Expert Systems With Applications 194 (2022) 116519

Contents lists available at ScienceDirect

Expert Systems With Applications

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

# Structure-aware deep learning for chronic middle ear disease

Zheng Wang [b,c,1], Jian Song [d,1], Ri Su [c], Muzhou Hou [c,*], Min Qi [e], Jianglin Zhang [a,*], Xuewen Wu [d]

[a] Department of Detmatology, Shenzhen Peoples Hospital, The Second Clinical Medica College, Jinan University. The First Affiliated Hospital, Southern University of Science and Technology, Shenzhen 518020, Guangdong, China
[b] Computer Science School, Hunan First Normal University, Changsha 410205, China
[c] School of Mathematics and Statistics, Central South University, Changsha 410083, China
[d] Department of Otorhinolaryngology of Xiangya Hospital, Central South University. Key Laboratory of Otolaryngology Major Disease Research of Hunan Province. National Clinical Research Centre for Geriatric Disorders, Department of Geriatrics, Xiangya Hospital, Central South University, Changsha 410008, China
[e] Department of Plastic Surgery, Xiangya Hospital, Central South University, Changsha 410008, China

ARTICLE INFO

*Keywords:*
Convolutional Neural Networks (CNNs)
Computed Tomography (CT)
Middle Ear (ME)
Chronic Suppurative Otitis Media (CSOM)

ABSTRACT

The main purpose of this paper was to develop a deep-learning method for the diagnosis of different chronic middle ear diseases, including middle ear cholesteatoma and chronic suppurative otitis media, based on computed tomography (CT) images of the middle ear. The origin of the dataset was the CT scans of 499 patients, which included both ears and selected by specialized otologists. The final dataset was constructed from 973 ears, which labeled by a professional otolaryngologist and classified into 3 conditions: MEC, CSOM and normal. The diagnostic framework, called the "Middle Ear Structure Identification Classifier"(MESIC), was consisted of two deep-learning networks with dissimilar functions: a "region of interest" area search network for extracting the special image of the middle ear structure and a classification network for finishing the diagnosis. The area under the curve (AUC), which means receiver operating characteristic curve (ROC), reflects the robustness of the algorithm by comparing its sorting effectiveness. According to simulation experiments, we chose Visual Geometry Group 16 (VGG-16) as the model's backbone. In our framework, the ROI search part exhibited an AUC of 0.99 on the right and 0.98 on the left. The classification part exhibited an average AUC of 0.96 for both sides based on VGG-16. The average precision (90.1%), recall (85.4%) and F1-score (87.2%) show the effectiveness of framework. This paper presents a deep-learning framework to automatically diagnose cholesteatoma and CSOM. The results show that MESIC can effectively and quickly classify these two common diseases through CT images, which can ameliorate the pressure of professional doctors and the practical problems of the lack of professional doctors in rural areas.

## 1. Introduction

Chronic middle ear diseases, which describe some common problems with the middle ear, play important roles in daily otorhinolaryngology practice due to their high incidence. This kind of disease represent a major cause of hearing loss, especially in developing countries (Hallberg et al., 2008; Bächinger et al., 2021). Surgery should be performed in some cases for the purpose of removing the lesions and infection to achieve an infection-free and dry ear (Shohet et al., 2002). Chronic suppurative otitis media (CSOM) is characterized by persistent inflammation of the ME or mastoid cavity (Acuin et al., 2004). Middle ear cholesteatoma (MEC) refers to uncontrolled growth of squamous keratinized epithelium in the ME, usually located in the tympanic cavity and/or tympanic sinus, mastoid cavity, or connective tissue below the epithelium (Nevoux et al., 2010; Rutkowska et al., 2017). The European Academy of Otology and Neurotology, in collaboration with the Japanese Otological Society (EAONO/JOS), produced a joint consensus document outlining the definition, classification and staging of

*Abbreviations:* CNNs, Convolutional Neural Networks; CT, Computed Tomography; MEC, Middle Ear Cholesteatoma; CSOM, Chronic Suppurative Otitis Media; ROC, Receiver Operating characteristic Curve; AUC, Area Under the Curve.

* Corresponding authors.
*E-mail addresses:* entsj@csu.edu.cn (J. Song), suricsu@csu.edu.cn (R. Su), houmuzhou@sina.com (M. Hou), qimin05@csu.edu.cn (M. Qi), zhang.jianglin@szhospital.com (J. Zhang).

[1] Zheng Wang and Jian Song contributed equally to this work.

https://doi.org/10.1016/j.eswa.2022.116519
Received 6 November 2021; Received in revised form 14 December 2021; Accepted 7 January 2022
Available online 15 January 2022

cholesteatoma (Castle et al., 2018). The differences in etiologies and lesion manifestations and treatments between these two diseases make diagnosis very important (Lustig et al., 2018).

Previous studies have demonstrated that high-resolution computed tomography (CT) of the temporal bone is currently an accurate diagnostic imaging method for cases of Chronic middle ear diseases because it clearly shows the middle ear structure and exhibits sensitivity in detecting the characteristic findings of middle ear lesions, including the extent and complications (Kusak et al., 2018). Generally, the hallmarks of CT imaging of CSOM and MEC are soft tissue mass-like opacities in the middle ear cavity and mastoid antrum. Association with smooth bony erosion of the ossicles and expansion of adjacent structures are the lesion characteristics of MEC (Molteni et al., 2019; Gaurano et al., 2004).

Artificial intelligence methods have made many contributions to the diagnosis of other diseases and the use of medical images, especially in the intelligent analysis of pulmonary medical images. An outbreak within the Champions League in 2020 resulted in numerous researchers using the method of artificial intelligence in different medical imaging applications as a method of rapid diagnosis (Tsiknakis et al., 2020; Apostolopoulos et al., 2020; Wang et al., 2020; Mei et al., 2020; Kanavati et al., 2020; Tang et al., 2020). Automatic analysis of medical images obtained from other parts of the body has also proven feasible (Younis et al., 2019; Wang et al., 2019; Fukae et al., 2020; Wang et al., 2021).

With the rapid development of deep learning, it is feasible to apply computer vision processing and recognition technology to medical image-related fields. This approach can greatly reduce the human cost and reduce the human error caused by repetition, fatigue or differences in knowledge to improve the general diagnosis rate. The purpose of this paper was to construct an automatic diagnosis framework of chronic ME disease by means of a convolutional neural network to provide doctors with an unbiased diagnosis reference before diagnosis.

This paper's primary contributions are as follows:

- We provide a new automatic detection direction for the detection of cholesteatoma and CSOM in Ear, Nose and Throat departments. Compared with ordinary endoscopic images, CT images can better reflect the details of the inner structure of the ME, and the classification effect is more obvious for these two diseases.
- To solve the problem of data imbalance caused by a small number of cholesteatoma cases in trainning, we decided to use image inversion to enhance the training samples of the classification network by observing the data characteristics.
- The process of our design is fully automatic. We only need to input the brain CT scan results of the corresponding patients to determine the patient's binaural disease results.
- We use a convolutional neural network to classify the CT images that have extracted the region of interest (ROI) efficiently.

This work could reduce the burden on doctors, and provide a feasible plan for intelligent diagnosis in the future.

## 2. Literature reviews

In today's hospitals, otolaryngologists usually obtain a comprehensive understanding of the structure of the ME through endoscopy and CT images and make a treatment plan based on the diagnosis. Some researchers have demonstrated the effectiveness of CT and diffusion-weighted magnetic resonance imaging in the detection of cholesteatoma and have shown that fusion CT and diffusion-weighted magnetic resonance imaging (CT-DWMRI) is superior to either diagnostic method alone in elucidating the location and bone relationship in different cases of MEC, making it a valuable surgical planning tool (Dutt et al., 2019). Based on CT images, CNNs have also been proven effective in judging otitis media and cholesteatoma (Wang & Li et al., 2019).

There have been many research achievements in the application of CNNs in otolaryngological medical image processing. Classification of tympanum perforations of different sizes using migration learning and Inception-V3 has been successfully implemented for this task (Habib et al., 2020). Using a large database, (Cha et al., 2019; Khan et al., 2020) the tympanic membrane and external auditory canal characteristics were divided into 6 categories covering most ear diseases with good, accurate values. Some researchers have realized the automatic segmentation of the cochlea and vestibule, calculated the ratio of hydrolytic endolymph, and automatically realized the diagnosis of Meniere's disease (Cho et al., 2020; Viscaíno et al., 2020).

Otosclerosis is a multifactorial bone disorder that affects the otic capsule with complex etiology.Pathologically, it is due to the primary localized bone resorption of the bone labyrinth, which is replaced by spongy bone hyperplasia with abundant blood vessels (Quesnel et al., 2018). Fujima et al. firstly used a variety of deep learning methods to analyze the temporal bone CT of patients with otosclerosis. Compared with radiologists, the analysis using GoogLeNet and ResNet proved the non-inferiority of deep learning in the diagnosis of otosclerosis (Fujima et al., 2019). Tan et al. used Logical Neural Network (LNN) to analyze otosclerosis images, which effectively reduced the misdiagnosis of fenestral otosclerosis (Tan et al., 2021).

Besides, the segmentation and positioning of related anatomical structures through deep learning is helpful for clinicians to further understand and learn the adjacent relationship between anatomical structures, select treatment methods and plan surgical routes (Yao et al., 2021). At present, deep learning can also be used for fine segmentation and localization of temporal bone structure through imaging, as well as feature extraction and differentiation of different lesions (Nikan et al., 2021; Li et al., 2020; Vaidyanathan et al., 2021).

## 3. Materials and methods

In this section, we describe the data sources, data processing, and the overall structure and details of the model.

### 3.1. Dataset and data preprocessing

This study was approved by the medical research and ethics committee of Xiangya Hospital, Central South University. The data were collected from 573 patients who underwent middle ear surgery at the Department of Otorhinolaryngology, Xiangya Hospital, from January 2018 to October 2020. Medical records were then reviewed to exclude any patient diagnosed with congenital malformation or any postoperative situation, as well as any patient missing a temporal bone CT scan.

According to the clinical diagnostic criteria for otitis media (Listed 1994) and a review of medical history records, preoperative examination findings (pure tone audiometry, temporal bone CT, ear endoscopy, etc.), intraoperative findings and postoperative pathological findings, three experienced otolaryngologists divided these operated ears into two groups: the CSOM group (only for resting phase) and the MEC group. In a few cases where MEC was present in combination with CSOM, the diagnosis of MEC was prioritized. For normal ears, the label was assigned according to the absence of ear discharge, hearing loss, or signs of inflammation on the imaging examination.

The composition of the dataset is shown in Table 1. In our dataset, the age range of patients was 5–72 years, with a mean ± SD of 38.75 ± 14.38 years. There were 198 (40%) men and 301 (60%) women. All

**Table 1**
Illustration of dataset.

| | Right | Left | Total |
|---|---|---|---|
| MEC | 62 | 46 | 108 |
| CSOM | 308 | 314 | 622 |
| Normal | 129 | 139 | 268 |
| Total | 499 | 499 | 998 |

patients were admitted to the hospital and underwent surgery during the hospitalization period, and had no other diseases such as granulation seed, secretory otitis media, or inner ear malformation.

To achieve a better training effect, we carried out systematic manual annotation of the original data under the guidance of cooperative professionals. We chose to make a cover for the ME structure and crop the ME structure with a $100 \times 100$ pixel box, in accordance with region of interest (ROI) labeling. This approach helped us train the first network. Fig. 1 presents an example regarding data preprocessing.

In the dataset, due to the small number of MEC instances, a data imbalance phenomenon was produced. To reduce the impact of data imbalance, we reused the MEC data. In the ME data images, the left and right sides are mainly distinguished by the outer contour, so we added inverted left ear MEC case data when training the right ear pathological classifier. We also used the same operation when training the right ear classifier after transfer learning. Fig. 2 shows the similarity of the left and right ME structures.

### 3.2. Overview of the framework

To realize the automatic output of pathological results from CT images, we designed two different networks to form a series structure to solve problems: region of interest (ROI) search net and classification net (C-Net). Fig. 3 presents the complete technological process and some details about ROIs and the middle ear structure identification classifier (MESIC).

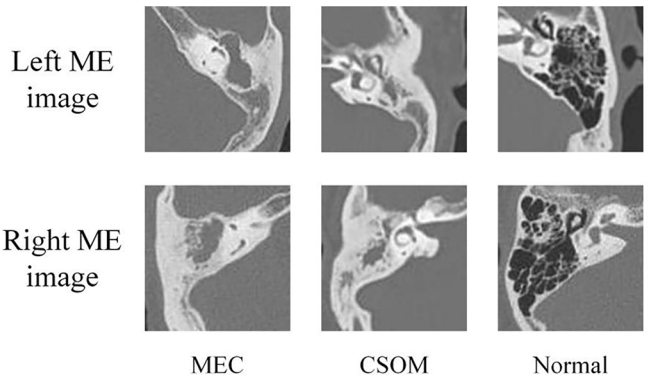



**Fig. 2.** Flip examples. Three 3 kinds of ME images from different sides. All examples show their differences.

### 3.3. Region of interest (ROI) search net

With the help of the mask region-based convolutional neural network (Mask R-CNN) (He et al., 2018), we realized automatic searching of the bounding box, which is the box where we search and cut the middle ear structure diagram. As Fig. 3 shows, the ROI part gives a clear structure of our method. The gray block is the region proposal network (RPN) of the Mask R-CNN, which can help us to roughly determine the middle ear structure. ROI-Align uses bilinear interpolation to overcome the misalignment problem that exists in the Faster R-CNN. In this way, we can obtain a fixed size (which depends on the training tags that we are preparing) feature map to produce the correct

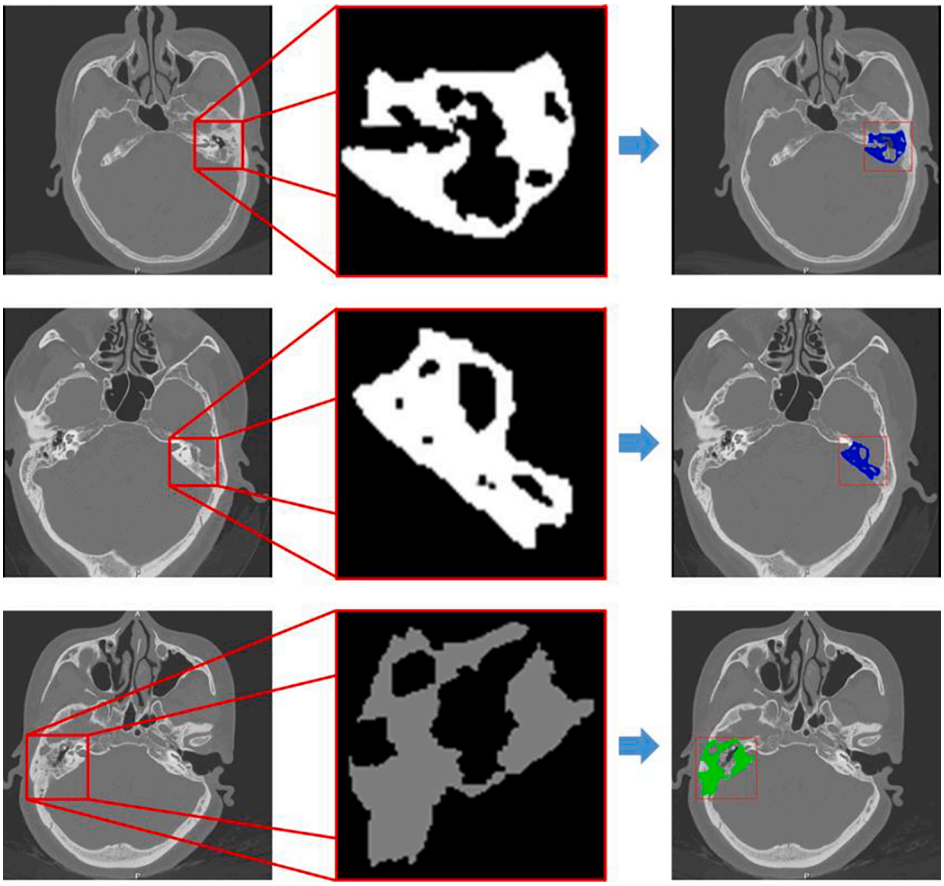


**Fig. 1.** Image preprocessing. Several preprocessing examples. Through expert tagging, we created training tags, indicated by the red box, for the ROI network. The white/gray pixel area is the true structure of the ME, which contains the characteristics of diseases. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

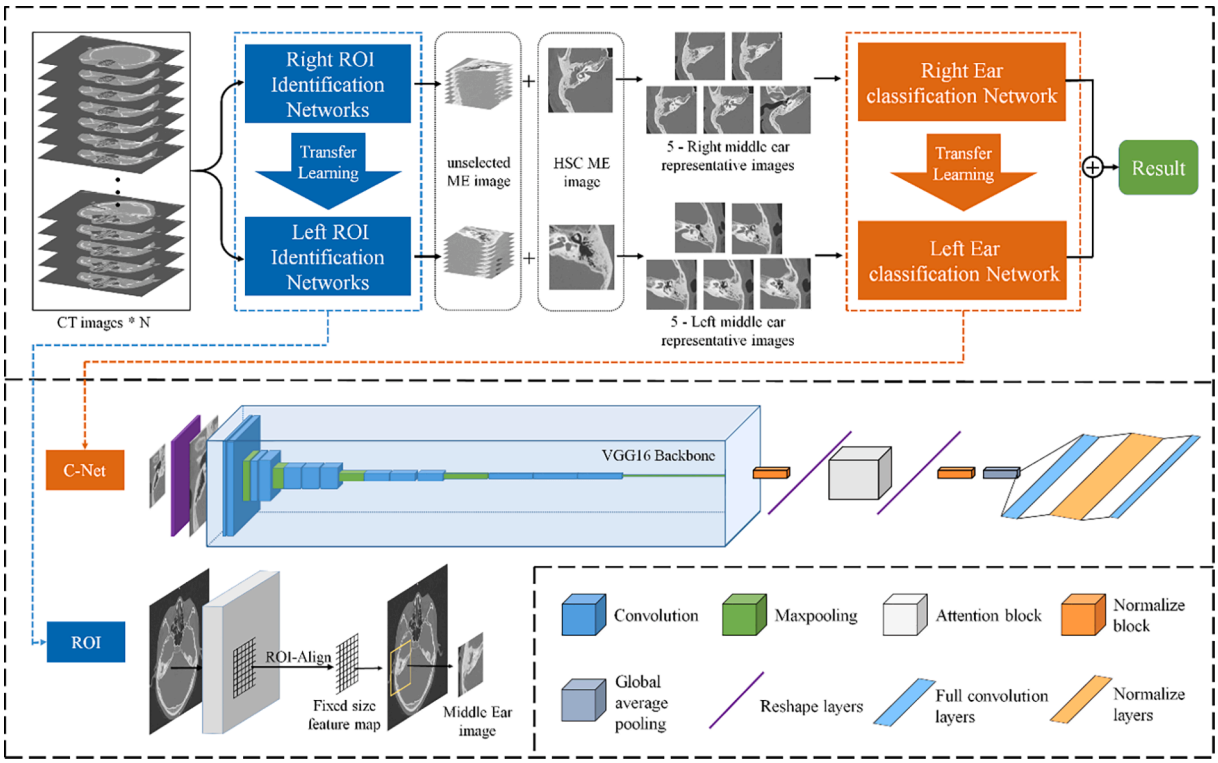



**Fig. 3.** An overview of the MESIC. The upper part of the figure shows the overall judgment process of the MESIC. After all CT layers are input, they will go through two networks and an automatic processing process, and finally the binaural results of the patient will be output. The lower part of the figure shows the detailed network structure corresponding to the upper part, which is the image extraction network constructed by the ROI method in the mask region-based convolutional neural network (MASK R-CNN) and the classification network with Visual Geometry Group 16 (VGG-16) as the backbone.

bounding box.

As Fig. 4 shows, we chose 5 special ME images to represent all structures of the ME. The chosed 5 ME images generates representative feature maps containing the region of the ME (as shown in Fig. 4). The selection of 1st image and 5th image consists of superior semicircular canal and the floor of external acoustic meatus respectively, and illustrated the upper and lower boundary markers of middle ear stucture. The 3rd image contains horizontal semicircular canal and internal auditory canal, which are two landmark structure. The 2nd image and 4th image locates the upper tympanum and the mastoid respectively, both of which are main parts of ME. In temporal bone CT scans, these structures demonstrate significant features that can be effectively learned by a deep learning model.

Because of the importance of the horizontal semicircular canal (HSC), the way that we extracted the special image was to use the HSC middle ear image as an anchor point, and we defined the ROI network to

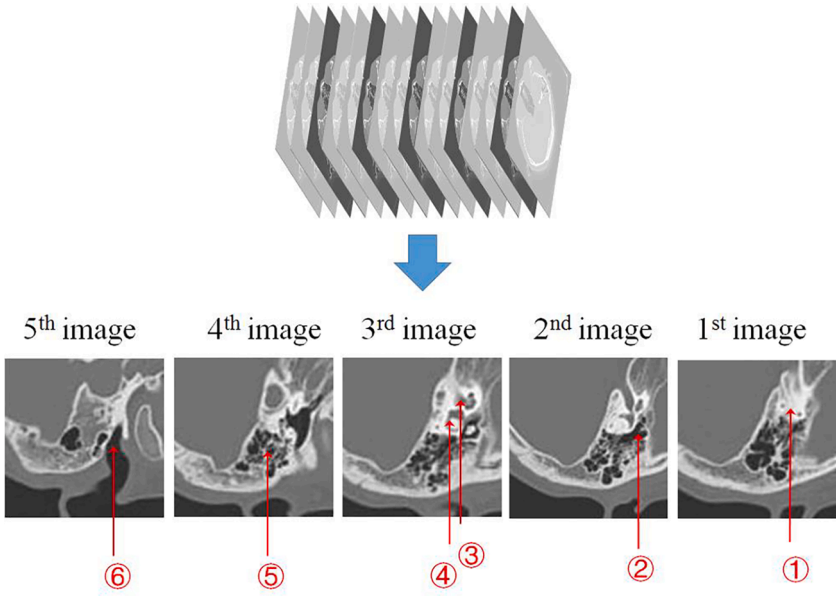



**Fig. 4.** The example of five Representative CT images (Left ear). ①Superior semicircular canal;②Upper tympanum;③Internal auditory canal;④Horizontal semicircular canal; ⑤ Mastoid;⑥ External acoustic meatus.

distinguish k pictures (k < N). We used k to obtain the step size, which was used with the anchor point to locate the 5 special images. The step size was defined by equation (1):

$$stepsize = \left[\frac{k}{5}\right] \tag{1}$$

where [ ] is an integer operation.

### 3.4. Classification net (C-Net)

For the extracted special ME images, we used the designed CNN backbone and restructured the final full convolution layer to complete our classification task. In the experimental phase, we used different network backbones to find the one that best suited our model. After experiments, we ultimately selected Visual Geometry Group 16 (VGG-16) (Simonyan et al., 2014) as the main structure of the classification network.

As shown in Fig. 3, the MESIC provides the network structure diagram. First, to properly apply the VGG-16 backbone, we adjusted the size of the input image from 100X100 to 256 × 256. Then, we obtained an 8 × 8 × 512 feature block through a series of convolution and pooling operations of some columns on the backbone network.

Although there was still some noise in the final selected class image, we applied the attention mechanism (Fu et al., 2019) in the final full convolution stage. As demonstrated in Fig. 5, given a input $X \in \mathbb{R}^{H\times W\times C}$, we generate three new feature maps A, B and C, respectively. Then we reshape them to $\mathbb{R}^{N\times C}$, where $N = H \times W$ is number of features. The resulting output $S \in \mathbb{R}^{H\times W\times C}$ calculates as follows:

$$S_j = \alpha \sum_{i=1}^{N} \left( \frac{exp(B_i \cdot C_j)}{\sum_{i=1}^{N} exp(B_i \cdot C_j)} A_i \right) + X_j \tag{2}$$

Where · perform a matrix multiplication between the transpose of B and C, $\alpha$ as a scale parameter is initialized as 0 and gradually learn weight. The obtained feature S at each position is a weighted sum of the features at all positions and original features. By considering P as the attention operator and R as the reshape operator, the above equation changes to the form:

$$S = P(X)R(X) + X \tag{3}$$

We first reconstructed the 8 × 8 feature map to a 64-dimensional vector and then applied the attention block to the reconstructed vector. We then restored the attention-processed vectors through the reverse process of remolding to 8 × 8 and sent them to the full convolution layer to extract depth features and complete the classification task.

In the final full convolution layer, the corresponding disease probability of a given image was output according to the following calculation formula:

$$p^c = \sum_k w_k^c \sum_{i,j} y^k(i,j) \tag{4}$$

Where $y^k(i,j)$ represent the activation of filter unit k in the last convolutional layer at position (i, j). $\sum_{i,j} y^k(i,j)$ is the result of global average pooling. $w_k^c$ is the weight and correspondes the class c for filter unit k. We set the given image bias to 0 as it has little to no influence on the classification task. Finally $p^c$ of the softmax is given by $\frac{\exp(p^c)}{\sum_c \exp(p^c)}$.

For the probability of the final output of the five feature maps, we adopted the medical concept of not missing the diagnosis even if the diagnosis was wrong. In the five special ME images, we used the concept of veto power; that is, if one was judged to be MEC or CSOM, the framework would definite diagnosis of this patient according to the corresponding disease even if the other four ME images were diagnosed as normal.

## 4. Experiment and results

In this section, we will describe all aspects of the experiment, including the Evaluation Metrics, training process, definition of the loss function and final experimental results.

### 4.1. Evaluation Metrics and network training process

Similar to many rapid diagnostic frameworks based on deep learning (Dutt et al., 2019), we used three metrics, precision, recall, and F1-score, to measure our final diagnostic effectiveness.

In the process of constructing the network, the most important consideration was the training of the two main networks. We used transfer learning to simplify the training of similarity classification networks on both sides. By fine-tuning the network, the two similar networks shared the same architecture but with different parameters. This approach could reduce the quantity of data needed to achieve a certain precision in the training process and effectively improve the accuracy of training. As mentioned above, we chose to train the right ear-related network first and then use the learned structure to carry out transfer learning to finish training the left ear-related network.

During model training, the dataset was randomly divided into a training set, verification set and test set at proportions of 70%, 20% and 10%, respectively. In the random partition process, we retained the corresponding proportions of the three different diseases and rounded them down to ensure that the framework did not conduct "bias" learning after the partition. Fig. 4 shows our use of separate datasets and the whole process of the experiment.

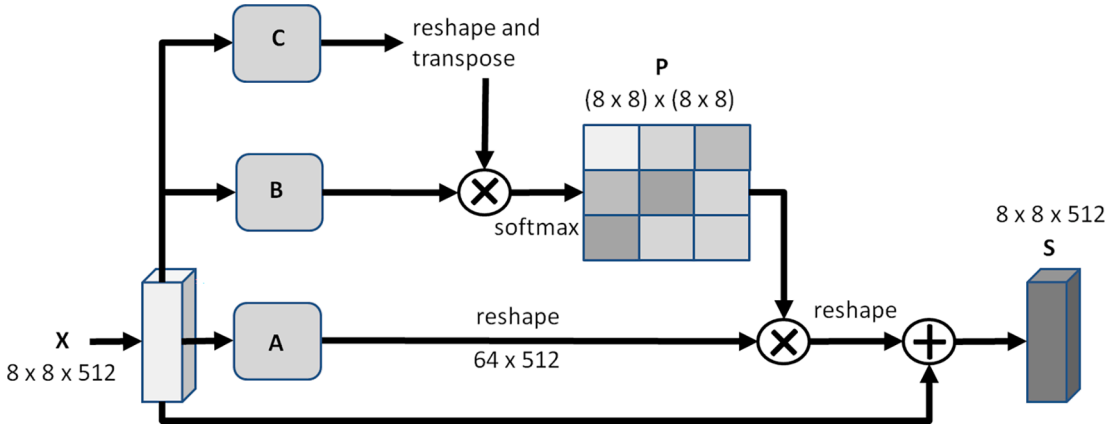



**Fig. 5.** The details of spatial attention mechanism.

### 4.2. Loss function

According to the requirements of different stages in the training process, we used different loss functions in different stages after adjustment. In the searching ROI phase, because our goal related to the image segmentation task and the dataset was imbalanced, we chose the "dice coefficient" (dice-coef) (Priyadarshini et al., 2018) as the loss function.

In the classification network, we used adaptive moment estimation (Adam) (Kingma et al., 2014) as the optimizer of the training classifier. The main idea of the algorithm is to calculate the update step size of the parameters by considering the first and second moments of the gradient. In the intermediate process of convolution, we used the rectified linear unit (ReLU) (Wang et al., 2016) activation function to add the nonlinear fitting function. In the final fine-tuning full convolution layer, we chose the common SoftMax (Grave et al., 2016) to strengthen our final selection probability. Its formula is as follows:

$$Loss = -\sum_{i=1}^{outputsize} y_i \cdot log\hat{y}_i \tag{5}$$

Where $\hat{y}_i$ is the output value, and $y_i$ is the ground truth value.

### 4.3. MESIC results

We used the part of Mask R-CNN method to build our network and search for an ROI presenting highly accurate results. Fig. 6 shows the corresponding ROC curves. This figure shows almost no difference between the ME structure automatically extracted from the network and the manually marked image, which proves the feasibility of our construction method and lays a good foundation for the classification of the following network.

Based on previous work, the MESIC exhibits a good classification effect. Table 2 reports the related results. The paper by Wang (Wang & Li et al., 2019) concluded that Inception-V3 (Szegedy et al., 2016) can effectively solve ME problems based on CT images. Therefore, we performed the network model from their study and VGG-16 for 5-fold cross validation experiments. From the results, CNNs with the VGG-16 backbone showed better results in the both sides classification task. Therefore, we chose VGG-16 as the preferred backbone of our structure based on the results because it presented a balanced and successful result on both sides of the classification.

Fig. 7 presents the confusion matrix of the classification results. The classification accuracy was higher for MEC and CSOM cases than normal cases, which is mainly due to the following three reasons. First, due to the uneven distribution of the datasets, some details of normal and mild CSOM cases also could not be learned. Second, the MEC image feature is obvious and could be effectively learned by computers. Third, based on the bottom line of no missed diagnosis in medicine, we set up a biased mechanism in the classification process, which led to low precision for normal cases.

**Table 2**
Backbone comparison on the test set.

| Sides | Backbone | Mean Metrics | | | | | |
|---|---|---|---|---|---|---|---|
| | | Pre (%) | Rec (%) | F1 (%) | Sen (%) | Spe (%) | p-value |
| Left | Wang & Li et al., 2019 | 83.8 ± 1.3 | 84.7 ± 1.4 | 82.6 ± 1.2 | 88.6 ± 0.1 | 91.2 ± 3.1 | 3.1e-26 |
| Right | Wang & Li et al., 2019 | 89.6 ± 0.9 | 82.2 ± 1.2 | 86.0 ± 0.9 | 87.1 ± 0.2 | 90.2 ± 3.6 | 8.4e-25 |
| Left | Ours | 90.5 ± 0.3 | 87.9 ± 0.3 | 89.7 ± 0.3 | 90.7 ± 0.1 | 94.5 ± 1.5 | 7.6e-29 |
| Right | Ours | 93.8 ± 0.3 | 95.1 ± 0.3 | 94.4 ± 0.3 | 96.7 ± 0.1 | 97.6 ± 0.8 | 1.7e-46 |

*Pre: precision; Rec: recall; F1: F1-score; Sen: Sensitivity; Spe: Specificity; p-value of Pearson correlation coefficient.*

In the training stage, we used transfer learning from the right ear net to the left ear net in two main parts, enabling the left ear correlation network to achieve better classification robustness in the training process with reduced data requirements. In other words, the network for the left ear is more robust than that for the right ear. We can see from the confusion matrix that the classification effect of the left ear network was more balanced and more accurate than that of the right ear.

Fig. 8 shows the receiver operating characteristic (ROC) curve results for all classification tasks. The area under the ROC curve (AUC) measures the robustness of the algorithm for a certain classification problem. From the results, we can see that the MEC task classification exhibited the best robustness regarding the MESIC, and it also exhibited good robustness in the other two classification problems. The ROC curve results prove that our framework has good generalizability for this problem. (See Fig. 9).

## 5. Discussion

This paper presents a framework to detect MEC and CSOM from CT images with improved accuracy. In the framework, we combined the part of MASK R-CNN to realize an intelligent search for the middle ear structure, increasing the accuracy of the algorithm for pathological recognition. By contrast experiments, we reconstructed a CNN classfier with VGG-16 as the backbone and applied the attention mechanism to make our method robust. Through experiments, we determined that the

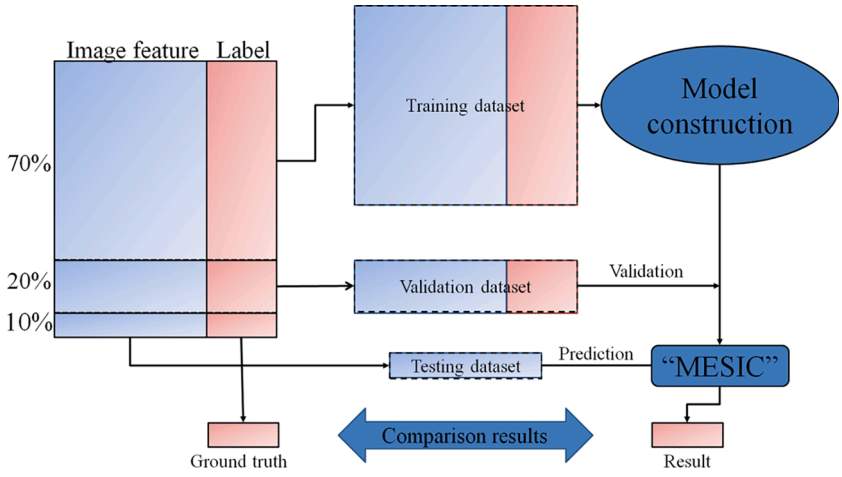



**Fig. 6.** Experimental process. We divided the dataset into three parts for different steps in the experiment.

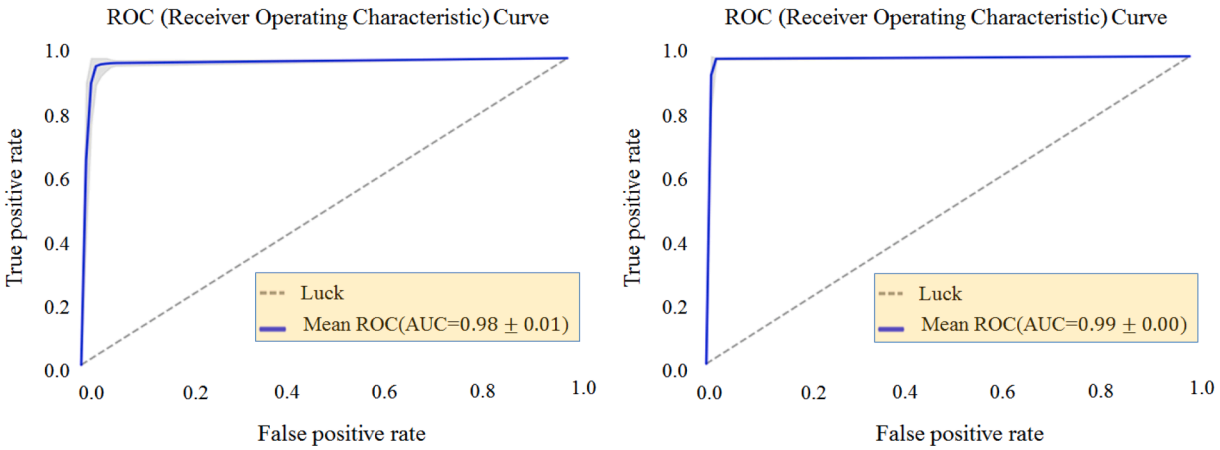
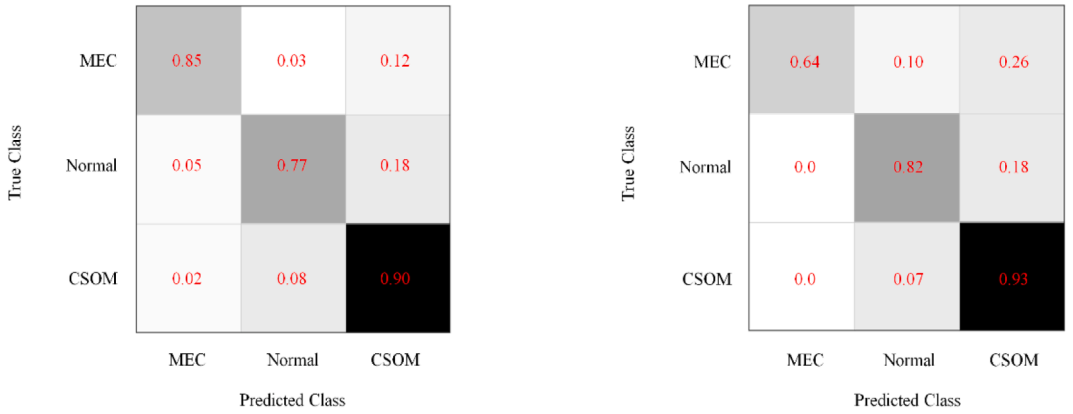
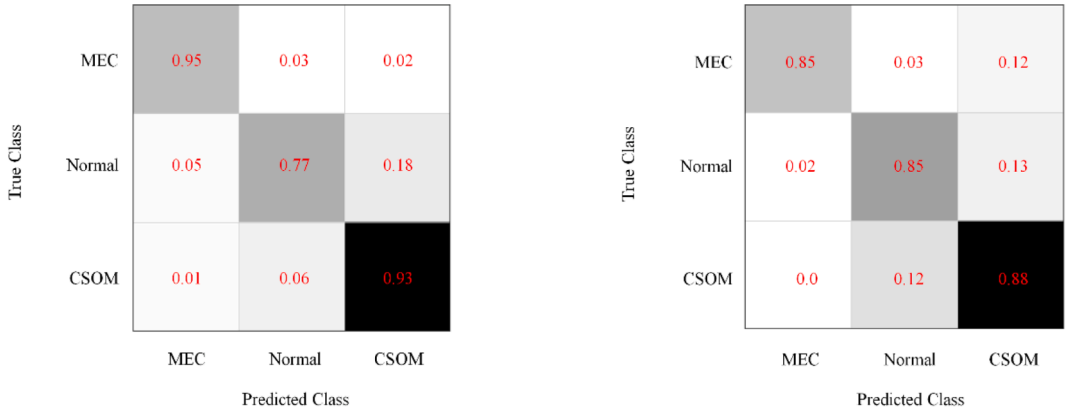



**Fig. 7.** ROC curve of the network results in the first part of the framework. The results demonstrate that the extraction of the ROI area has high accuracy.




(a) Right result based on Inception-V3

(b) Left result based on Inception-V3




(c) Right result based on VGG-16

(d) Left result based on VGG-16

**Fig. 8.** Confusion matrixes of different classification results. The higher the percentage of test samples for classification is, the darker the color of the corresponding blocks. The diagonal line of blocks indicates cases of correct classification.

average accuracy of MESIC was 90.2%, the average recall rate was 85.4%, and the average F1-score was 87.3%. In terms of performance, the average AUC for the classification of MEC, CSOM, and normal cases was 0.985, 0.950, and 0.935, respectively. The results show that our framework can effectively and quickly classify CT numbers (CTNs) through CT images, which can overcome the fatigue of professional doctors and the practical problems of a shortage of professional doctors in rural areas.

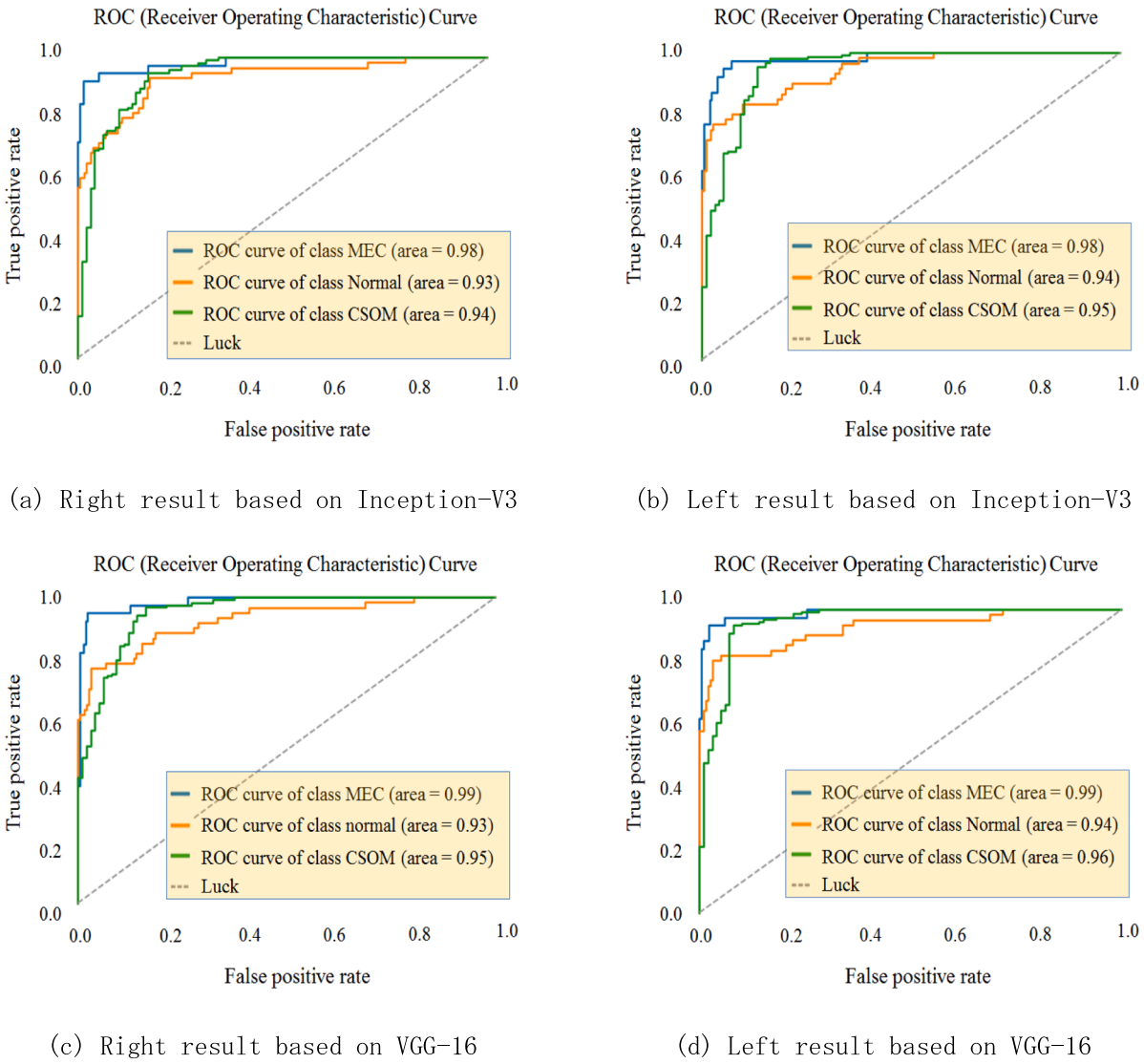



**Fig. 9.** ROC curves of different classification results. The area under the ROC curve (AUC) shows the robustness of the algorithm.

The framework is helpful in otolaryngology for daily diagnosis and relief of sitting pressure, but it has some limitations. First of all, this framework is based on the representative CT scan level for diagnosis, which means that when the patient's disease is in the early stage, the lesion images of CSOM and MEC may not exist at the representative level, which may lead to missed diagnosis. Secondly, because CT images scan the patient from top to bottom, there is natural continuity, and the information between these images is not taken into account in this framework. Finally, during routine medical visits, the image data that patients receive is usually physically printed rather than electronic (which can only be transmitted through personal accounts on the healthcare platform). Therefore, the framework is more suitable for comprehensive inspection using local detection instruments and displaying diagnostic reference results in printed data.

In future research, on the premise of improving the accuracy of the algorithm, we envisage using the upper and lower relationship between each layer of CT images to allow more ME images to enter the classifier and return the number of image layers corresponding to the disease. This hypothesized approach could help doctors better find a patient's lesions to provide a good recommendation for preoperative planning. At the same time, with the popularization of smart phones and the development of the network, the online consultation platform can be linked with the cloud digital hospital deployed in the framework to realize the effect of independent intelligent diagnosis for individual patients. This prospect also puts more stringent requirements on the speed of the framework diagnosis and the lightness of the network volume, which is the direction of improvement in the future work of the framework.

*CRediT authorship contribution statement*

**Zheng Wang:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Revising the paper. **Jian Song:** Investigation, Resources, Data curation. **Ri Su:** Software, Data curation. **Muzhou Hou:** Resources, Writing – review & editing. **Min Qi:** . **Jianglin Zhang:** Methodology, Investigation, Resources, Data curation, Writing – review & editing. **Xuewen Wu:** .

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

This work was supported by Scientific Research Fund of Hunan Provincial Education Department (grant number 20C0402), Hunan First Normal University (grant number XYS16N03), the Projects of the National Natural Science Foundation of China (grant number 82073019 and 82073018), the China Postdoctoral Science Foundation (grant number 2021M693566, 2021T140751), The science and technology innovation Program of Hunan Province China (grant number 2020RC2013), Hunan Province Natural Science Foundation (grant number 2021JJ41017,2021JJ31108).

## References

Hallberg, L., Hallberg, U., & Kramer, S. (2008). Self-reported hearing difficulties, communication strategies and psychological general well-being (quality of life) in patients with acquired hearing impairment. *Disability and rehabilitation., 30*, 203–212. https://doi.org/10.1080/09638280701228073

Bächinger, D., Großmann, W., Mlynski, R., & Weiss, N. (2021). Characteristics of health-related quality of life in different types of chronic middle ear disease. *European Archives of Oto-Rhino-Laryngology., 278*, 1–6. https://doi.org/10.1007/s00405-020-06487-6

Shohet, J., & Jong, A. (2002). The management of pediatric cholesteatoma. *Otolaryngologic clinics of North America., 35*, 841–851. https://doi.org/10.1016/S0030-6665(02)00052-X

Acuin, & Jose. (2004). Chronic suppurative otitis media: burden of illness and management options. Geneve World Health Organization.

Nevoux, J., Lenoir, M., Roger, G., Denoyelle, F., Pointe, H., & Garabédian, E.-N. (2010). Childhood cholesteatoma. *European annals of otorhinolaryngology, head and neck diseases., 127*, 143–150. https://doi.org/10.1016/j.anorl.2010.07.001

Rutkowska, J., Ozgirgin, N., & Olszewska, E. (2017). Cholesteatoma Definition and Classification: A Literature Review. *The Journal of International Advanced Otology., 13*. https://doi.org/10.5152/iao.2017.3411

Castle, J. (2018). Cholesteatoma Pearls: Practical Points and Update. *Head and Neck Pathology., 12*. https://doi.org/10.1007/s12105-018-0915-5

Lustig L R & Limb C J & Baden R. (2018)..Chronic otitis media, cholesteatoma, and mastoiditis in adults. UpToDate Waltham, MA (citirano 145 2019).

Kusak, A., Rosiak, O., Durko, M., Grzelak, P., & Pietruszewska, W. (2018). Diagnostic imaging in chronic otitis media: Does ct and mri fusion aid therapeutic decision making? - a pilot study. *Otolaryngologia polska. The Polish otolaryngology, 72*(5), 1–5.

Molteni, G., Fabbris, C., Molinari, G., Alicandri-Ciufelli, M., Presutti, L., Paltrinieri, D., & Marchioni, D. (2019). Correlation between pre-operative CT findings and intra-operative features in pediatric cholesteatoma: A retrospective study on 26 patients. *European Archives of Oto-Rhino-Laryngology., 276*, 1–8. https://doi.org/10.1007/s00405-019-05500-x

Gaurano, J., & Joharjy, I. (2004). Middle Ear Cholesteatoma: Characteristic CT Findings in 64 Patients. *Annals of Saudi medicine., 24*, 442–447. https://doi.org/10.5144/0256-4947.2004.442

Tsiknakis, N., Trivizakis, E., Vassalou, E., Papadakis, G., Spandidos, D., Tsatsakis, A., … Marias, K. (2020). Interpretable artificial intelligence framework for COVID–19 screening on chest X–rays. *Experimental and therapeutic medicine., 20*. https://doi.org/10.3892/etm.2020.8797

Apostolopoulos, Ioannis & Bessiana, Tzani. (2020). Covid-19: Automatic detection from X-Ray images utilizing Transfer Learning with Convolutional Neural Networks.

Wang, Z., Xiao, Y., Li, Y., Jie, Z., Lu, F., Hou, M., & Liu, X. (2020). Automatically discriminating and localizing COVID-19 from community-acquired pneumonia on chest X-rays. *Pattern Recognition., 110*, Article 107613. https://doi.org/10.1016/j.patcog.2020.107613

Mei, X., Lee, H.-C., Diao, K.-Y., Huang, M., Lin, B., Liu, C., … Yang, Y. (2020). Artificial intelligence–enabled rapid diagnosis of patients with COVID-19. *Nature Medicine., 26*, 1–5. https://doi.org/10.1038/s41591-020-0931-3

Kanavati, F., Toyokawa, G., Momosaki, S., Rambeau, M., Kozuma, Y., Shoji, F., … Tsuneki, M. (2020). Weakly-supervised learning for lung carcinoma classification using deep learning. *Scientific Reports., 10*. https://doi.org/10.1038/s41598-020-66333-x

Tang, Yu-Xing & Tang, You-Bao & Peng, Yifan & Yan, Ke & Bagheri, Mohammadhadi & Redd, Bernadette & Brandon, Catherine & lu, Zhiyong & Han, Mei & Xiao, Jing & Summers, Ronald. (2020). Automated abnormality classification of chest radiographs using deep convolutional neural networks. npj Digital Medicine. 3. 10.1038/s41746-020-0273-z.

Younis, Haseeb & Bhatti, Muhammad & Azeem, Muhammad. (2019). Classification of Skin Cancer Dermoscopy Images using Transfer Learning. 1-4. 10.1109/ICET48972.2019.8994508.

Wang, Z., Meng, Y.u., Weng, F., Yinghao, C., Lu, F., Liu, X., … Jie, Z. (2019). An Effective CNN Method for Fully Automated Segmenting Subcutaneous and Visceral Adipose Tissue on CT Scans. *Annals of Biomedical Engineering., 48*. https://doi.org/10.1007/s10439-019-02349-3

Fukae, J., Isobe, M., Hattori, T., Fujieda, Y., Kono, M., Abe, N., … Koike, T. (2020). Convolutional neural network for classification of two-dimensional array images generated from clinical information may support diagnosis of rheumatoid arthritis. *Scientific Reports., 10*. https://doi.org/10.1038/s41598-020-62634-3

Wang, Z., Xiao, Y., Weng, F., Li, X., Zhu, D., Lu, F., … Meng, Y.u. (2021). R-JaunLab: Automatic Multi-Class Recognition of Jaundice on Photos of Subjects with Region Annotation Networks. *Journal of Digital Imaging., 34*. https://doi.org/10.1007/s10278-021-00432-7

Dutt, S., Bartley, A., Bassett, P., Singh, A., Lingam, R., & Hall, A. (2019). Surgical mapping of middle ear cholesteatoma with fusion of computed tomography and diffusion-weighted magnetic resonance images: Diagnostic performance and interobserver agreement. *International journal of pediatric otorhinolaryngology*. https://doi.org/10.1016/j.ijporl.2019.109788

Wang, Y.-M., Li, Y., Cheng, Y.-S., He, Z.-Y., Yang, J.-M., Xu, J.-H., … Ren, D. (2019). Deep Learning in Automated Region Proposal and Diagnosis of Chronic Otitis Media Based on Computed Tomography. *Ear and Hearing., 41*, 1. https://doi.org/10.1097/AUD.0000000000000794

Habib, A.-R., Wong, E., Sacks, R., & Singh, N. (2020). Artificial intelligence to detect tympanic membrane perforations. *The Journal of Laryngology & Otology., 134*. https://doi.org/10.1017/S0022215120000717

Cha, D., Pae, C., Seong, S.-B., Choi, J., & Park, H.-J. (2019). Automated diagnosis of ear disease using ensemble deep learning with a big otoendoscopy image database. *EBioMedicine., 45*. https://doi.org/10.1016/j.ebiom.2019.06.050

Khan, M. A., Kwon, S., Choo, J., Hong, S. M., Kang, S.-H., Park, I.-H., … Hong, S. (2020). Automatic detection of tympanic membrane and middle ear infection from oto-endoscopic images via convolutional neural networks. *Neural Networks., 126*. https://doi.org/10.1016/j.neunet.2020.03.023

Cho, Y., Cho, K., Park, C., Chung, M., Kim, J., Kim, K., … Chung, W.-H. (2020). Automated measurement of hydrops ratio from MRI in patients with Ménière's disease using CNN-based segmentation. *Scientific Reports., 10*. https://doi.org/10.1038/s41598-020-63887-8

Viscaíno, Michelle & Maass, Juan & Delano, Paul & Torrente, Mariela & Stott, Carlos & auat cheein, Fernando. (2020). Computer-aided diagnosis of external and middle ear conditions: A machine learning approach. PLOS ONE. 15. e0229226. 10.1371/journal.pone.0229226.

Quesnel, A. M., Ishai, R., & McKenna, M. J. (2018 Apr). Otosclerosis: Temporal Bone Pathology. *Otolaryngol Clin North Am., 51*(2), 291–303. https://doi.org/10.1016/j.otc.2017.11.001

Fujima, N., Shimizu, Y., Yoshida, D., Kano, S., Mizumachi, T., Homma, A., … Shirato, H. (2019). Machine-Learning-Based Prediction of Treatment Outcomes Using MR Imaging-Derived Quantitative Tumor Information in Patients with Sinonasal Squamous Cell Carcinomas: A Preliminary Study. *Cancers., 11*, 800. https://doi.org/10.3390/cancers11060800

Tan, W., Guan, P., Wu, L., Chen, H., Li, J., Ling, Y., … Yan, B. (2021 Jun). The use of explainable artificial intelligence to explore types of fenestral otosclerosis misdiagnosed when using temporal bone high-resolution computed tomography. *Ann Transl Med., 9*(12), 969. https://doi.org/10.21037/atm-21-1171

Yao, X., Sun, K., Bu, X., Zhao, C., & Jin, Y. (2021 Dec). Classification of white blood cells using weighted optimized deformable convolutional neural networks. *Artificial Cells, Nanomedicine, and Biotechnology, 49*(1), 147–155. https://doi.org/10.1080/21691401.2021.1879823

Nikan, S., Van Osch, K., Bartling, M., Allen, D. G., Rohani, S. A., Connors, B., … Ladak, H. M. (2021). PWD-3DNet: A Deep Learning-Based Fully-Automated Segmentation of Multiple Structures on Temporal Bone CT Scans. *IEEE Transactions on Image Processing, 30*, 739–753. https://doi.org/10.1109/TIP.2020.3038363

Li, X., Gong, Z., Yin, H., Zhang, H., Wang, Z., & Zhuo, L. (2020). A 3D deep supervised densely network for small organs of human temporal bone segmentation in CT images. *Neural Netw., 124*, 75–85. https://doi.org/10.1016/j.neunet.2020.01.005

Vaidyanathan, A., van der Lubbe, M. F. J. A., Leijenaar, R. T. H., van Hoof, M., Zerka, F., Miraglio, B., … Lambin, P. (2021). Deep learning for the fully automated segmentation of the inner ear on MRI. *Scientific Reports, 11*(1), 2885. https://doi.org/10.1038/s41598-021-82289-y

He, Kaiming & Gkioxari, Georgia & Dollar, Piotr & Girshick, Ross. (2018). Mask R-CNN. IEEE Transactions on Pattern Analysis and Machine Intelligence. PP. 1-1. 10.1109/TPAMI.2018.2844175.

Simonyan, Karen & Zisserman, Andrew. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv 1409.1556.

Fu, Jun & Liu, Jing & Tian, Haijie & Li, Yong & Bao, Yongjun & Fang, Zhiwei & Lu, Hanqing. (2019). Dual Attention Network for Scene Segmentation. 3141-3149. 10.1109/CVPR.2019.00326.

Priyadarshini, Ishaani & Jha, Sudan & Kumar, Raghavendra & Smarandache, Florentin & son, le. (2018). Neutrosophic Image Segmentation with Dice Coefficients. Measurement.

Kingma, Diederik & Ba, Jimmy. (2014). Adam: A Method for Stochastic Optimization. International Conference on Learning Representations.

Wang, Pu., Ruiquan, G.e., Xuan, X., Cai, Y., Wang, G., & Zhou, F. (2016). Rectified-Linear-Unit-Based Deep Learning for Biomedical Multi-label Data. *Interdisciplinary Sciences, Computational Life Sciences, 9*. https://doi.org/10.1007/s12539-016-0196-1

Grave, Edouard & Joulin, Armand & Cissé, Moustapha & Grangier, David & Jégou, Hervé. (2016). Efficient softmax approximation for GPUs.

Szegedy, Christian & Vanhoucke, Vincent & Ioffe, Sergey & Shlens, Jon & Wojna, ZB. (2016). Rethinking the Inception Architecture for Computer Vision. 10.1109/CVPR.2016.308.