Detection of Children's Molars Based on Noise Filter

Zhiwen Yan
South China Normal University
Ying Chen
South China Normal University
Jinlong Song
South China Normal University
Jia Zhu (jiazhu@zjnu.edu.cn)
Zhejiang Normal University
Jianbo Li
Southern Medical University

Research Article

Keywords: molars, noise filter

Posted Date: January 11th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1200020/v1

Detection of children's molars based on noise filter

Zhiwen Yan[1], Ying Chen[1], Jinlong Song[1], Jia Zhu[2,*], and Jianbo Li[3,*]

[1]School of Computing Science, South China Normal University, Guangzhou, China.
[2]Key Laboratory of Intelligent Education Technology and Application of Zhejiang Province, Zhejiang Normal University, Zhejiang, China.
[3]Stomatological Hospital, Southern Medical University, Guangzhou, China.
[*]jiazhu@zjnu.edu.cn, mz04ljb@139.com

ABSTRACT

Pit and fissure sealant is for children aged seven to twelve years to prevent molars from becoming caries. In this paper, we propose a new detection framework to identify whether children need pit and fissure sealing. We divide the framework into two parts: molar detection and molar classification. According to the characteristics of teeth, we propose to use the clustering method to filter the bounding box in the object detection part. In the region divided by clustering, we only keep one detection frame in the same category. In the classification part, we propose a noise filtering layer based on wavelet transform for feature extraction. We map the training samples to another space in the training process based on metric learning to increase the distance between categories and improve the accuracy of classification.

Introduction

Most of the children in primary and permanent tooth replacement period are at high risk of dental caries. In this period, the newly erupted permanent teeth of children are not fully developed, such as enamel mineralization and low resistance to acid substances secreted by bacteria. These problems make children vulnerable to dental caries. The results[1] showed that 90% of molars caries occurred in pit and fissure, and pit and fissure sealing was the most effective method to prevent pit and fissure caries. On the one hand, the pit and fissure are closed that can help the tooth surface is easier to clean, and the external cariogenic bacteria can not enter. On the other hand, the nutrient source of the original bacteria was cut off, and the bacteria died gradually. The purpose of our model is to detect whether molars have caries and whether children need pit and fissure sealant.

Most researchers judge dental caries by X-ray images[2,3]. Our dataset is taken by the patient himself, and the images are labeled by professional doctors. Hence, our model not for clinical diagnosis. We provide the app for users to use on their mobile phones. After users upload their photos, we analyze the photos and then feedback the information to users. The research we have done is for practical application, and it is a new exploration. Therefore, we seek the cooperation of the hospital to get the dataset. After the hospital annotates the data. Most researchers[4–7] use artificial intelligence technology to study medicine. We are also study the deep learning model for this dataset. The image category of the dataset is shown in Fig. 1.

In this paper, we propose a new framework for this dataset. We divide the framework into two parts. The first part is an object detection model, which aims to identify the molars in the image, and then cut the identified molars. However, results of molar detection have the situation that multiple bounding boxes to mark the same tooth. Therefore, we propose to cluster teeth according to the position information of the teeth. We keep only the bounding box with the highest score in each category, and the score comes from the score got during object detection. The second part is the classification of molars. We put the molars of the first part into the classification model. We divide the categories into three categories. The one is the need for pit and fissure sealing, the other is the do not need pit and fissure sealing, and the last is caries. For a detailed introduction of these three categories, please see Section 4.1 in part IV. However, the different category of tooth characteristics are difficult to distinguish, and the classification effect of the simple classifier is not ideal. Some researchers[8] have also proposed using filters to remove image noise to improve image quality. According to the idea of DWT[9] and residual shrinkage network[10], we design a noise filter layer to filter irrelevant information during feature extraction. According to the idea of metric learning, we improved the EPNet[11] to encode the known category information and the unknown category information together, and then classify the results of the unknown category. Feature encoding increases the distance between categories, which is more conducive to classification. Our dataset and source code will be published later.

In summary, our main contributions are as follows:

1) We propose a noise filter layer. The noise filtering layer adds the small deep learning network to the wavelet transform, and the layer achieves the filtering of image noise by learning the threshold of filtering noise.

2) We apply the improved EPNet in metric learning to general classification tasks. We use the improved EPNet to

Figure 1. The teeth in the bounding boxes represent the molars that the model needs to detect. (1) Teeth are caries. (2) Teeth need to have pit and fissure sealing. (3) Teeth do not require to have pit and fissure sealing.

classification after encoding for feature, which shows a significant effect in the experimental result.

3) We provide a new dataset in stomatology. The dataset is about seven years old children's molar dataset, which is helpful to explore the development of children's tooth health problems by the deep learning method.

Related works

The main task of object detection is to find all the objects of interest in the image, including two subtasks: object location and object classification. There are two methods of object detection. The first is the two-stage method, in which some researchers have proposed the R-CNN[12] algorithm of pre-selecting candidate regions (ROI). This method merging regions with similar in the image, which reduces the amount of calculation. However, due to the overlapping of ROI, many repeated feature extraction and calculation are carried out, so fast R-CNN[13] is proposed. The model extracts the features of the whole image once, and then ROI is generated and pooled. The region proposal network is used instead of the candidate region method to generate ROI faster. Although the two-stage method has high accuracy, they are inefficient. The second is the one-stage method. For the sliding window detector, the window can be regarded as the initial guess, and the boundary and category can be predicted at the same time. In recent years, yolov5[14] and SSD[15] are also one-stage detectors. Based on the peculiarity of the oral cavity dataset, we choose yolov5 as our object detection model. At the same time, we modified yolov5 to make it more suitable for our practical use.

Wavelet transform comes from signal theory. Wavelet transform is a local transform of space and frequency, so it can extract information from signal effectively. The wavelet transforms in the image are generally used in image compression, image denoising and so on. However, some researchers[9] proposed to use the wavelet transform to replace the pooling layer and achieved excellent results. Inspired by the residual shrinkage transform network[10] and wavelet transform, we put forward a noise filter layer by combining the two.

Metric learning is to learn the metric distance function for a specific task according to different tasks. The method of metric learning calculates the similarity between the images of all known tags and an unknown tag image to determine what category the unknown tags belong to. This kind of method has been used in small sample classification. Some researchers[11] propose to use metric learning to encode the distance between a set of image features, and then use the encoded features for classification, which can expand the distance between different categories and improve the classification task. We apply this method to the general classification task and get excellent results.

Proposed method

All methods were performed in accordance with the relevant guidelines and regulations of Scientific Reports. All experimental protocols were approved by Stomatological Hospital of Southern Medical University. We confirming that informed consent was obtained from all subjects and their legal guardians. Our framework is divided into two parts to introduce: molar detection and molar classification. The frame diagram is shown in Fig 2.

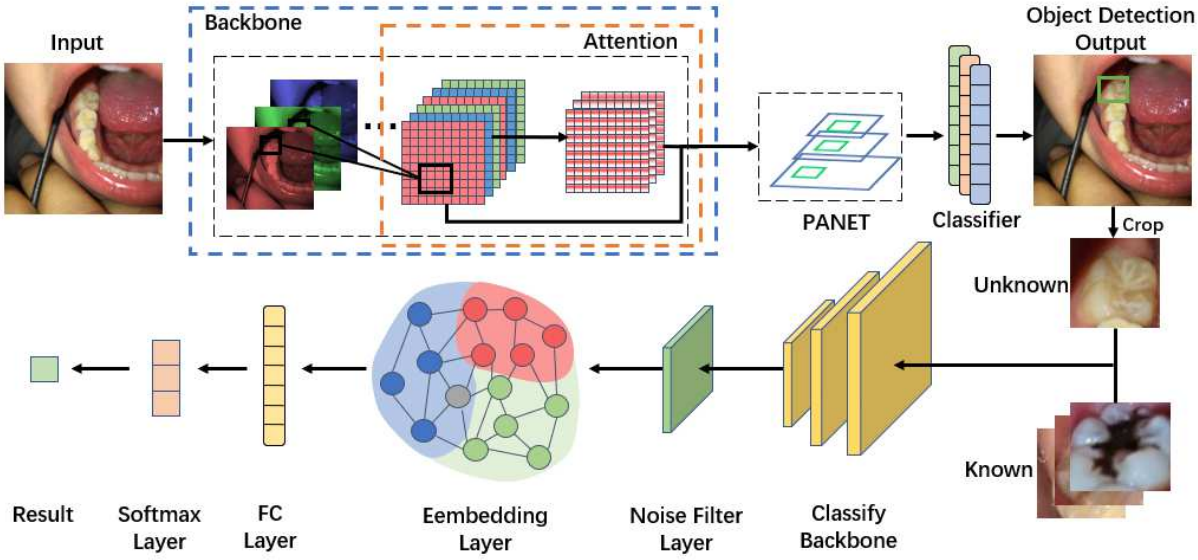

Figure 2. Flow chart of molar detection framework.

Molar Detection

In this paper, our object detection model is improved on the basis of yolov5[14]. The main purpose of molar detection is to identify the first molar and the second premolar. Identifying the second premolar can increase the relative position information to help identify the first molar. We also do a comparative experiment in Table 1. Our improvement in yolov5 has two parts. In the first part, the original yolov5 result recognizes the object outside the oral cavity as the tooth in, so we use the attention mechanism to limit the recognition area in the oral cavity. We add attention mechanism between backbone model and PANET, the high-dimensional features extracted from backbone model are weighted by attention mechanism to obtain the final high-dimensional features, which selectively giving higher weight to the image in the target area. The specific formula of attention mechanism is as follows:

$$F^S = \sigma\left(w_1\left(w_0\left(F^C_{avg}\right)\right) + \left(w_1\left(w_0\left(F^C_{\max}\right)\right)\right)\right), \tag{1}$$

$$F = \sigma\left(f^{7*7}\left(\left[F^S_{\text{avg}}; F^S_{\max}\right]\right)\right), \tag{2}$$

where F^C_{avg} and $F^C_{\max}$ are the results of putting feature f into average pooling layer and maximum pooling layer respectively. Here, w_1 and w_0 are learnable parameters, and σ is the sigmoid function. The feature F^S obtained from the channel attention mechanism is multiplied by F element by element, and then put into the average pooling layer and the maximum pooling layer to obtain F^S_{avg}, $F^S_{\max}$. Among them, [;] denotes the concatention of inputs. f^{7*7} is a 7x7 convolution layer.

In the second part, we modify the bounding box screening mechanism of object detection. In the test phase of object detection, it is possible that over one detection box can mark the same tooth at the same time. We encode the position information of the first molars and then cluster them. In the same region, we select only the teeth with the highest score in the object detection for each type of teeth, and the number of clusters is less than 4. Specifically, in the process of inference, we first filter the bounding boxes according to the confidence level of the detected objects, and only the bounding boxes larger than the threshold are retained. Then, according to the category (molar1 or molar2) of the objects in the bounding box, they are divided into two groups. For each group, the MeanShiftmethod[16] is used for clustering. For each sample that has not been visited, we classify the data within the threshold radius into one group with the sample as the centre. According to the vector

sum of all samples in the current category relative to the sample centre, the class centre is adjusted until the sample centre does not change too much. In a clustered category region, only keeps one data in each category of objects, which is in line with the actual situation.

Classification of Molars

Due to the features of the first molars are difficult to distinguish, the general classification method is not ideal. Therefore, we improve the model from two aspects. On the one hand, because there is too much irrelevant information in the tooth image, we need to filter the image noise. Consequently, we propose a noise filtering layer based on wavelet transform and residual shrinkage network. We first decompose the obtained feature layer and retain the low-frequency information after decomposition, because the low-frequency information contains the key information of the image and represents the smooth part of the image. The high-frequency information often contains image noise and details. First, we use DWT to decompose the image features to obtain high-frequency information and low-frequency information. Then, we use the soft threshold filtering of the residual shrinkage network[10] to filter the high-frequency information noise. Finally, we use IDWT to combine high-frequency and low-frequency information. The size of image features does not change before and after filtering. The specific formula is as follows:

$$\mathbf{X}_{ll} = \mathbf{L}\mathbf{X}\mathbf{L}^T, \mathbf{X}_{lh} = \mathbf{H}\mathbf{X}\mathbf{L}^T, \tag{3}$$

$$\mathbf{X}_{hl} = \mathbf{L}\mathbf{X}\mathbf{H}^T, \mathbf{X}_{hh} = \mathbf{H}\mathbf{X}\mathbf{H}^T, \tag{4}$$

$$\mathbf{Y}_{ij} = sgn(\mathbf{X}_{ij})S(R(R(\mathbf{X}_{ij}\mathbf{W}_1)\mathbf{W}_2)), \tag{5}$$

$$\mathbf{Y} = \mathbf{L}^T\mathbf{X}_{ll}\mathbf{L} + \mathbf{H}^T\mathbf{Y}_{lh}\mathbf{L} + \mathbf{L}^T\mathbf{Y}_{hl}\mathbf{H} + \mathbf{H}^T\mathbf{Y}_{hh}\mathbf{H}, \tag{6}$$

where L and H are the low frequency and high frequency filters of Haar in wavelet transform, which can be seen in[9]. Here, R represents the relu activation function, and S represents the sigmod activation function.

On the other hand, we encode the extracted image features, which can increase the space between categories and make the image easier to classify. The coding here is modified on the basis of EPNet[11]. There are five steps to encode features. The first step is to calculate the similarity between features and construct a similarity matrix. The similarity formula is as follows:

$$A_{ij} = \exp\left(-\left\|x_i - x_j\right\|^2 / 2\sigma^2\right) \text{ if } i \neq j, \tag{7}$$

where x_i and x_j represent different image feature. According to EPNet, $\sigma^2 = Var(\left\|x_i - x_j\right\|^2)$ which help to stabilize training. Here $A_{ii} = 0$.

The second step is to compute the Laplacian of the adjacency matrix,

$$L = D^{-\frac{1}{2}}AD^{-\frac{1}{2}}, \quad D_{ii} = \sum_j A_{ij}, \tag{8}$$

where D represent a diagonal matrix.

In the third step, we use the formula of propagation matrix in label propagation formula[17],

$$P = (I - \alpha L)^{-1}, \tag{9}$$

where I is the identity matrix, and $\alpha \in \mathbb{R}$ which in our experiment set to 1.

The fourth step is to calculate the score matrix of coding features in different categories.

$$Z = PY, h_i = \arg\max_j Z_{ij}, \tag{10}$$

where Y is a matrix of size N×C, and each row represents the one-hot encoding of the category label. If the tag data is unknown, the value of the data corresponding to row all zeros. Here, arg max is a function which can calculate the subscript with the largest value in a group of data, and h_i represents the category label most similar to the input data.

In the fifth step, we use relation value and similar category information to encode features.

$$\widetilde{\mathbf{x}}_i = \sum_j P_{ij} Y_{j,h_i} \mathbf{x}_j, \text{ if } i \neq j, \tag{11}$$

$$\widetilde{\mathbf{x}}_i = \widetilde{\mathbf{x}}_i + x_i. \tag{12}$$

Our $\widetilde{\mathbf{x}}_i$ feature represents the weighted sum of neighbor nodes of its similar category.

EXPERIMENTS

Dataset

The dataset is collected by a hospital. All images are voluntarily submitted by patients to the hospital and agreed to be used for research. Moreover, the images are annotated by professional doctors. As shown in Fig. 1, the hospital marks the first molar and the second premolar to get the coordinates of these molars. Then, the doctors classified the first molars into three categories: caries, need pit and fissure sealing and do not need pit and fissure sealing. The appearance of cavities on the surface of molars belongs to the caries category. The molars that have cavities, obvious pit and fissure and filling agent on the surface of molars are belonged to the pit and fissure sealing category. The molars that do not erupt completely, the surface of molars is relatively flat, the molars have done pit and fissure sealing, and the molars have teeth filling agent are belonged to the do not need pit and fissure sealing category. The oral cavity dataset has 4146 images, We use 80% of all images for training, while the rest 20% for testing. The molars dataset has 2957 images. The dataset of molars is obtained by cutting the labels in oral cavity dataset. Doctors had selected and removed the unclear molars in the oral cavity. Among them, 991 caries images, 1215 needed pit and fissure sealing images, and 763 did not need pit and fissure sealing images. We use 80% of all images for training, while the rest 20% for testing.

Evaluating Indicator

The indexes of object detection are mAP, mAP50, mAP75, mAR, mAR50, mAR75. AP means that x is taken as [50,95], and 10 values with the interval of 5 are taken as the average statistical results. AP_X represents the proportion of correct predictions in all bounding boxes when the IOU value of the detection box and the label box is greater than x%. AR_X represents the accuracy rate when the IOU value of the detection box and the label box is greater than x%. In addition, in order to measure whether an object is repeatedly detected in the test results, we also define the evaluation index T, which is formulated as follows:

$$T = \frac{correct^2}{total \cdot detect}, \tag{13}$$

where *correct* represent the number of detection boxes that have been identified correctly, *total* represents the number of labelled detection boxes, *detect* represent the total number of bounding boxes output by the model. On the one hand, the more labelled boxes are correctly identified, the larger the index value will be. On the other hand, if the label box is repeatedly recognized by multiple bounding boxes, the index value will be reduced.

The indexes of image classification include Accuracy, Precision and Recall. Precision and Recall are averaged based on the values obtained for each category.

Implementation Details

Our training is divided into two parts. The first part is object detection. The backbone network model we used in object detection is yolov5. At the same time, we train with the improved attention module. In the training process, batch size is 32, epoch is 200, and Adam optimizer is used. The MeanShift method[16] is only used in the inference process. In addition, we named the second premolar as molar1 and the first molar as molar2 in the experiment.

The second part is molar classification. The main network used in molar classification is DenseNet-121[18]. In the process of training, object detection and molar classification are trained separately, and the images of molar classification are directly cut out from the marked oral cavity images. In this paper, the batch size of the training is 64, epoch is 200, and Adam optimizer is used. We first extract 21 images from different categories in the data as known categories, and then extract the images in

a category randomly from the remaining data as the location category. We extract all the categories with the largest number of categories as an epoch. If the number of other categories is not enough, the data will be extracted in a cycle. We put the extracted data into DenseNet. We add noise filter layer after every block of DenseNet, and replace pooling layer with DWT Pooling layer[9]. At the same time, we add EPNet[11] to the last full connection layer of DenseNet for image feature coding. In the test stage, we first extract the features of all training images before entering the EPNet layer, and then cluster the features of each category separately. We use k-means clustering to sum and average the data obtained from each cluster to get the center points of each category in the cluster. A category could produce many centre points. We save these central points and then putting these points into EPNet together with the unknown-image to classify in the test stage.

Result Analysis

Table 1. Ablation Experiment of the model. A means that the attention mechanism is added to the model. O means only the first premolar was detected. M means that mean shift method has been applied.

Method	Baseline	Baseline(O)	Baseline(A)	Baseline(A+M)
AP	46.2	44.5	46.7	46.7
AP_50	92.5	37.2	93.2	**93.2**
AP_75	40.2	89.7	40.8	40.8
AR	48.4	46.7	48.6	48.6
AR_50	99.7	98.1	99.7	**99.7**
AR_75	70.3	64.3	70.8	70.8
AP_molar1	92.7	90.1	93.5	**93.5**
AP_molar2	94.3	_	95.4	**95.4**
T	0.848	0.803	0.848	**0.908**

As shown in Table 1, the experimental results have a significant improvement after the original object detection model add attention mechanism. However, the object detection model with the clustering method did not improve the indicator other than T. Due to the universal detection index of object detection is only used to detecting the target object, which does not consider the situation that different detection boxes to identify the same object. We propose the T indicator for this problem, and it can better reflect the situation that detects the target object in actual use. As shown in Fig. 3, we also put the images identified as special errors in baseline into our modified model to detect, and its result was correct.

Table 2. Results of different classifiers. In the table, A stands for accuracy, P stands for precision, and R stands for recall.

Model	A(%)	P(%)	R(%)
DenseNet-121[18]	82.1	81.6	82.4
ResNet-18[19]	78.6	78.3	78.8
ResNet-34	79.2	78.7	79.9
MobileNetV3[20]	80.6	80.0	81.3
WRN[21]	81.2	80.1	81.5
Our	**87.6**	**88.4**	**87.3**

In the experiments of the classification model, we conducted two groups of comparative experiments. As shown in Table 2 and Table 3, they are the comparative experiments of general classifiers and our classifiers and the ablative experiments of our classifiers. We can see from the table that our model is better than others, which shows the validity of the methods used in our classification model. To demonstrate which method was effective in the experiment, we performed eight ablation experiments.

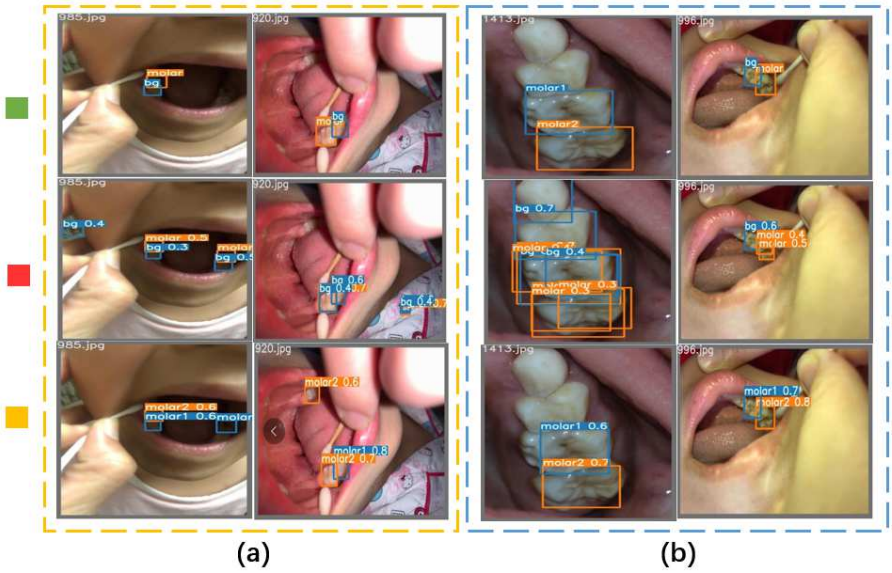

Figure 3. Visual analysis graph of experimental results. The green rectangle represents the original annotated image, the red rectangle represents the result of the baseline model, and the yellow rectangle represents the result of our model. (a) The result analysis graph of the sample which the same tooth is marked for multiple detection boxes. (b) The result analysis graph of the sample which model mistakenly recognizes objects outside the oral cavity as teeth.

Table 3. Ablation experiment of classifier.

	1	2	3	4	5	6	7	8
DWT Pooling		✓			✓	✓		✓
Noise Filter			✓		✓		✓	✓
EPNet				✓		✓	✓	✓
ACC(%)	82.1	83.7	84.2	83.4	84.6	85.7	86.5	**87.6**

We can see that the noise filter method is the most effective. However, when DWT Pooling method is combined with it, the effect is not significantly improved. This is because their purpose is basically the same, both to filter noise. However, the difference is that DWT Pooling method has no parameters and uses the algorithm to filter. Noise filter method filters feature by learning parameters. The noise of the whole layer has been reduced by combining the EPNet layer with the noise filtering layer. The effectiveness of the EPNet layer also depends on whether the feature extraction layer obtains the key features. If the key features can be obtained, it is more conducive to EPNet to increase the gap between different categories and improve the classification results.

Conclusions

The framework in this paper mainly to check the health of children's molar. The framework in this paper is mainly divided into two parts. The first part is to use the object detection model to identify molars. We propose to increase the attention mechanism in object detection to solve the problem that misidentified the extraoral objects as a molar. For the question that the multiple bounding boxes marking the same molar, we cluster the location information of the bounding box in the test phase, and then screen the bounding box according to the score of the bounding box. The second part is the classification of molars. Because the features of different categories of molars are similar, we propose to add a noise filtering layer in the feature extraction network. At the same time, we use the EPNet to encode the extracted features. In the future, we will expand the dataset to all age groups and detect the health score of each tooth.

References

1. Naaman, R., El-Housseiny, A. A. & Alamoudi, N. The use of pit and fissure sealants—a literature review. *Dent. journal* **5**, 34 (2017).
2. Datta, S., Chaki, N. & Modak, B. Neutrosophic set-based caries lesion detection method to avoid perception error. *SN Comput. Sci.* **1**, 63 (2020).
3. Geetha, V., Aprameya, K. & Hinduja, D. M. Dental caries diagnosis in digital radiographs using back-propagation neural network. *Heal. Inf. Sci. Syst.* **8**, 1–14 (2020).
4. Zerouaoui, H. & Idri, A. Reviewing machine learning and image processing based decision-making systems for breast cancer imaging. *J. Med. Syst.* **45**, 1–20 (2021).
5. Agnes, S. A., Anitha, J., Pandian, S. I. A. & Peter, J. D. Classification of mammogram images using multiscale all convolutional neural network (ma-cnn). *J. medical systems* **44**, 1–9 (2020).
6. Agarwal, M. *et al.* A novel block imaging technique using nine artificial intelligence models for covid-19 disease classification, characterization and severity measurement in lung computed tomography scans on an italian cohort. *J. Med. Syst.* **45**, 1–30 (2021).
7. Rocha, J., Cunha, A. & Mendonça, A. M. Conventional filtering versus u-net based models for pulmonary nodule segmentation in ct images. *J. medical systems* **44**, 1–8 (2020).
8. Majeeth, S. S. & Babu, C. N. K. Gaussian noise removal in an image using fast guided filter and its method noise thresholding in medical healthcare application. *J. medical systems* **43**, 1–9 (2019).
9. Li, Q., Shen, L., Guo, S. & Lai, Z. Wavelet integrated cnns for noise-robust image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7245–7254 (2020).
10. Zhao, M., Zhong, S., Fu, X., Tang, B. & Pecht, M. Deep residual shrinkage networks for fault diagnosis. *IEEE Transactions on Ind. Informatics* **16**, 4681–4690 (2019).
11. Rodríguez, P., Laradji, I., Drouin, A. & Lacoste, A. Embedding propagation: Smoother manifold for few-shot classification. *arXiv preprint arXiv:2003.04151* (2020).
12. Girshick, R., Donahue, J., Darrell, T. & Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 580–587 (2014).
13. Ren, S., He, K., Girshick, R. & Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, 91–99 (2015).
14. Jocher, G. *et al.* ultralytics/yolov5: v3.0, DOI: 10.5281/zenodo.3983579 (2020).
15. Liu, W. *et al.* Ssd: Single shot multibox detector. In *European conference on computer vision*, 21–37 (Springer, 2016).
16. Comaniciu, D. & Meer, P. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on pattern analysis machine intelligence* **24**, 603–619 (2002).
17. Zhou, D., Bousquet, O., Lal, T. N., Weston, J. & Schölkopf, B. Learning with local and global consistency. In *Advances in neural information processing systems*, 321–328 (2004).
18. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700–4708 (2017).
19. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778 (2016).
20. Howard, A. *et al.* Searching for mobilenetv3. In *Proceedings of the IEEE International Conference on Computer Vision*, 1314–1324 (2019).
21. Zagoruyko, S. & Komodakis, N. Wide residual networks. *arXiv preprint arXiv:1605.07146* (2016).