevus (12,875 images), actinic keratosis (867 images), dermatofibroma (239 images), basal cell carcinoma (3323 images), Vascular lesions (253 images), benign keratosis (2624 images), melanoma (4522 images)
ISIC MSK	It is a subdataset of ISIC. This dataset contains 225 RGB dermoscopic images, acquired from various international hospitals with the help of different devices.
ISIC UDA	It is another subdataset of ISIC. We have collected 557 images having 446 training and 111 testing samples from ISIC-UDA dataset.
ISBI 2016 [105]	The data set is mainly divided into melanoma benign and malignant. The data set contains 273 malignant images and 1006 benign images.
ISBI 2017 [106]	It contains 2750 images, with 2200 training and 550 testing samples. The ISBI-2017 dataset has three disease classes: melanoma, keratosis and benign.
ISIC 2018	The data set contains seven different types of skin diseases: melanoma (271 images); melanocytic nevus (2061 images); basal cell carcinoma (151 images); benign keratosis (345 images); dermatofibroma (36 images); Actinic keratosis (113 images); Vascular disease (45 images)

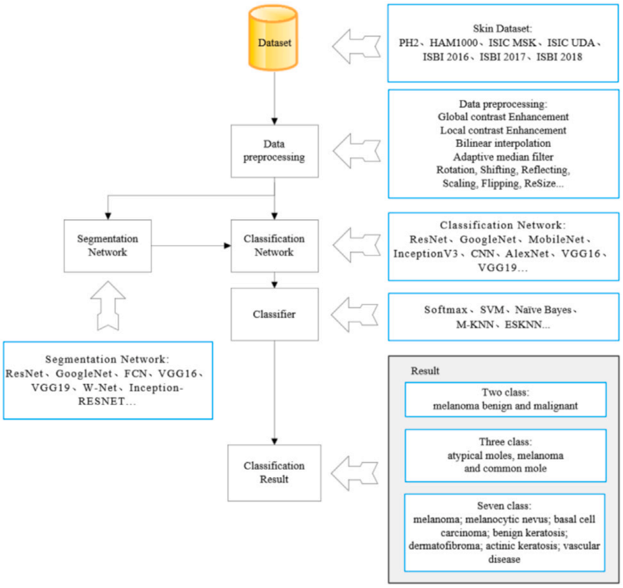

Dataset
Skin Dataset:
PH2、HAM1000、ISIC MSK、ISIC UDA、
ISBI 2016、ISBI 2017、ISBI 2018
Data preprocessing
Data preprocessing:
Global contrast Enhancement
Local contrast Enhancement
Bilinear interpolation
Adaptive median filter
Rotation, Shifting, Reflecting,
Scaling, Flipping, ReSize...
Segmentation Network
Classification Network
Classification Network:
ResNet、GoogleNet、MobileNet、
InceptionV3、CNN、AlexNet、VGG16、
VGG19...
Classifier
Softmax、SVM、Naïve Bayes、
M-KNN、ESKNN...
Segmentation Network:
ResNet、GoogleNet、FCN、VGG16、
VGG19、W-Net、Inception-
RESNET...
Result
Two class:
melanoma benign and malignant
Three class:
atypical moles, melanoma
and common mole
Seven class:
melanoma; melanocytic nevus; basal cell
carcinoma; benign keratosis;
dermatofibroma; actinic keratosis; vascular
disease
Classification
Result

Fig. 10. Basic flow chart of multiple skin lesions recognition.

direct classification methods. The classification algorithm determines the type of skin lesion based on the retrieved feature feedback. Gessert et al. [107] proposed to use dermoscopic images and additional patient metadata to solve the skin lesion classification problem by selecting a series of deep learning models (including EfficientNets, SeNet and ResNet WSL) through search strategies. Jasil et al. [108] used convolutional neural network to classify skin lesions based on transfer learning and used three popular architectures: InstantV3, VGG16 and VGG19 to obtain 74%, 77% and 76% accuracy respectively. Bhardwaj et al. [109] used algorithms based on “Average of *N* models” using InceptionV, Deep-CCN and MobileNet models. Replace the last Softmax function with a support vector machine (SVM). The final accuracy, precision and recall reached 86%, 80% and 60%, respectively. Ratula et al. [110] used transfer learning on four architectures: VGG16 [22], VGG19, MobileNet [111] and Inception V3 [112]. Using the HAM10000 data set for training, verification and testing, the classification accuracy rates reached 87.42%, 85.02%, 88.22% and 89.81%, respectively. Sevli et al. [113] proposed a method to classify 7 different skin lesions in the HAM10000 dataset based on the CNN model. The classification accuracy rate of this model is 91.51%. Bian et al. [114] proposed a multi-view filter transfer learning (MFTL) method for cross-domain skin lesion classification, using ResNet as the network structure to extract features, and measure the contribution of features to the target domain based on Wasserstein distance. Using softmax as the classifier, an accuracy of 91.8% was obtained on the ISIC 2017 data set. Alqudah et al. [115] used GoogLeNet, AlexNet, transfer learning and optimized gradient descent adaptive momentum learning rate (ADAM). They applied them to segmented and non-segmented lesion images from the ISIC database. The segmented images had the highest classification accuracy of 92.2%. Andre et al. [116] applied Google-Net, a deep learning model developed by Google, to a dataset containing 129,450 clinical images. The overall accuracy was 72.1 ±0.9%. Sarkar et al. [117] proposed a melanoma classification algorithm based on depth separable residual convolution. The accuracy on the PH2 dataset was 99.50% and the sensitivity was 100%.

As shown in Table 9, several articles employ the strategy of first segmentation and then classification to identify the region of interest. The acquired features are then classified using various classifiers. To improve the contrast of the original dermoscopic images, Afza et al. [118] first fuse locally and globally enhanced images. First, segment the lesion image, then use ResNet-50 to learn features via transfer learning. The collected features are further optimized by the modified grasshopper optimization algorithm after being extracted from the global average pooling layer. The features are then classified by a Naive Bayes classifier, which achieves 95.40%, 91.1% and 85.50% accuracy on the three datasets (PH2, ISBI2016 and HAM1000), respectively. Khan et al. [119] proposed a color optimization method (OCF) for skin lesion segmentation. Then a deep convolutional neural network model (DCNN) is used for classification. This model is inspired by the VGG and AlexNet models. DCNN is evaluated on ISIC 2016, ISIC 2017 and ISIC 2018 datasets. The classification accuracies reached 92.1%, 96.5% and 85.1%, respectively. Akram et al. [120] used mean-based methods for segmentation. DenseNet 201, Inception-Resnet-v2 and Inception-V3 are adopted for discriminative feature selection. The classification accuracies on the four datasets of PH2, ISIC MSK, ISIC UDA and ISBI 2017 are 98.8%, 99.2%, 97.1% and 95.9%, respectively. Jayapriya et al. [121] used a hybrid VGG-16 and GoogLeNet neural network framework to improve segmentation accuracy. The classification uses a combination of deep residual networks and handcrafted features. Feature extraction is performed on the segmented lesions. Support vector machines are used for classification. The accuracy rates on ISBI 2016 dataset and ISIC 2017 dataset reach 88.92% and 85.3%, respectively. Khan et al. [122] first used the segmentation structure of MASK Recurrent Convolutional Neural Network (MASK R-CNN). In the classification stage, a 24-layer convolutional neural network structure is designed. Finally, the optimal CNN features are provided to the Softmax classifier for final classification. It achieves 86.5% accuracy on the HAM10000 dataset. Ge et al. [123] used a fully convolutional neural network (FCN) and a specific convolutional neural network (CNN). FCN is used for segmentation. CNN is used on the segmented raw image ROI to extract features. Finally, the fully connected layer is used to judge

Table 8
Statistical table of multiple skin lesions recognition methods.

Class	Dataset	DL model	Performance metrics
7: AK, BCC, BKL, VASC, MEL, NV, DF	HAM10000	InceptionV + Deep-CCN + MobileNet [110]	Accuracy: 86% Precision: 80% Recall: 60%
		VGG16 [110]	Accuracy: 87.42%
		VGG19 [110]	Accuracy: 85.02%
		CNN [113]	Accuracy: 91.51%
	ISIC 2018	InstantV3 [108]	Accuracy: 74%
		VGG16 [108]	Accuracy: 77%
		VGG19 [108]	Accuracy: 76%
3: Melanoma, Keratosis, Benign	ISIC 2017	ResNet-50[114]	Accuracy: 91.8%
		AlexNet [115]	Accuracy: 92.2%
		GoogleNet [115]	Accuracy: 89.8%
		Google Inception [116]	Sensitivity: 96%
		V3 CNN [116]	Specificity: 94%
2: Benign, Malignant	ISIC	ResNet [117]	Accuracy: 99.50%
	PH2	ResNet [117]	Accuracy: 96.77%

Table 9
First segmentation and then classification skin multi-lesion recognition method statistical table.

Segmentation mode	Detection mode	Dataset	Performance metrics	Papers
Simple linear iterative clustering (SLIC)	ResNet-50	PH2 ISBI2016 HAM1000	Accuracy: 95.40% Accuracy: 91.1% Accuracy: 85.50%	[118]
optimized color feature (OCF)	DCNN	ISBI 2016 ISBI 2017 ISBI 2018	Accuracy: 92.1% Accuracy: 96.5% Accuracy: 85.1%	[119]
Mean and mean deviation based segmentation	InceptionV3	PH2 ISIC MSK ISIC UDA ISBI-2017	Accuracy: 98.8% Accuracy: 99.2% Accuracy: 97.1% Accuracy: 95.9%	[120]
VGG-16 GoogLeNet	DRN SVM	ISBI 2016 ISBI2017	Accuracy: 88.92% Accuracy: 85.3%	[121]
FCN	CNN	ISBI 2016	Accuracy: 93%, Sensitivity: 92%, Specificity: 94%	[123]
Hybrid Fusion CNN	CNN	HAM10000 ISIC2018 ISIC2019	Accuracy: 87.0% Accuracy: 94.6% Accuracy: 91.3%	[125]
CNN	KELM	HAM10000	Accuracy: 90.7%	[126]

the benign and malignant melanoma. Khouloud et al. [124] chose to segment in the classification step first. Two new deep learning network structures, W-Net and Inception-RESNET, are used. The segmentation and classification problems are solved separately. On the PH2 dataset, the sensitivity, specificity, accuracy and precision reached 98.5%, 99%, 98.50% and 97.5%, respectively. Muhammad et al. [125] first segmented the lesion region using a hybrid segmentation model made up of two segmentation networks, fused the findings using joint distribution probability and marginal distribution function, and then used a 30-layer CNN to classify. Some researchers concentrated on feature selection after segmentation as well [126]. The feature selection of the lesion area is carried out using the evolutionary algorithm to screen the most discriminant features and then classify them in order to improve the overall classification results. Furthermore, there is still some work being done to improve the effect of lesion segmentation [127].

3.5. Lung multiple-lesion recognition

Lung lesion is one of the most prevalent infectious disorders in clinical medicine, with a short start cycle and complex etiology, affecting infants and older persons with inadequate immunity the most. According to the World Health Organization, over 800,000 individuals died from lung lesions globally in 2016. As a result, prompt detection and treatment of lung abnormalities is critical.

In several clinical studies, CT scans are used to analyze and detect features of lung disease, as CT provides a clearer representation of the morphology and nature of lung shadows and masses, providing important evidence for classification and severity assessment of lung lesions. On the other hand, chest X-rays can be used to identify lung lesions, which is a more practical and cost-effective solution, although not as sensitive as chest CT images, X-ray imaging devices are ubiquitous in hospital emergency rooms, public healthcare facilities and even rural clinics. Some examples of X-ray, ultrasound and CT images of pulmonary lesions are shown in Fig. 11.

With the development of artificial intelligence technology, several researchers have applied machine learning approach to the detection and diagnosis of multiple-lesions in the lung. The machine learning approach can accurately measure the cumulative pneumonia load of the disease by quantitatively analyzing the morphology, extent, density, and other key imaging features of the lesion, which helps physicians quickly determine the clinical condition.

In previous work, the task of identifying multiple lesions in the lung has focused on the multiple-lesion classifications for new coronary pneumonia, other types of pneumonia and healthy individuals. Julián et al. [130] proposed a method for evaluating different deep neural networks using chest X-ray images to distinguish between control, pneumonia or COVID-19 groups, and the network structure was also reconstructed to allow gradient-based localization estimation, which was used to find interpretable models after training. Chowdhury et al. [131] evaluated eight different popular and previously reported CNN-based methods with transfer learning techniques for the classification of healthy and pneumonia patients using chest X-ray images. Horry

Table 10
Comparison of different diagnostic models for lung multi-lesions.

Modality	Network	Category	Results	Paper
X-ray	2D-CNN (COVID-Net2)	COIVD-19, Pneumonia, Normal	ACC = 0.915	Julia et al. [130]
	2D CheXNet	COIVD-19, Pneumonia, Normal	ACC = 0.9779	Choudhuri et al. [131]
	2D-CNN (CovAI-Net)	Pneumonia, Covid-19+ve, Covid-19-ve	AUC = 0.986 (Stage1), AUC = 0.972 (Stage2)	Mishara et al. [132]
	2D U-net + Faster RCNN	Exudation, Calcification, Nodule, Miliary tuberculosis, Free pleural effusion, Encapsulated pleural effusion	Montgomery: AUC = 0.977 and ACC = 0.926 Shenzhen: AUC = 0.941 and ACC = 0.902	Xie et al. [133]
CT	3D COVNet	COIVD-19, Pneumonia, Normal	AUC = 0.98 (COIVD-19) AUC = 0.95 (Pneumonia) AUC = 0.96 (Normal)	Li et al. [134]
	2D ResNet-50 (DRENet)	COIVD-19, Pneumonia, Normal	AUC = 0.95	Song et al. [135]
	ResNet-18	COIVD-19, Pneumonia, Normal	ACC = 0.867	Xu et al. [129]
	3D-ResNet (PARL)	COIVD-19, Pneumonia, Normal	AUC = 0.933	Wang et al. [136]
	2D-CNN (M3Lung)	COIVD-19, Pneumonia, Normal, H1N1	ACC = 0.9521	Qian et al. [137]
X-ray, CT	GAN + CNN (DL-CRC)	COIVD-19, Pneumonia, Normal	AUC = 0.9525	Sakib et al. [138]
	Genetic Deep Learning CNN	COIVD-19, Non COVID, Normal	ACC = 0.9884	Babukarthik et al. [139]
X-ray, Ultrasound, CT	2D-CNN (VGG19)	COIVD-19, Pneumonia, Normal	F1 = 0.87 (X-ray), F1 = 0.99 (Ultrasound), F1 = 0.78 (CT)	Horry et al. [128]

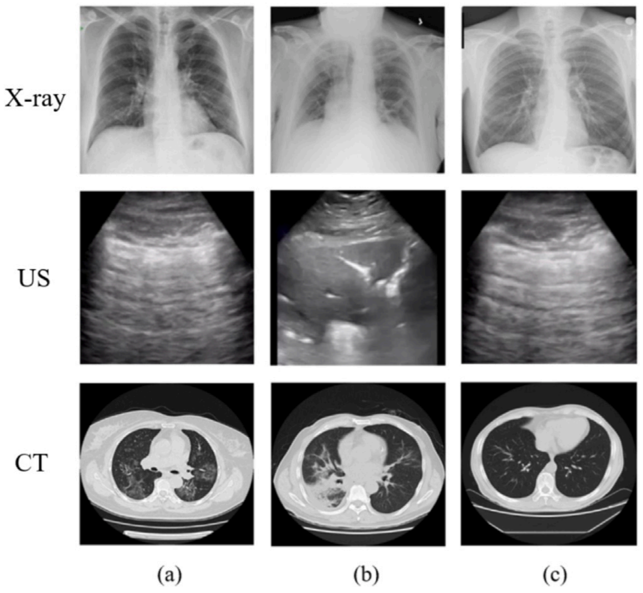


Fig. 11. Examples of X-ray, ultrasound and CT images [128,129]. The first row is a typical X-ray image, the second row is a typical ultrasound image, and the third row is a typical CT image of (a) COVID-19; (b) pneumonia; and (c) a healthy individual.

et al. [128] proposed to perform COVID-19 detection using CNN models of transfer learning techniques for images from image modalities such as X-ray, Ultrasound and CT scans.

In addition, other works have divided the multiple-lesion recognition task into a multiple-stage approach. Mishra et al. [132] proposed a two-stage intelligent diagnosis system for medical images, in which in the first stage, chest X-ray images were classified as pneumonia and non-pneumonia, and in the second stage, pneumonia images were further classified as Covid-19 positive and Covid-19 negative. Li et al. [134] proposed a U-Net-based segmentation method for extracting lung regions as regions of interest, and developed a neural network framework for detecting COVID-19 to extract both 2D local and 3D global representative features. Xie et al. [133] proposed an intelligent diagnostic system for medical images that can automatically segment lung regions by learning scalable pyramidal structures to detect segmented lung multiple categories of tuberculosis lesions, thus achieving the auxiliary analysis and diagnosis of tuberculosis. Song et al. [135] proposed a three-stage lung CT intelligent diagnosis system, first by extracting the main regions of the lung to avoid noise caused by different lung contours, then a detailed relation extraction neural network was designed to obtain image-level predictions, and finally, the image-level predictions were aggregated to achieve person-level diagnosis.

Based on the multiple-stage approach, some works incorporate additional attention techniques to improve lung multifocal recognition performance. Xu et al. [129] proposed to segment candidate infection regions from a pulmonary CT image set using a three-dimensional CNN model, and then categorized these separated images into COVID-19, Influenza-A viral pneumonia, and irrelevant to infection groups, together with the corresponding confidence scores using a location-attention classification model, and finally calculated the infection type and total confidence score of the case using a Noisy-or Bayesian function. Wang et al. [136] proposed to generate a soft lesion-aware map with remarkable lesion localization ability using the convolutional feature map of the detection subnetwork, and then this soft lesion-aware map is fed into the type classification subnetwork to make it pay attention to the lesion regions. Qian et al. [137] proposed a multi-task multi-slice CAD system with an attention mechanism for multiple-class lung pneumonia screening with CT imaging, which consists of two-dimensional slice- and patient-level CNN networks, the former aiming to seek feature representations from abundant CT slices and the latter one allowing recover the temporal information through feature refinement and aggregation between different slices.

Furthermore, SAKIB et al. [138] proposed a DL-CRC framework for data enhancement algorithm using COVID-19 data to generate synthetic COVID-19-infected chest X-ray images to enhance the robustness of the training model by adaptively employing GAN and data augmentation methods. BABUKARTHIK et al. [139] proposed an independent and continuous learning algorithm for generating a DCNN architecture spontaneously for extracting features for classifying them between COVID-19 and normal images.

In summary, a variety of studies have been proposed for diagnosis of pulmonary multilocular lesions with generally encouraging results, and the most recent studies related to this are listed in Table 10. As

Table 11
Comparison results of multiple abdominal disease detection and classification tasks.

Task type	Data	Modality and sample size	Metrics	Paper
Classification	Private	CT (182 case)	Sensitivity 0.857; Specificity 0.924	[140]
	Private	CT (145 case)	Accuracy 0.973; Sensitivity 0.964; Specificity 0.982	[141]
	Private	CT (172 case)	Accuracy 0.899; Recall 0.890; Precision 0.951; F1-score 0.920	[142]
	Private	CT (179 case)	Accuracy 0.76; Sensitivity 0.852; Specificity 0.608; PPV 0.825; NPV 0.587	[143]
	Private	US (1299 case)	Accuracy 0.856	[144]
	Private	Multiphase CT (206 case)	AUC 0.889	[145]
	Private	CT (206 case)	Accuracy 0.728	[146]
	Private	Endoscopic image (1331 case)	Accuracy 0.8922	[147]
	Private	Endoscopic image (787 case)	AUC (Ulcer/Cancer) 0.95/0.97; Accuracy (Ulcer/Cancer) 0.85/0.9	[148]
	Private	Wireless endoscopy image (12,000 case)	Recall 0.9907; Specificity 0.99; Precision 0.9905; AUC 0.993;FPR 0.003;Accuracy 0.993	[149]
	Private	Endoscopic image (8000 case)	Accuracy 0.9738; Recall 0.9715; Precision 0.9727; F1-score 0.9721	[150]
Detection	Public	CT (4427 case)	Recall 0.958; Precision 0.902; Dice 0.926	[151]
	Public	CT (4427 case)	Recall 0.915; Precision 0.932; Dice 0.898; AVD 0.349; VS 0.942	[152]
	Public	CT (4427 case)	Dice 0.741; Sensitivity 0.709	[153]
	Public	CT (4427 case)	Recall 0.883; Precision 0.947; Dice 0.912; Different axis 1.747/1.555	[154]
	Public	CT (4427 case)	Sensitivity 0.9071	[155]
	Public	CT (4427 case)	AUC 0.81	[156]
	Public	CT (4427 case)	Sensitivity 0.9249	[157]
	Public	CT (4427 case)	Sensitivity 0.9077	[158]
	Public	CT (4427 case)	Sensitivity 0.6128	[159]
Segmentation	Public	CT (4427 case)	mAP 0.64	[160]
	Public	CT (4427 case)	mAP 0.678	[161]
Segmentation, Classification	Private	CT (225 case)	Accuracy 0.98; Sensitivity 1; Specificity 0.9772; Jaccard index 0.95; DSC 0.9743	[162]
Labeling, segmentation	Public	CT (4427/1018/131/127 case)	Sensitivity 0.504	[163]
Detection, Segmentation	Public	CT (4427 case)	Sensitivity 0.702	[164]
Detection, labeling, segmentation	Public	CT (4427 case)	Sensitivity 0.84; AUC 0.9601; F1 0.4563; Distance /diameters average error 1.4138/1.96	[165]
Detection, Classification	Public Private	US (367 case /177 case)	AUC 0.916	[166]

a next step, screening studies with AI technology will help to detect early pulmonary multi-lesion and improve the diagnostic accuracy for radiologists. Likewise, AI technology severity prediction is very important and can help in clinical decision making for estimating ICU events or treatment planning.

3.6. Abdomen multiple-lesion recognition

Clinically, the abdomen contains many key human body functioning systems, such as the digestive and urinary systems, as well as numerous significant organs, such as the liver, kidneys, and stomach. The most common abdominal disorders are liver cysts, hepatitis, liver cancer, kidney cysts, kidney cancer, gastric ulcers, and so on. Different clinical evaluations of the abdomen are classified into two types. The first are examinations with a fixed imaging angle, such as CT and ultrasound, and the second is a particular examination with an adjustable multi-angle, such as endoscopes. Routine examinations are most commonly utilized for structural abnormalities and disorders such as cancer, cysts, and nodules. Endoscopy is most commonly used to examine the mucosa and surrounding tissues of the digestive tract, including the esophagus and gastrointestinal tract.

The use of deep learning technology in the diagnosis of many forms of abdominal disorders is mostly focused on lesion detection and disease classification. The challenge of segmenting multi-target images of multiple types of lesions is particularly difficult due to the extremely complex background of the abdominal cavity. There are currently few connected studies. More studies will begin with target detection and then proceed to lesion segmentation in ROI. Recently, there has been a lot of work done on target detection, including lesions in the abdomen area. This is primarily owing to the release of the DeepLesion public dataset, which contains eight distinct locations such as the lung, mediastinum, belly, liver, pelvis, soft tissue, kidney, and bone. The lesion has been identified. The research of disease classification focuses on the classification of benign and malignant tumors, the classification of tumor subtypes, and the classification of tumors and other benign lesions. Table 11 summarizes the work and results comparison of abdominal multiple-lesion recognition.

3.6.1. Recall improvement strategy

When compared to accuracy, the target recognition task in medical imaging frequently necessitates a greater recall criterion. As a result, the majority of the work's advancements are focused on improving feature fusion and screening, as well as network structure design. K. Yan et al. [165] completed the simultaneous training of detection, segmentation, and labeling for several tasks in their feature fusion and screening work, and suggested a 3D continuous context feature fusion technique that improved detection sensitivity. Gc A et al. [160] fused the discrepancies between multi-resolution features on this basis to improve the 3D context feature fusion procedure. Cai J et al. [163] used the multi-scale feature fusion extracted by the proposed anchorless proposal network and further carried out the feature re-fusion between

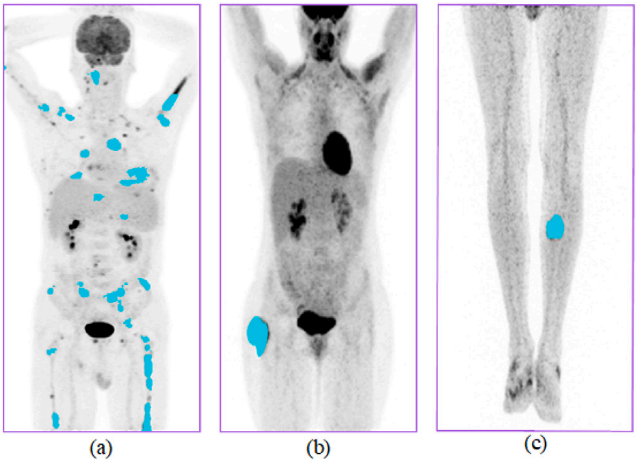


Fig. 12. Examples of Whole Body Multiple Diseases. Where (a) is lymphoma, (b)(c) is soft tissue sarcoma, and the blue area indicates the tumor region. The background is the whole body PET image.

different data sets, and they have a good performance on different data. Z. Wu and others [161] make full use of the shape information in the context to better fuse the features between slices and improve its utilization. The innovation of network structure mainly includes functional modules and multi-stage structures with new structural blocks as typical features. Refs. [151–153] combined channel attention and spatial attention mechanisms to design a variety of attention modules to improve detection sensitivity. Literature [154–156] combined with existing Faster-RCNN, VGG, and other networks to implement multi-stage detection tasks. The literature [157–159,164] further optimizes the index from multi-axial convolution kernel, overall network setting, combined graph cut, or traditional methods.

3.6.2. Accuracy improvement strategy

The majority of research on the classification of various abdominal disorders is focused on diverse diseases of the same organ. In the classification of liver-related diseases, the literature [140–142,162,166] first extracts the ROI of the target area through detection or segmentation and then fuses multi-model features for classification. In the study of kidney diseases, the jobs [143–146] use multi-phase enhanced CT data, multi-angle decomposition, and multi-model feature optimization methods to improve the accuracy of classification. In gastrointestinal diseases, due to the particularity of color dynamic images, the preprocessing process often only contains ROI interception, and the model is lightened from the perspective of parallel small models, real-time communication is possible, and the model can be applied to clinical practice [147–150].

4. Whole-body multiple-lesion recognition

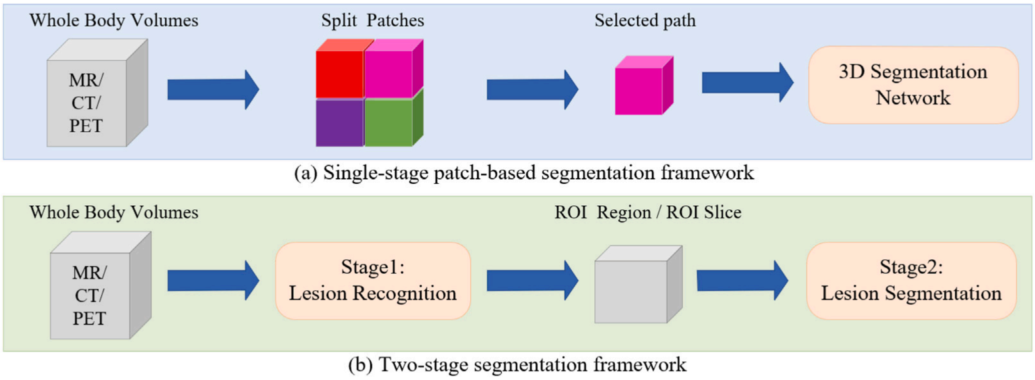
For some whole body multiple-lesions, such as lymphoma and soft tissue sarcoma. Lymphoma occurs in all lymphatic systems throughout the body, and soft tissue sarcoma occurs in the nodal tissues throughout the body. Fig. 12 shows examples of lymphoma and soft tissue sarcoma, where Fig. 12(a) shows an example of lymphoma and Fig. 12(b)(c) show examples of soft tissue sarcoma. The blue colored area is lesion. In clinical practice, PET images or PET-CT images are generally chosen as an auxiliary imaging technique for diagnosis because of the variable location. PET-CT images use the sensitivity of PET images for lesion to coarsely locate, while using the anatomical information of CT images to further accurately locate the lesion. Currently, most of the deep learning-based segmentation methods for whole-body multidisease are based on PET-CT images.

Deep learning-based whole-body multiple-lesion recognition can be divided into two strategies, a two-stage segmentation strategy and another single-stage patch-based segmentation strategy. Both strategies are designed to solve the problem of excessive computational burden caused by the need to input whole-body data for segmentation. Fig. 13 shows the flow diagram of the two segmentation strategies. The two-stage segmentation strategy includes two parts, lesion recognition and lesion segmentation, where the purpose of lesion recognition is to remove a large number of regions unrelated to the lesion, and then reduce the computational burden of lesion segmentation, and the results obtained in the first stage are generally called ROI regions or ROI slices. Table 12 shows the comparison of some of the two-stage methods. Generally speaking, the first stage of the network is more simple, because after the first stage of location is completed, the results of the first stage will generally be used as the central area to do external expansion operations to serve as input for the second stage. Xu et al. [167] used the traditional anatomical automatic recognition (AAR) method to obtain ROI regions. Hu et al. [168] used a simple 2D U-Net to obtain ROI regions. While Yuan et al. [169] directly manually selected the nasal, thoracic and abdominal regions where lymphoma may occur as ROI regions. In the second stage of segmentation, three methods were selected to perform lesion segmentation with different variants of the U-Net. The approach of Li et al. [170] differed from the first three in that an arbitrary deep learning model was used in the first stage to obtain a initial result, and then an adaptive level set was done based on this to obtain the precise segmentation result.

Single-stage strategy generally uses a 3D network to fuse lesion detection and segmentation. In order to solve the problem of computational burden proliferation caused by 3D networks, a patch strategy is generally adopted. The raw whole body image is chunked into several small patches and fed into the network, and then the results are concatenated in order to obtain the final segmentation result. Among them, the patch size used by Huang et al. [171] is the largest, but the network he used is a simplified version of U-Net that reduces the number of layers and feature maps. Table 13 shows the comparison of some of the single-stage methods.

5. How far are we from medical practice?

The approaches presented previously in this study are all strictly retrospective experiments, in which the public data set is separated into training and test sets, and the general criterion are determined on the test set. Although these methods have obtained outstanding performance in terms of the test set's evaluation criterion, whether they may be employed in medical practice and how their performance in medical practice will be investigated remains to be seen. The most straightforward method is to directly compare with an expert doctor. Chen et al. [176] compared the AI film reading system to radiologists in the detection of non-small cell carcinoma using lung CT, and discovered that while the sensitivity of the AI reading system was substantially greater than that of radiologists, the false positive rate was equally high. The sensitivity of the AI reading system was 94.12%, with a false positive rate of 22.22%. Radiologists had a sensitivity of 72.94% and a false positive rate of 7.93%. Similarly, Hayashida et al. [177] compared the AI system to 20 doctors in the BI-RADS multi-classification task based on breast ultrasound, and discovered that it performed significantly better than doctors. Kazemzadeh et al. [178] discovered that the sensitivity of AI systems in detecting active tuberculosis based on chest X-rays is higher than that of radiologists, but the false positive rate is also higher. Given deep learning's great sensitivity in the detection and classification of lesions, some researchers propose that an AI system be employed as an auxiliary tool to find missing lesions in radiation reports [179]. The model training data set and the clinical comparison data set come from the same source, making the examples mentioned above almost from the same source. However, in clinical practice, a specific medical institution may not be able to collect enough training



(a) Single-stage patch-based segmentation framework

(b) Two-stage segmentation framework

Fig. 13. Two different whole-body multiple-lesion segmentation frameworks.

Table 12
Summary of the two-stage segmentation method on whole-body multiple-lesion recognition.

Method	Dataset	Dice	Lesion recognition	Lesion segmentation
Xu et al. [167]	63 cases of lymphoma PET-CT datasets	0.77	Anatomical automatic recognition (AAR)	DiSegNet
Hu et al. [168]	80 cases of lymphoma PET-CT datasets	0.7115	2D U-Net	Adversarial Networks
Li et al. [170]	90 cases of lymphoma PET datasets	0.8796	Any deep learning model	Adaptive Level Set
Yuan et al. [169]	45 cases of soft tissue sarcoma PET-CT datasets	0.7303	Manual selection of the nasal, abdominal and thoracic regions where lymphoma occurs	Mixture Network

Table 13
Summary of the single-stage patch-based segmentation method on whole-body multiple-lesion recognition.

Method	Dataset	Dice	Patch size $(H \times W \times N)$	Segmentation network
Paul et al. [172]	733 cases of lymphoma PET-CT datasets	0.73	96 × 24 × 112	3D U-Net
Huang et al. [171]	173 cases of lymphoma PET-CT datasets	0.72	128 × 128 × 256	Slim U-Net
Li et al. [173]	80 cases of lymphoma PET-CT datasets	0.7284	128 × 128 × 6	DenseX-Net
Hu et al. [174]	109 cases of lymphoma PET-CT datasets	0.6664	48 × 96 × 480	3D ResUNet
Neubauer et al. [175]	51 cases of soft tissue sarcoma PET-CT datasets	0.772	256 × 256 × 16	DenseUNet

data. As a result, AI system mobility is an important consideration. Han et al. [180] demonstrated the robustness and generalization of the X-ray-based lesion detection CAD system by training on data collected by one institution and conducting clinical verification at another. However, Pedrosa et al. [181] performed the same experiment on the X-ray-based COVID-19 detection system and discovered that the model's performance in extraterritorial data is significantly lower than that of the source data set. This demonstrates that, before an AI detection or classification system is widely employed in medical practice, its robustness and generality must be prioritized. Furthermore, the imbalance of medical data has a significant impact on the functioning of AI systems. In a study of six classification issues based on digital otoendoscopic images, Cha et al. discovered that, despite data augmentation and other technologies, the AI system is significantly inferior than radiologists in cases of class imbalance [182].

Some researchers deploy AI detection or classification systems to actual medical institutions. Bridge et al. [183] deployed the MR-based stroke detection AI system to three medical institutions and compared its detection outcomes for a long time with radiologists. If radiologists' results were considered the ground-truth, their AUROC values in the three medical institutes were 0.964, 0.981, and 0.998, respectively. Hong et al. [184] introduced to the hospital CAD software based on X-ray detection of pneumothorax. After utilizing it for a while, they came to the same conclusion as Kazemzadeh et al. [178]. In order to test and evaluate the AI system based on X-ray for COVID-19 and other pneumonia detection, Sevli recruited 10 clinical volunteer doctors. Sevli eventually demonstrated that the system's accuracy in actual use achieved 99.06% [185].

Due of the difficulties of segmentation compared to classification and detection, it is far from medical practice. Andrea et al. [186] compared a CT-based head and neck organ segmentation system to clinicians and discovered that the highest overlap was 1.00 and the lowest overlap was 0.56. According to comparative investigation, the AI system performs as well as physicians in large organs but poorly in tiny organs. As a result, Shu et al. proposed that the large ones be segmented using AI while the little ones be segmented manually [187]. However, in clinical use, determining the size of the segmented object prior to segmentation is problematic. Therefore, some researchers use AI segmentation system as an auxiliary tool for manual segmentation. Zhong et al. [188] discovered that when segmenting head and neck tumors based on CT scans, the AI system's segmentation result is used as the first result, and then manual segmentation is performed, reducing the time it takes radiologists to segment a single case from hours to minutes. Shirokikh et al. [189] came to similar conclusions in brain tumor segmentation using MR. Furthermore, Min et al. [190] discovered that the Dice value of AI system segmentation results and radiologist manual segmentation results can be employed as a manual delineation quality assurance in clinical application.

6. Challenges and future directions

Although deep learning has made significant progress in the task of multiple-lesion recognition for computer-aided diagnosis, a comparison of various literature related to deep learning segmentation reveals that there are certain difficulties and challenges in the development and evolution of deep learning segmentation networks at this stage. The establishment of more accurate, efficient and robust recognition models based on deep learning ideas and methods still deserves more in-depth research. The current improvement in the quality of medical image multiple-lesion recognition is mainly due to the advantages of network models in image representation learning capability and the efficiency of processing large-scale data under existing computational techniques. Most of the current multiple-lesion recognition algorithms for medical imaging task scenarios cannot meet the requirements of medical applications, and the algorithms require a large amount of annotated data and repeated annotations.

Future medical imaging for multiple-lesion recognition requires more in-depth research in the following difficulties and challenges.

(1) Since tissues in medical images have no clear edges, textures, and colors like natural images, the visual textures of lesions and normal tissues are difficult to be distinguished from each other. Moreover, redundant background information near lesion areas can particularly interfere with the ability to represent target visual features. Due to the variability and complexity of lesions, there is also great variation in sample textures within classes. The analysis of the above computer-aided diagnosis of multiple-lesion in medical imaging can conclude that there are similarities in the models or methods that apply deep learning or the fusion of multiple algorithms for lesion recognition. And most of the current approaches are in the theoretical stage and have not been practically applied to clinical diagnosis.

(2) Different tasks in medical image analysis require different data annotations, and there are few datasets suitable for deep learning models. Moreover, medical image datasets are usually small in data size, and the size of training data directly affects the training effect of deep learning models, and little training data tends to cause overfitting, which makes the trained models perform poorly on other datasets. Current transfer learning techniques can migrate suitable samples from the source domain to the target domain by sample-based or feature-based transfer learning, but both methods are limited by the number of samples in the source domain.

(3) Unsupervised learning methods provide a convenient way to perform image classification tasks, such as avoiding misdiagnosis cases due to the personal factors of doctors without manual feature extraction. However, research on this approach is still mainly focused on the study of conditions with high incidence rates, i.e., a large amount of patient data, while research on some rare conditions is rare. This may be due to the self-learning and adaptive nature of deep learning, which requires a large amount of sample data for learning in order to more accurately acquire disease image features for further disease diagnosis.

In response to these difficulties, the following tasks are key to enabling further development of deep learning research in medical imaging.

(1) Medical image segmentation will encounter smaller inter-class differentiation and larger intra-class variability. It is an urgent problem to design a network that can focus the network attention on the target region and optimize the feature representation of the network for the characteristics of medical images. The attention mechanism can be introduced in the network to make the extracted target regions more compact and the distance between them and the background features as large as possible. We prefer to build a unified lightweight segmentation model, which can meet different performance and accuracy requirements. At the same time, it solves the problem that the current deep learning algorithms require very good hardware devices for the number of parameters and the amount of computation. On the other hand, it is also convenient for researchers to innovate in algorithms.

(2) In response to the insufficient amount of data, the research and application of unsupervised learning are enhanced based on supervised learning. Combining the image data generated by the generative adversarial network framework with the original data to participate in model training can improve the model performance, which is especially important for medical image analysis. How to reasonably divide the original data and the generated data to achieve optimal performance of the trained model is an important issue to be addressed now and in the future. The adversarial transfer learning can effectively increase the number of samples through a learning-based data augmentation method and is not affected by the number of samples in the source domain. At present, the method is only initially applied in medical image analysis and has not been fully and widely applied yet. However, it is expected to become a hot spot in this research field and an important direction for future research because of its obvious advantages and great potential.

(3) With the development of medical technology, the amount of medical image data will be even larger, and it is impossible to annotate all the data with limited medical resources. Therefore, semi-supervised learning algorithms have great prospects for future development and will obtain more attention and research. At present, in the semi-supervised medical segmentation field, the performance of existing algorithms is still far from the effect of fully supervised learning, due to the information in unlabeled data being difficult to be utilized. On the one hand, it is not accurate to define the relationship between data distribution and posterior by only simple assumptions, and semi-supervised learning strategies using certain specific priors cause some performance degradation on data from other distributions. On the other hand, even though most algorithms impose filtering constraints on the information in unlabeled data, the training process inevitably learns the wrong information, which leads to potential performance degradation. In summary, proposing new semi-supervised learning algorithms for deeper mining of useful information in unlabeled data is an important research direction for medical image segmentation tasks.

In addition, there are only two relatively large public datasets for multiple-lesion recognition. For the deep learning field, a large number of medical images usually need to be trained over a long period to achieve better performance, so more large-scale image datasets are needed to contribute to multiple-lesion recognition.

Of course, the current deep learning methods are also in the development period, and the maturity and breakthrough of each technology is the process of continuous experimentation, innovation and improvement. At present, both computer hardware technology and medical photography technology provide good basic conditions for deep learning to deal with medical imaging problems. In conclusion, the self-learning advantage of deep learning makes it possible to improve the classification effect while extracting features. It is believed that as the method continues to mature and improve, it will definitely provide better help for disease diagnosis in medical image analysis.

7. Conclusions

In this survey, we systematic summarize the deep learning based multiple-lesion recognition methods. As opposed to most existing reviews, we conduct the broadest survey, rather than covering only particular domains. In this paper, we review the recognition of multiple-lesion in multiple organs including brain, eye, skin, breast, lung and abdomen. Also, we review the whole-body multiple-lesion recognition methods. Then, we summarize and review the problems that still exist in the current deep learning-based methods for multiple-lesion recognition. Finally, this paper presents trends and opportunities for new developments around multiple-lesion recognition methods, and hoping that this survey will provide an scientific community background for future research.