Original Article

The use of explainable artificial intelligence to explore types of fenestral otosclerosis misdiagnosed when using temporal bone high-resolution computed tomography

Weimin Tan[1#^], Pengfei Guan[2#^], Lingjie Wu[2#^], Hedan Chen[1#^], Jichun Li[1^], Yu Ling[1^], Ting Fan[2^], Yunfeng Wang[2^], Jian Li[3^], Bo Yan[1^]

[1]School of Computer Science, Shanghai Key Laboratory of Intelligent Information Processing, Fudan University, Shanghai, China; [2]ENT Institute and Otorhinolaryngology Department of Affiliated Eye and ENT Hospital, State Key Laboratory of Medical Neurobiology, Fudan University, Shanghai, China; [3]Clinical Laboratory Center, Children's Hospital of Fudan University, Shanghai, China

Contributions: (I) Conception and design: W Tan, P Guan, L Wu, H Chen; (II) Administrative support: Y Wang, J Li, B Yan; (III) Provision of study materials or patients: P Guan, L Wu; (IV) Collection and assembly of data: W Tan, P Guan, L Wu, H Chen, J Li; (V) Data analysis and interpretation: W Tan, P Guan, H Chen, J Li, Y Ling; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

[#]These authors contributed equally to this work.

Correspondence to: Bo Yan. School of Computer Science, Shanghai Key Laboratory of Intelligent Information Processing, Fudan University, Shanghai 200433, China. Email: byan@fudan.edu.cn; Yunfeng Wang. School of Computer Science, Shanghai Key Laboratory of Intelligent Information Processing, Fudan University, Shanghai 200433, China. Email: yunfengwang@fudan.edu.cn; Jian Li. Clinical Laboratory Center, Children's Hospital of Fudan University, Shanghai 201102, China. Email: lijianjulialee@126.com.

Background: The purpose of this study was to explore the common characteristics of fenestral otosclerosis (OS) which are misdiagnosed, and develop a deep learning model for the diagnosis of fenestral OS based on temporal bone high-resolution computed tomography scans.

Methods: We conducted a study to explicitly analyze the clinical performance of otolaryngologists in diagnosing fenestral OS and developed an explainable deep learning model using 134,574 temporal bone high-resolution computed tomography (HRCT) slices collected from 1,294 patients for the automatic diagnosis of fenestral OS. We prospectively created an external test set with 31,774 CT slices from 144 patients, which contained 86 fenestral OS ears and 202 normal ears and used it to evaluate the performance of our otosclerosis-Logical Neural Network (LNN) model to assess its potential clinical utility. In addition, we compared the diagnostic acumen of seven otolaryngologists with the otosclerosis-LNN approach in the clinical test set, which was mixed with 78 fenestral OS and 62 normal ears. Finally, to evaluate the assisting value of the model, the seven participants were again invited to classify all cases in the clinical test set after referring to the diagnostic results of the model, to which they were blinded.

Results: The diagnostic performance of otologists was not satisfactory, and those CT samples which were misdiagnosed had similar characteristics. Based on this finding, we defined three subtypes of fenestral OS lesions that are suitable for clinical diagnosis guidance: "focal", "transitional", and "typical" fenestral OS. The most encouraging result is that the model achieved an area under the curve (AUC) of 99.5% (per-ear-sensitivity of 96.4%, per-ear-specificity of 98.9%) on the prospective unknown external test. Furthermore, we used this model to assist otologists and observed a consistent and significant improvement in diagnostic performance, especially for the newly defined focal and transitional fenestral OS, which led to the initial high misdiagnosis rate.

Conclusions: Our findings of the fine-grained classification of fenestral OS could have implications for

^ ORCID: Weimin Tan, 0000-0001-7677-4772; Pengfei Guan, 0000-0002-3592-2377; Lingjie Wu, 0000-0002-6577-2162; Hedan Chen, 0000-0003-1582-3029; Jichun Li, 0000-0003-4906-8244; Yu Ling, 0000-0003-3329-3359; Ting Fan, 0000-0001-5978-7803; Yunfeng Wang, 0000-0001-7627-0929; Jian Li, 0000-0003-1551-412X; Bo Yan, 0000-0003-0256-9682.

future diagnosis and prevention programs. In addition, our deep OS localization network is an effective approach providing assistance to otologists to deal with the significant challenge of the misdiagnosis of fenestral OS.

Keywords: Fenestral otosclerosis; high-resolution computed tomography; artificial intelligence (AI); deep learning

Submitted Mar 24, 2021. Accepted for publication May 28, 2021.
doi: 10.21037/atm-21-1171
View this article at: http://dx.doi.org/10.21037/atm-21-1171

Introduction

According to the World Health Organization (WHO), over 5% of the world's population (about 466 million people) have disabling hearing loss and it is estimated that by 2050 over 900 million will have disabling hearing loss (1). Hearing loss has a variety of causes including genetic causes, complications at birth, infectious disease, chronic ear infections, the use of particular drugs, exposure to excessive noise, and aging. Any disease that can affect sound transmission may lead to hearing loss, of which otosclerosis (OS) is one that is both easy to be ignored and leads to progressive hearing loss.

Known also as otospongiosis, OS is an inner ear disease characterized by primary cavernous degeneration of the labyrinthine cysts. Clinical OS is not uncommon and occurs most commonly among Caucasians with an incidence of 0.3–1.2%, followed by Asians around 0.006–0.5% (2-4). Histologically, the incidence of OS is about 2.5%, which is much higher than the clinical incidence (5-10). The most frequent type of otosclerosis was fenestral otosclerosis, accounting for 91.8% of otosclerosis. It can be inferred that the incidence of fenestral otosclerosis is about 0.005% to 0.5% in clinic.

OS can be divided into stapedial OS, cochlear OS, and mixed OS based on the different locations and scope of the lesions. Cochlear OS is the terminal form of OS, which is not difficult to diagnose because of its typical clinical and CT manifestations, and the treatment is limited to the wearing of hearing aids or cochlear implantation (11). OS first effects and is most commonly found in the anterior area of the vestibular window, which causes stapedial OS, resulting in conductive hearing loss due to fixation of the stapes footplate. In addition to the diagnosis of OS based on clinical symptoms, signs, and audiological examination, the diagnostic value of high-resolution CT has also been widely recognized. With a positive diagnosing rate of 74% to 95.1% (10,12,13), high-resolution computed tomography (HRCT) is considered the first choice for the diagnosis of OS.

However, stapedial OS is often misdiagnosed, and most likely to be misdiagnosed as sensorineural deafness, congenital stapes fixation and tympanosclerosis. Missed diagnosis can also occur when combined with other ear diseases such as chronic otitis media (COM). According to Huang's retrospective study of 37 cases, the incidence of misdiagnosis of fenestral otosclerosis is around 27% (14). Several reasons may account for this. Firstly, the clinical manifestation of stapedial OS is not typical, and most patients see a doctor with simple hearing loss. Secondly, compared with common ear diseases such as chronic suppurative otitis media and middle ear cholesteatoma, the incidence of OS is relatively low. Thirdly, and most importantly, although temporal bone CT is a highly specific examination in otology, stapedial OS lesions are not obvious using this method because the lesions are very small. A misdiagnosis may carry serious implications as untreated hearing loss may affect the ability of patients to communicate with others, hinder their daily activities and ability to work, and cause loneliness and depression. Wrongly diagnosed conditions may also subject patients to potentially unnecessary or harmful medical treatment.

Multiple etiologies have been postulated, including, genetic, hereditary, sex, ethnicity, pregnancy, and viral infections. Sodium fluoride is prescribed to slow the disease, however, the efficacy is still controversial. Surgical correction of sound transmission disorder caused by stapes fixation is an effective method to improve hearing. Also, hearing aids are used in many patients to improve their hearing. The timely and accurate diagnosis of stapedial OS renders successful treatment with stapes implantation more likely and often results in the recovery of complete normal hearing function. Therefore, it is necessary to design an intelligent diagnosis system based on temporal bone CT scans to improve stapedial OS diagnostic efficiency.

In recent years, artificial intelligence (AI) technology represented by deep learning has achieved great success in various fields. Deep learning is a machine learning technique that automatically learns the most informative and representative features from input images given a large data set of labeled examples, which avoids man-induced factors such as extensive preprocessing and extraction of handcrafted visual features. This technique uses an optimization algorithm called back propagation to indicate how a machine should change its internal parameters to best predict the desired output of an image. Recently, introducing deep learning technique into computer-aided detection or diagnosis (CAD) has attracted great attention and is considered to be one of the revolutionary directions for future medical development.

Deep learning based CAD of CT abnormalities has achieved great success in various fields (15,16) such as integrating chest CT with other medical information for detecting patients with COVID-19 (17), lung nodule detection based on CT images (18), monitoring organs at risk delineation in CT images (19), and low-dose CT image reconstruction for improving lesion detectability (20). However few studies have been devoted to diagnosing ear diseases using CT volume. Mei *et al.* developed an approach for rapidly diagnosing disease by using chest CT and the clinical history (17) and achieved an area under the curve (AUC) of 0.92 and a sensitivity of 84.3% when tested on 279 patients. These results showed chest CT is an important method for screening early suspected COVID-19 infection patients. Hwang *et al.* used a deep learning model to classify 54,221 chest radiographs into normal and abnormal, and achieved an AUC of 0.979 (0.973–1.000) for image classification and 0.972 (0.923–0.985) for lesion detection (21). Shan *et al.* proposed an iterative CT enhancement deep neural network for improving low-dose CT image quality (20) and demonstrated a superior visual quality compared with three commercial algorithms. Collectively, these studies demonstrate the feasibility and effectiveness of utilizing CT images for disease diagnosis.

The CT value of the fissula ante fenestram of patients with stapedial OS is significantly lower than that of normal people, and developing a diagnostic system of stapedial OS based on AI is a promising choice. However, there are few studies on AI in the diagnosis and treatment of otological disease using CT images. To date, only one retrospective study based on temporal bone CT has described the application of AI in the distinction between COM and middle ear cholesteatoma (22). This is because the imaging manifestations of ear diseases are much smaller in size than those of other diseases. Therefore, it is a great challenge to train an AI model to detect stapedial OS lesions because lesions are extremely "indistinct" in CT.

We conducted a study to assess the clinical performance of otolaryngologists to diagnose OS and made an in-depth analysis of the results. We found that otolaryngologists have a clear concentration tendency to diagnose some ear samples and based on this finding, we then subdivided the types of OS and newly defined them as focal, transitional, and typical fenestral OS. We contend this new fine-grained classification can provide a guide for screening OS in clinical settings. In addition, the unsatisfactory results of otolaryngologists to diagnose OS prompted us to develop a novel explainable OS localization deep neural network [otosclerosis-Logical Neural Network (LNN)] for the automatic diagnosis of fenestral OS in temporal bone high-resolution CT images. The otosclerosis-LNN model demonstrates a promising diagnosis result that outperforms the diagnostic level of otolaryngologists. Interestingly, it also shows a similar distribution in the diagnosis of our newly defined fenestral OS types on the test set. Finally, we used the otosclerosis-LNN model to assist all otolaryngologists and found this improved their diagnosis level, especially for focal and transitional fenestral OS which had an initially high misdiagnosis rate.

Our study is the first to apply deep learning techniques to extract the region of interest from whole-volume HRCT scans of temporal bones for diagnosing fenestral otosclerosis. We validated the proposed model to be an effective computer-aided diagnosis model of fenestral otosclerosis in a large-scale study containing 31,774 CT slices from 144 patients. The model achieved an AUC of 99.5% (per-ear-sensitivity of 96.4%, per-ear-specificity of 98.9%) on the collected test set, indicating great diagnostic performance. Furthermore, we used this model to assist otologists and observed a consistent and significant improvement in their diagnostic performance. We present the following article in accordance with the STARD reporting checklist (available at http://dx.doi.org/10.21037/atm-21-1171).

Methods

Dataset collection

The study was conducted in accordance with the

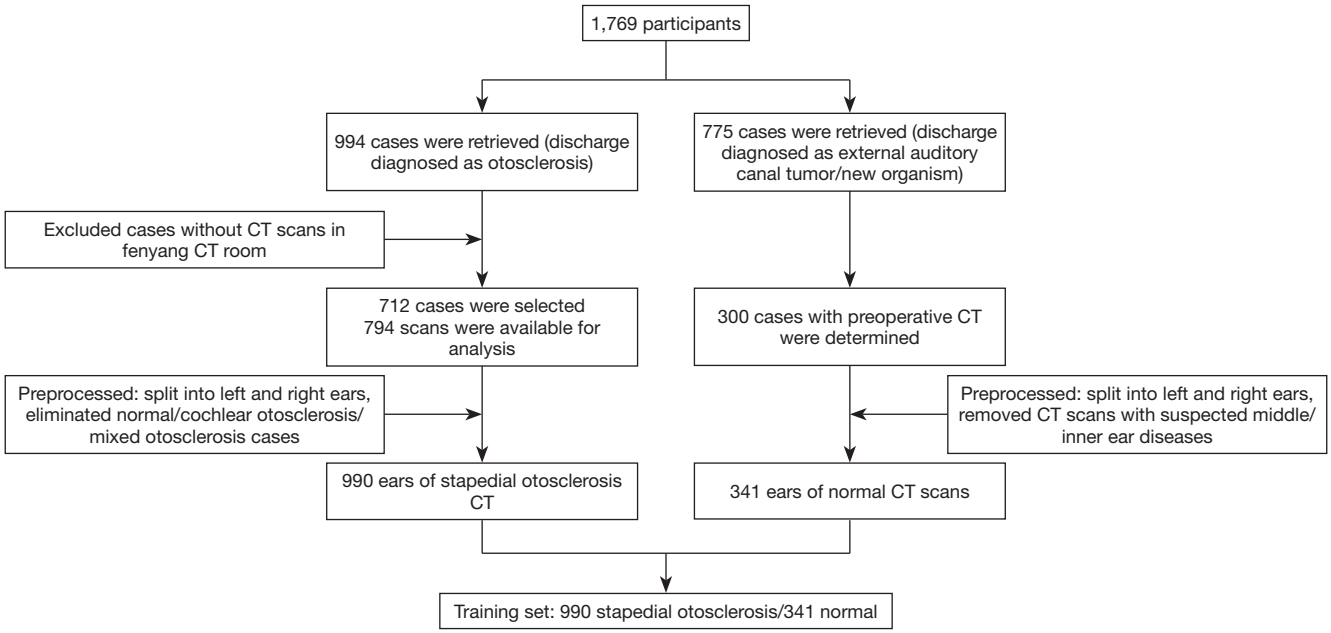


Figure 1 Flow chart of the training set collection.

Declaration of Helsinki (as revised in 2013). The study was approved by institutional committee of Eye and ENT Hospital, Fudan University (No. 2020005) and informed consent was taken from all the patients. All authors had access to the study data and reviewed and approved the final manuscript. All training CT scans were collected from the central CT room of Eye and ENT Hospital, Fudan University and were carried out by a 128-channel multidetector SOMA TOM Definition Edge CT scanner (Siemens Inc., Munich, Germany). Each CT scan contained the complete structure of the temporal bone. Axial sections of 0.6 mm thick were obtained with the following CT scanner settings: collimation of 128 mm × 0.6 mm, field of view of 220 mm × 220 mm, pitch of 0.8 mm, matrix size of 512×512, voltage of 120 kV, and current of 240 mAs. The number of axial CT slices in each scan was around 100 and all images were downloaded from the Cloud server and saved in a 512×512 size and DCM format for training.

A total of 994 cases in the training set who received artificial stapes implantation in the central area of Eye and ENT Hospital, Fudan University from July 16, 2014, to November 13, 2019, were identified (see *Figure 1* and Table S1). Among them, 712 cases that underwent one or more CT scans were selected, and a total of 794 scans were available for analysis. After removing the normal/cochlear OS/mixed OS ears and operated ears, 990 ears of stapedial OS CT were obtained. Finally, we sought cases with external auditory meatus tumor/new organism resection in the central area of Eye and ENT Hospital, Fudan University from July 16, 2014, to November 13, 2019. We obtained 775 cases, determined 300 cases, removed the suspected middle ear/inner ear disease ear CT scans by checking the medical history and examination results, and finally pretreated 341 ears of CT scans as a normal control study.

All testing CT scans were collected from the new CT room of Eye and ENT Hospital, Fudan University from December 1, 2019, to April 30, 2020. These scans were carried out by a 128-channel multidetector SOMA TOM Definition Flash CT scanner (Siemens Inc., Munich, Germany) and with the same CT scanner settings, axial sections of 0.6 mm thick were obtained. All images were downloaded from doctors' workstation and saved in a 512×512 size and JPG format for testing.

Manual labelling

To train the otosclerosis-LNN model, we manually annotated the ear axis CT in the training set and developed a Bounding-Box Marker annotation software to facilitate the annotation process of otologists. Each CT was then preprocessed, split into left and right ears, and

combined with the preprocessing algorithm to determine a rectangular clipping box. This was to ensure that the structure of the inner ear, middle ear, and outer ear could be displayed on each level, thus removing irrelevant areas and enlarging related areas to facilitate further labelling. For stapedial OS region labelling, otologists consulted the case information (mainly surgical records) and image reports to ensure the accuracy of labelling and annotated each CT slice with special attention to the vestibular window area. For the reduced bone mineral density shadow (the location of stapedial OS), experienced doctors used a rectangular red box to mark as close as possible to the focus and after the annotation of each mark, a new folder was be generated for the AI model to learn. The non-OS CT samples were then annotated. We first ensured there was no OS by carefully checking the medical history and imaging reports, then marked a rectangular area around the stapes footplate with a green box because the location of stapedial OS is relatively fixed. This space is relatively large, covering the possible location of stapedial OS lesions, and contains more labelling layers, covering the possible layers of fenestral OS. Similarly, the annotation result of each ear CT image was placed in the newly generated folder for AI model training.

Model architecture

The main purpose of this study was to develop an explainable OS localization deep neural network (otosclerosis-LNN) for the detection and triage of ears based on whole-volume temporal bone HRCT scans. The average number of CT slices per patient was 103 , and only about 3% contained lesions, indicating CT slices containing fenestral OS accounted for a very small proportion of the total number of CT slices in the collected dataset. The stapedial OS region is also small in the whole CT slice, accounting for less than 0.2%, resulting in a severe class imbalance problem. To deal with this, we decomposed the fenestral OS diagnosis of ear CT volume diagnosis into three stages; a pre-processing stage, OS detection network, and post-processing stage (*Figure 2A*,*B*). The first stage is a conventional image processing algorithm, which is used to filter out non-ear CT slices and automatically crop out areas that may be ears and there was a relative class balance between normal and fenestral OS in the cropped ear slices. Following this, the cropped ear regions were fed into the deep detection network for localizing the OS area. There will usually be several continued slices and discrete slices of the input 3D head CT scan detected including OS lesions and normal structure. Therefore, the post-processing stage was to heuristically fuse all CT diagnosis results, producing the final ear diagnosis result, which will be normal or fenestral OS.

Pre-processing stage

A conventional image processing algorithm was developed to automatically crop out ear regions from the temporal bone HRCT slices. The first step was based on the first temporal bone CT image, and the algorithm binarized this with a brightness threshold of T_1=50. Those regions with brightness greater than T_1 were maintained and resized to 122×364 using the Bicubic algorithm. The right and left ear areas in those regions were then cropped out in the pattern of [x1, y1, x2, y2], where [1, 181, 184, 295] were for the right ear and [184, 185, 364, 295] were for the left ear. Finally, the cropped ear areas were resized to 888×1,496 and normalized to 0–1 facilitating the detection of the following detection network.

Detection network

The detection network was used to detect fissula ante fenestram areas in the cropped ear CT slices produced by the pre-processing stage. The overall network structure of the detection network was mainly based on the Faster-RCNN model (23), which is a classic deep network model for detecting natural objects. We modified this by replacing its backbone using the pre-trained VGG-19 model (24), which helped to extract more general and discriminative visual features (512×25×42). These visual features were used to generate the features of each bounding box region in the input CT slice by the region proposal network (RPN) and region of interest pooling (RoI Pooling) (23,25). By performing bounding box regression and classification, we obtained the OS detection results for each CT slice. In general, there were multiple continued slices and discrete slices of the input 3D head CT scan detected including OS lesions. Therefore, we needed a post-processing stage to heuristically fuse all CT detection results, producing the final ear diagnosis result, which was normal or OS.

Post-processing stage

This stage output the final ear diagnosis result (normal or fenestral OS) by fusing the detection results of all CT

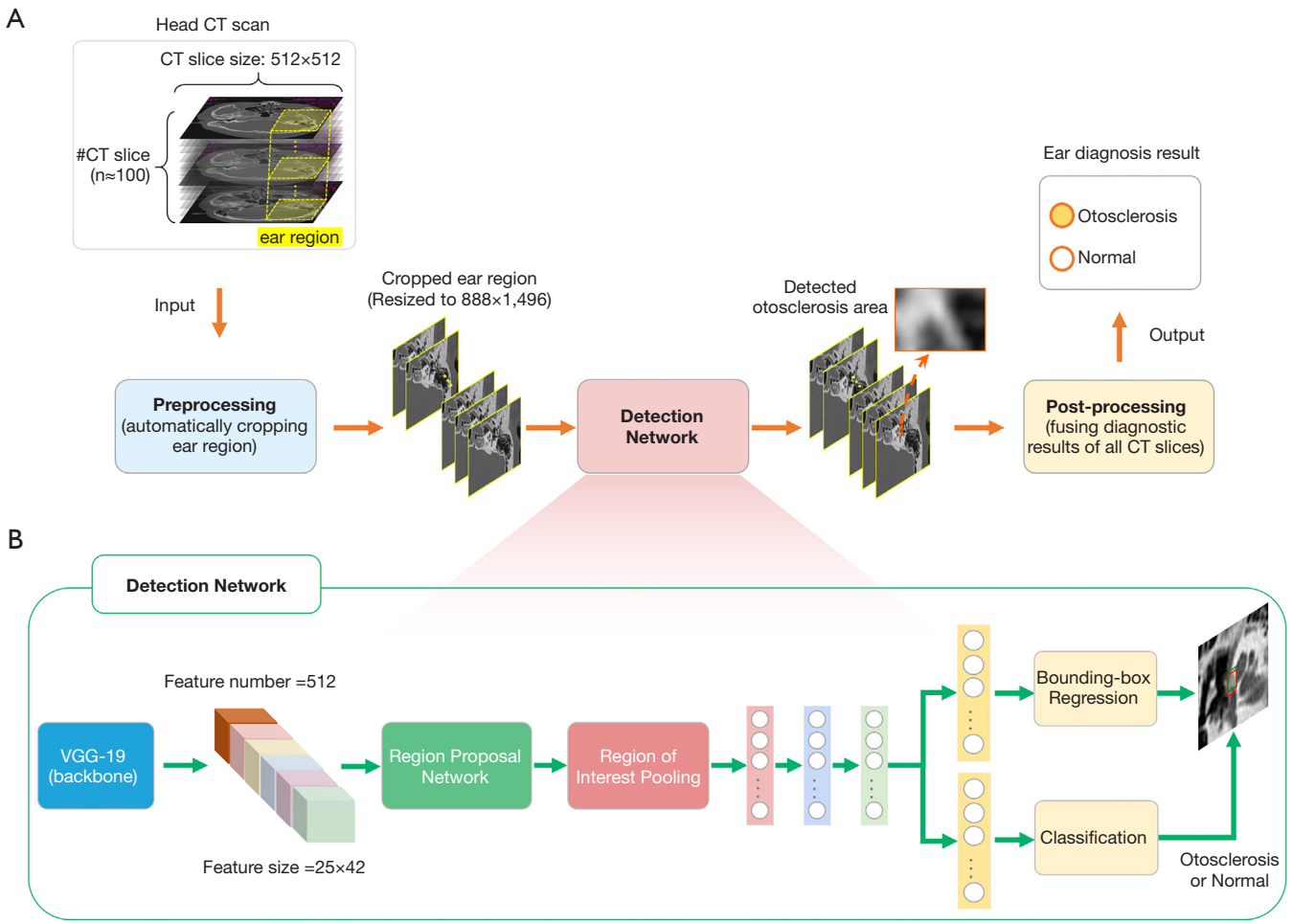


Figure 2 Overview of our proposed otosclerosis localization neural network (otosclerosis-LNN). (A) Processing pipeline of the otosclerosis-LNN model. This mainly consists of three stages: the pre-processing stage for automatically cropping out possible ear regions by using a conventional image processing algorithm, the detection stage for localizing otosclerosis lesion areas by using a deep learning-based detection network, and the post-processing stage for outputting the final ear diagnosis result (normal or otosclerosis) by heuristically fusing the diagnostic results of all CT slices. (B) Detailed structure of the detection network. LNN, Logical Neural Network.

slices. Firstly, those CT slices with a low probability of normal and OS bounding box (less than threshold P_1=0.5) were filtered out. The purpose of this step was to make a preliminary screening and directly exclude the results with a significantly low confidence level, thus avoid the selection of the subsequent longest subsequence. Therefore, the threshold setting could be relatively relaxed. The subsequence with the longest consecutive CT layers was then selected from the remaining stratification results. If there were multiple subsequences with the same length, the one with higher maximum confidence was selected. Since the fissula ante fenestram area is continuous, the purpose of this step was to determine the area of the stapes footplate in the CT layer. Next, the confidence of the results in the subsequence was filtered, and all the results whose confidence probability was less than the threshold P_2=0.99 were eliminated. The purpose of this step was to select the most valuable results in the subsequence for subsequent diagnosis, so the threshold setting should be relatively high. Finally, we calculated the ratio between the number of slices of the selected subsequences detected as lesions and the total number of remaining slices. We considered that stapedial OS was present when the ratio was greater than 0.25 and if not, the ear was considered to be normal. The purpose of this step was to synthesize all CT slice results to obtain the final diagnosis result.

Training details

The otosclerosis-LNN model was trained on the collected training set. Data augmentations such as random horizontal flipping and adding noise were employed in the training process to enhance the generalization ability of the model. We used the Adam optimizer (26) with a small initial learning rate 1e-5 to minimize the bounding box regression and classification losses. The output of the otosclerosis-LNN model represented the probability that the CT slices were OS or normal; if the probability was greater than 0.5, the slice was classified as fenestral OS and otherwise, as normal. We trained the otosclerosis-LNN model on a Nvidia RTX 2080Ti graphics processing unit with 8 Gb of memory and Intel i5-8400@2.80GHz with 24GB RAM with the training batch size set to 4. The number of training iterations was 269 epochs, which took about three days (68 hours) and the model was implemented with the PyTorch library (27) and Python 3.7 language.

Statistical analysis

We evaluated the performance of our otosclerosis-LNN model to assess its potential clinical utility by prospectively collecting an external test set. Given the temporal bone HRCT volume of an ear, the model localized the position of OS lesions in the input CT volume and output its probability that the ear was fenestral OS or normal. According to the setting confidence for CT slices, ROC was obtained. The sensitivity, specificity, accuracy, PPV, and NPV of this approach on the external test set were 96.43%, 98.86%, 98.07%, 97.59%, and 98.30%, respectively. In addition, we compared the otosclerosis-LNN approach with seven otolaryngologists (two chief physicians, three associate chief physicians, one attending doctor, and one resident doctor) on the clinical test set. Each otolaryngologist independently classified each case as fenestral OS or normal, there was no limit to the screening time of each doctor, and only the diagnostic accuracy was assessed. Finally, we evaluated the ability of the otosclerosis-LNN model to assist otolaryngologists in diagnosis. All otolaryngologists were blind to the performance of the model and after three months, all participants were invited to classify all cases in the clinical test set after referring to the diagnostic results of the model.

Results

Study of otolaryngologists' performance in clinically diagnosing OS

To explore the clinical performance of otolaryngologists in diagnosing OS, we invited seven otolaryngologists to assess their diagnostic level of stapedial OS on the collected test set (see *Figure 3* for the collection process). Participants included two chief physicians with over 20 years of experience, two associate chief physicians with about 15 years of experience, one associate chief physician with about 10 years of experience, one attending physician with about 10 years of clinical experience, and one resident with about 2 years of experience. Each doctor was required to independently diagnose stapedial OS and normal temporal bone by only reading CT scans in the absence of other clinical information. It is important to note that before we invited the seven doctors to participate in the test, we also invited other doctors who had just entered the hospital or had worked for 5 years to take the test. As most of the latter group did not know the CT manifestation of fenestral OS, their results were a right or wrong guess, which fully reflected the challenges in the clinical diagnosis of fenestral OS and the high diagnostic level of the seven invited otolaryngologists.

The external test set was prospectively collected as shown in *Figure 3*. By April 30, 2020, a total of 42 cases of bilateral stapedial OS and two of unilateral stapedial OS were collected (see Table S2). The operative ear diagnosis was based on what was seen during the operation, which was considered as the gold standard. The contralateral ear diagnosis was based on the comprehensive analysis of imaging experts reports and medical history, forming 86 ears with stapedial OS in the external test set. In addition, the information of outpatients in new area of Eye and ENT Hospital, Fudan University in April 2020 was collected, and 100 cases of completely normal CT images were included in the test set according to the diagnosis report of imaging experts combined with clinical manifestation and examinations.

In summary, the external test set consisted of 86 ears of stapedial OS and 202 normal ears. To reduce the diagnostic deviation caused by fatigue in reading the images, we randomly selected 30 cases from 100 cases with normal CT and 40 cases from 44 cases with stapedial OS, finally

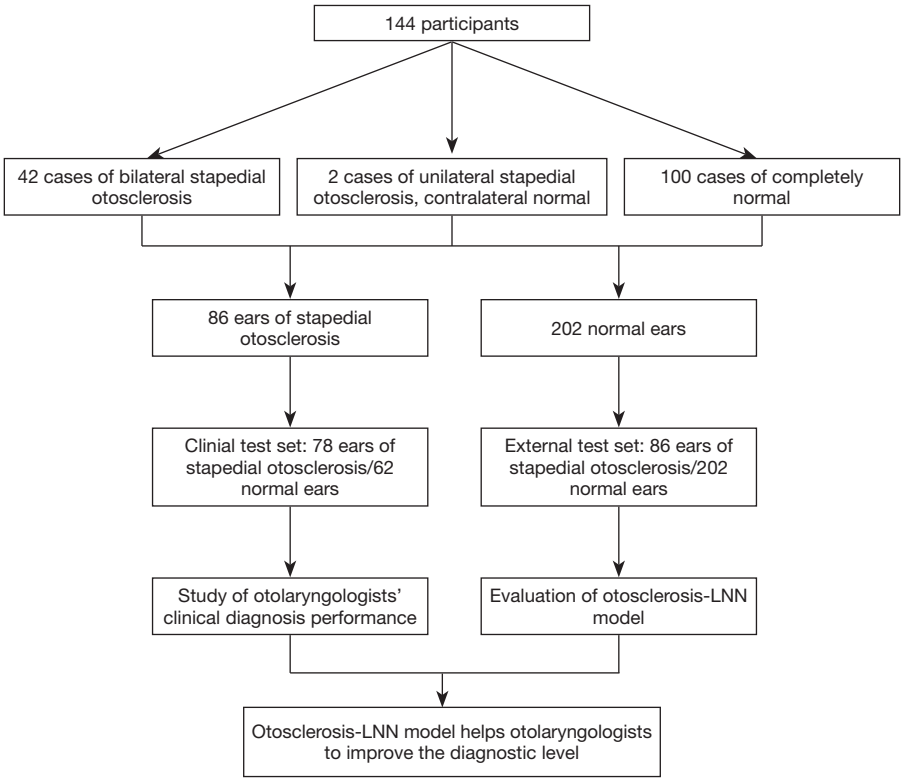


Figure 3 Flow chart of the collection of the retrospective clinical test set. LNN, Logical Neural Network.

forming the clinical test set for doctors to read as 78 ears of stapedial OS mixed with 62 normal ears. The prospective and complete collection of OS cases admitted in the new area of Eye and ENT Hospital, Fudan University began on December 1, 2019 and was part of our project *Artificial intelligence in diagnosis and treatment decision-making of middle ear diseases based on temporal bone CT scans* (Registration number: ChiCTR1900027535) and strictly carried out in accordance with the plan. This trial was verified on November 17, 2019, and detailed information can be seen on the website, Chinese Clinical Trial Registry.

This clinical study demonstrated the performance of otolaryngologists to diagnose stapedial OS was unexpectedly low (see *Figure 4*). The two chief physicians achieved the best diagnostic average results in terms of sensitivity and accuracy (Figure S1 and Table S3) and the two associate chief physicians achieved average sensitivity and specificity of 70.07% and 67.70%, respectively. The attending physician (~10 years) achieved sensitivity and specificity of 67.90% and 62.90%, respectively, and the resident (~2 years) achieved sensitivity and specificity of 66.70% and 100%, respectively. The results show that while compared with other doctors, the chief physicians demonstrated a higher diagnostic level in the test set, there is still much room for improvement and that misdiagnoses occur frequently.

Defining new lesion types of fenestral OS based on the analysis of otolaryngologists' diagnosis results

The unsatisfactory performance of otolaryngologists promoted us to further analyze what types of fenestral OS led to misdiagnoses. We counted the diagnosis results of each ear by each doctor in the test set and obtained the average diagnosis accuracy and variance of each ear (*Figure 5A*). It should be noted that we did not include the diagnosis of the attending physician in the definition process. The attending physician pays more attention than others to the facial nerve, as he believes that the fenestral OS will blur the horizontal segment of the facial nerve in images. Although this has a certain diagnostic efficacy, we believe that only typical fenestral OS will have this manifestation. As a precaution, we removed his diagnosis during the definition process, and the diagnosis of the

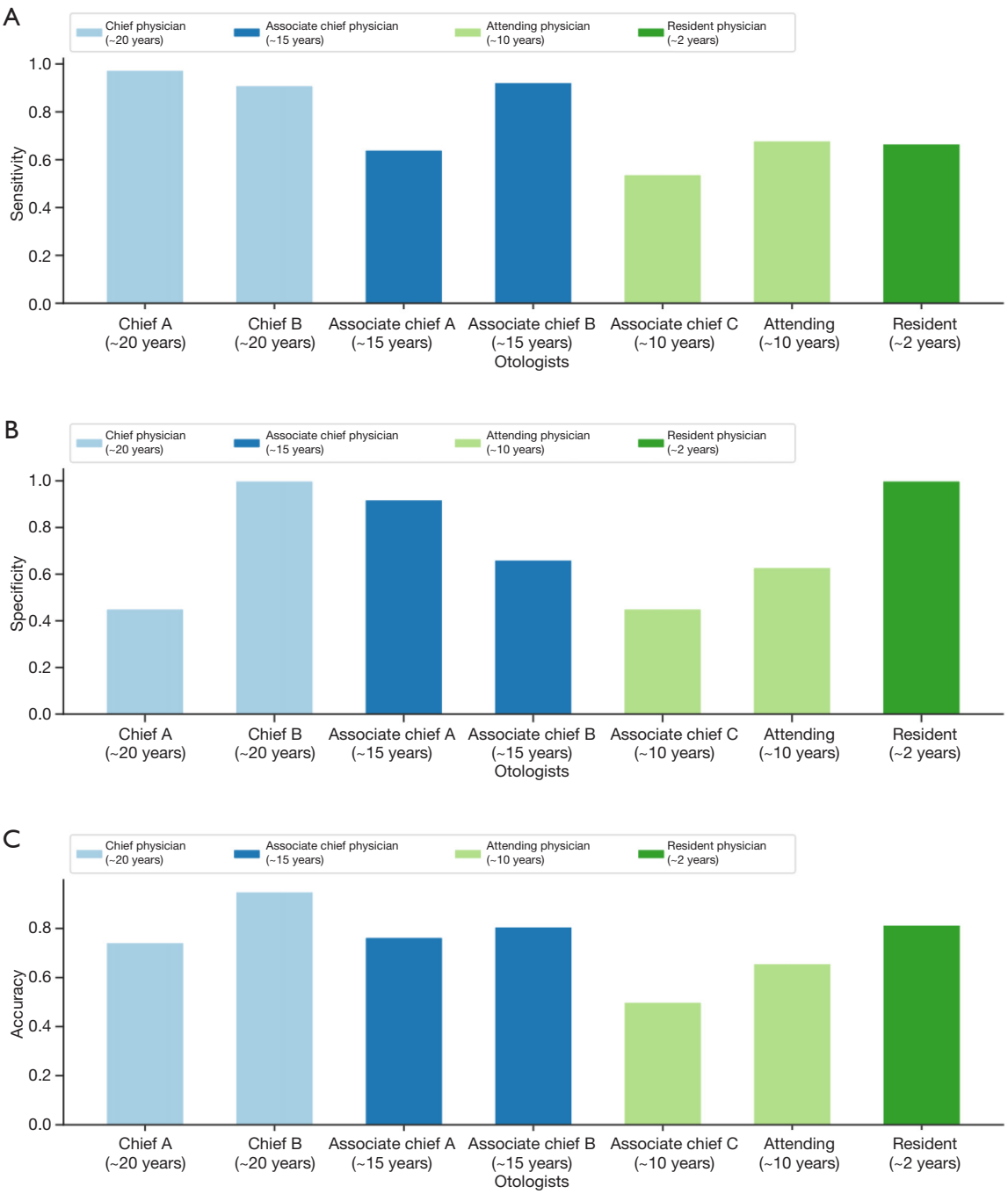


Figure 4 Demonstration of the performance of otolaryngologists to diagnose otosclerosis using the retrospective test set. (A,B,C) Show the sensitivity, specificity, and accuracy of seven otolaryngologists, respectively.

other six otologists contributed to this new subdivision. According to their average diagnostic accuracy, we further classified the fenestral OS lesions (see *Figure 5B*) with an average diagnostic accuracy of 0.33, 0.5 or 0.67, 0.83, or 1.0 as "focal", "transitional", or "typical" fenestral OS, respectively. These three types were highly related to the misdiagnosis rate of otologists, and we counted the number of these three types of OS in the test set (*Figure 5C* and

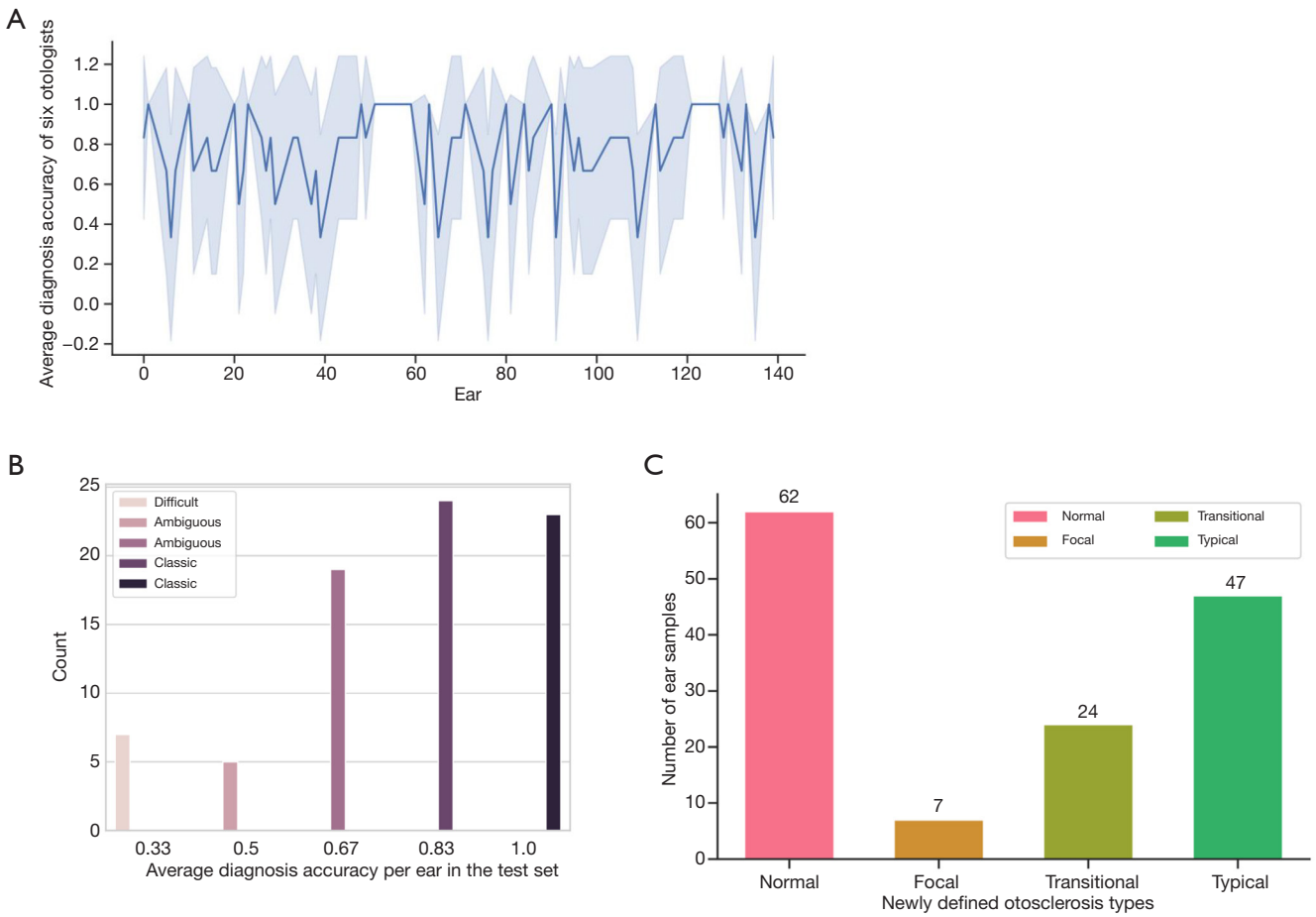


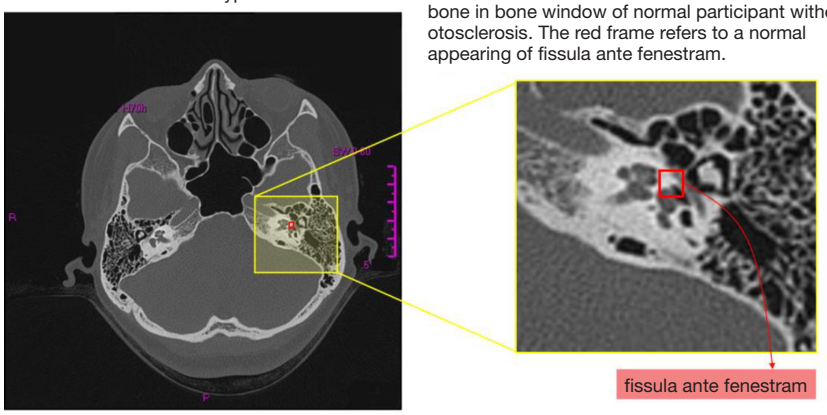
Figure 5 Statistical analysis of the otolaryngologists' diagnosis findings. Based on this, the otologists further classified the lesion types of fenestral OS as "typical", "transitional", and "focal" fenestral OS. (A) Demonstration of the average diagnostic accuracy of seven otolaryngologists for 140 ears in the clinical test set. The shade denotes the standard deviation of each ear. (B) Counting the number of ears based on its diagnostic accuracy. The ears with a diagnostic accuracy of 0.33, 0.5 or 0.67, 0.83, or 1.0 are defined as the focal, transitional, and typical types of fenestral OS, respectively. (C) Demonstration of the number of our newly defined fenestral otosclerosis types in the test set. OS, otosclerosis.

Figure S2).

We found that there was an obvious concentration trend in the diagnosis results of doctors for some ear samples. For example, some ears were misdiagnosed by most otolaryngologists, and some were correctly diagnosed with high confidence. The diagnosis results of some ears showed large variance, indicating that the otolaryngologists have differed considerably in the diagnosis of these ears. Based on this finding, we brought these CT images to experienced otologists for further discussion and study. The otologists found that these images with different diagnostic accuracy demonstrated a relatively slight appearance difference (see *Figure 6*).

The appearance of typical fenestral OS in high-resolution CT (HRCT) images is of a hypodense demineralised plaque in the region of the fissula ante fenestram, or as heaped-up bony otosclerotic plaques causing oval window narrowing. Thickened stapes footplate can also be seen. Transitional fenestral OS is ambiguous to diagnose as HRCT shows mild new bone involving the stapes footplate, narrowing of the oval window, a slight thickening of the stapes footplate, or new bony plaque formation at the fissula ante fenestram. Focal fenestral OS is also difficult to diagnose because the HRCT shows very subtle demineralization at the fissula ante fenestram and there is no thickening of the stapes plate.

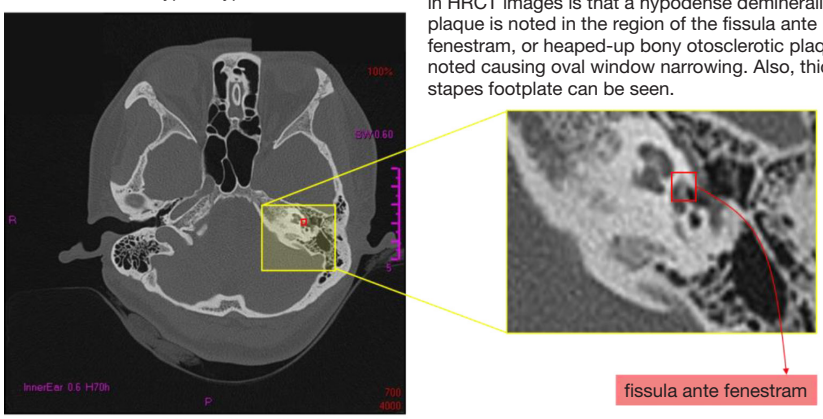
Normal type



Appearance characteristic. Axial view of a high resolution computed tomography image of temporal bone in bone window of normal participant without otosclerosis. The red frame refers to a normal appearing of fissula ante fenestram.

CT number: CT202004030088

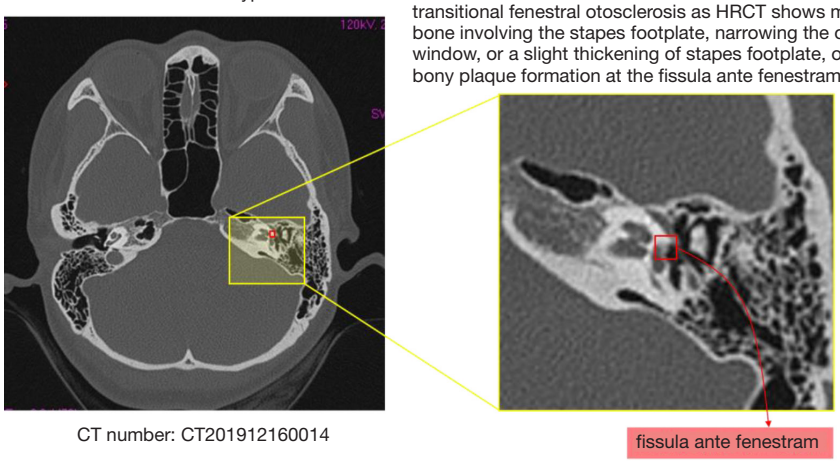
Typical type



Appearance characteristic. The appearing of this type in HRCT images is that a hypodense demineralised plaque is noted in the region of the fissula ante fenestram, or heaped-up bony otosclerotic plaques are noted causing oval window narrowing. Also, thickened stapes footplate can be seen.

CT number: CT201901020066

Transitional type



Appearance characteristic. It is ambiguous to diagnose transitional fenestral otosclerosis as HRCT shows mild new bone involving the stapes footplate, narrowing the oval window, or a slight thickening of stapes footplate, or new bony plaque formation at the fissula ante fenestram.

CT number: CT201912160014

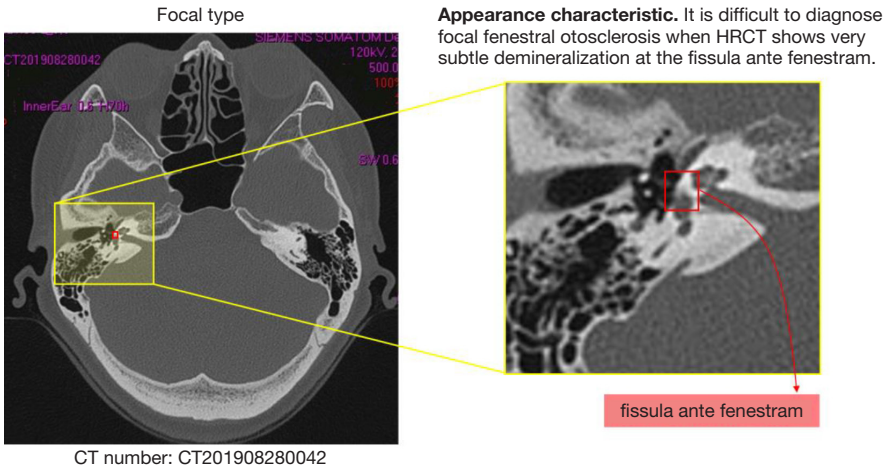


Figure 6 Visualization of the four types of the fissula ante fenestram region in HRCT images. The three types of stapedial otosclerosis demonstrated different detailed CT appearances, where the "focal" type showed similar CT features as the "normal" type, which may easily lead to a misdiagnosis. HRCT, high-resolution computed tomography.

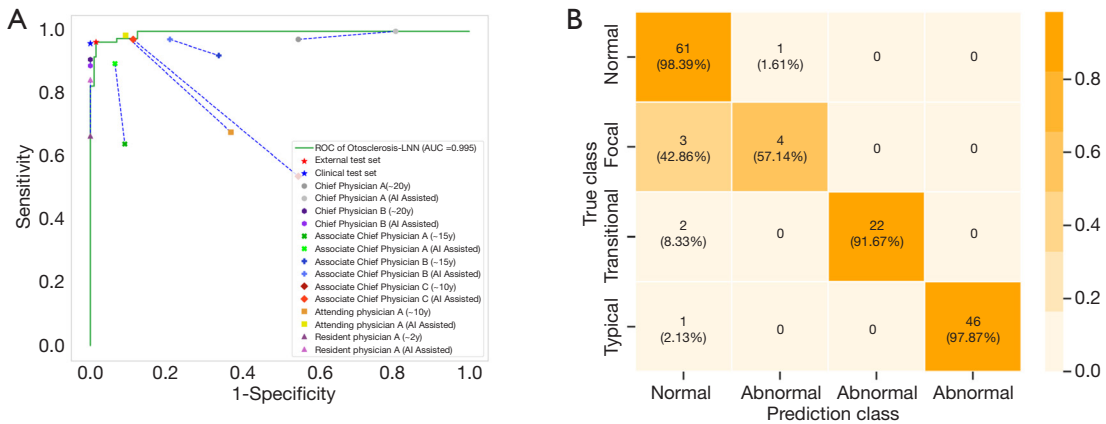


Figure 7 Assisting otolaryngologists to screen fenestral OS with the assistance of the otosclerosis-LNN model. (A) Demonstration of the diagnostic performance of the otosclerosis-LNN model, all otolaryngologists, and their diagnostic performance assisted by the model. (B) We showed another type of confusion matrix for demonstrating the detailed prediction results of our newly defined fine-grained otosclerosis types. OS, otosclerosis; LNN, Logical Neural Network.

Development and validation of an OS localization deep neural network based on temporal bone high-resolution CT images

The unsatisfactory results of the study evaluating the performance of otolaryngologists in clinical diagnosis results caused us to develop a CAD system to help otologists obtain a better diagnosis performance. For this purpose, we developed an explainable OS localization deep neural network (otosclerosis-LNN) for the detection and triage of ears based on whole-volume temporal bone HRCT scans (see *Figure 7A,B* for detailed model architecture). To develop the model, we collected 134,574 CT slices from 1,294 patients who underwent a CT examination at Eye and ENT Hospital, Fudan University from July 16, 2014, to November 13, 2019 (Table S3). All scans were performed by radiologists using a standard chest CT protocol and the dataset included 994 OS patients and 300 normal patients (see *Figure 1* for the collection process). All CT scans were preprocessed, split into left and right ears, and each CT image was cut into a rectangular region (containing ear

structure) to facilitate further labelling of the lesions.

In the labelling process, we carefully checked the case information (mainly surgical records and imaging reports), eliminated normal/cochlear OS/mixed OS temporal bone CTs, and finally obtained 990 ears of stapedial OS. The label (diagnosis) of the surgical ear was marked according to the intraoperative diagnosis while for the unoperated ear, the label was a comprehensive diagnosis based on imaging reports of an imaging expert and medical history. The average number of CT slices per patient was 103, and only about 3% contained visible lesions. Additionally, the proportion of fenestral OS in the total number of CT slices of temporal bone was 1.83% (see the rightmost figure in Figure S3). Such a small proportion indicates that CT slices containing fenestral OS account for a very small proportion of the total number of CT slices in the collected dataset and shows the number of normal and non-ear CT slices is large and represents a serious class imbalance problem. Class imbalance typically poses a great challenge to designing model structures (25,28), updating model parameters for convergence, and maintaining the balance of test performance for each category.

To deal with this extreme class imbalance problem, we employed an end-to-process strategy that could effectively turn this into a class balance problem by fully exploiting the advantages of conventional image processing algorithms and deep learning technology. Specifically, we used ear CT slices instead of whole temporal bone CT slices to train the deep learning-based OS detection network because the proportion of fenestral OS in the total number of ear CT slices of temporal bone was 71.21% (see the rightmost figure in Figure S3). Therefore, the class imbalance problem was largely alleviated.

To evaluate the effectiveness of our otosclerosis-LNN, we conducted the fenestral OS localization experiment on the collected test set containing 31,774 CT slices from 288 ears (see *Figure 3* and Table S1). The whole-volume temporal bone HRCT scans were fed into the otosclerosis-LNN model. The model then output the positions of fenestral OS in the inputted ear CT scans and the classification result of normal or fenestral OS for that ear. The AUC, sensitivity, and specificity of our otosclerosis-LNN approach were 99.5%, 96.4%, and 98.9%, respectively. *Figure 7A* and Figures S4,S5 show the comparison with all otolaryngologists, which had a higher screening performance than both the two chief physicians (~20 years) and other doctors.

It is also interesting to understand the diagnostic performance of the otosclerosis-LNN model for our newly defined three types of fenestral OS mentioned in the above section. We demonstrated another type of "confusion matrix" (see *Figure 7B*), where the vertical axis denotes the true class containing four types and the horizontal axis denotes the predicted class containing only two types (normal or abnormal) because the output of the otosclerosis-LNN model is two types, i.e., normal or fenestral OS. This new type of confusion matrix clearly demonstrated the detailed prediction results for the newly defined fine-grained OS types. The matrix demonstrated that the otosclerosis-LNN model achieved good diagnostic performance for transitional and typical types, while showing ambiguous judgment when dealing with focal ear CT scans. This result indicates that compared with doctors, the improvement of diagnostic performance of the model is mainly due to the improvement of diagnostic accuracy on the transitional ears. Simultaneously, we observed a gradual decline trend of the diagnostic accuracy of the model for typical, transitional, and focal types, which indirectly indicated the further classification of fenestral OS discussed in the above section was reasonable. In addition, we found that the serious class imbalance problem did not lead to significant class bias in the test results of our otosclerosis-LNN model.

Improving the ability of otolaryngologists to screen fenestral OS with the assistance of the otosclerosis-LNN model

We further conducted an experiment using the otosclerosis-LNN model to assist otolaryngologists in screening fenestral OS in temporal bone CT scans. Firstly, we set up a week for memory washing out, then after three months, we invited the seven otolaryngologists to again screen fenestral OS in the temporal bone CT scans of the test set with reference to the results of the otosclerosis-LNN model while the diagnostic performance of the otosclerosis-LNN model was blind to the otolaryngologists. The experimental results (see *Figure 8*, Figure S6 and Table S4) demonstrated that the diagnostic performance of all otolaryngologists was significantly improved with the assistance of the otosclerosis-LNN model. The category of significant improvement was in the focal and transitional ear samples, probably because these two types are the main reasons for the high rate of misdiagnoses. Overall, the average diagnostic ability of the seven otolaryngologists improved in all lesion types.

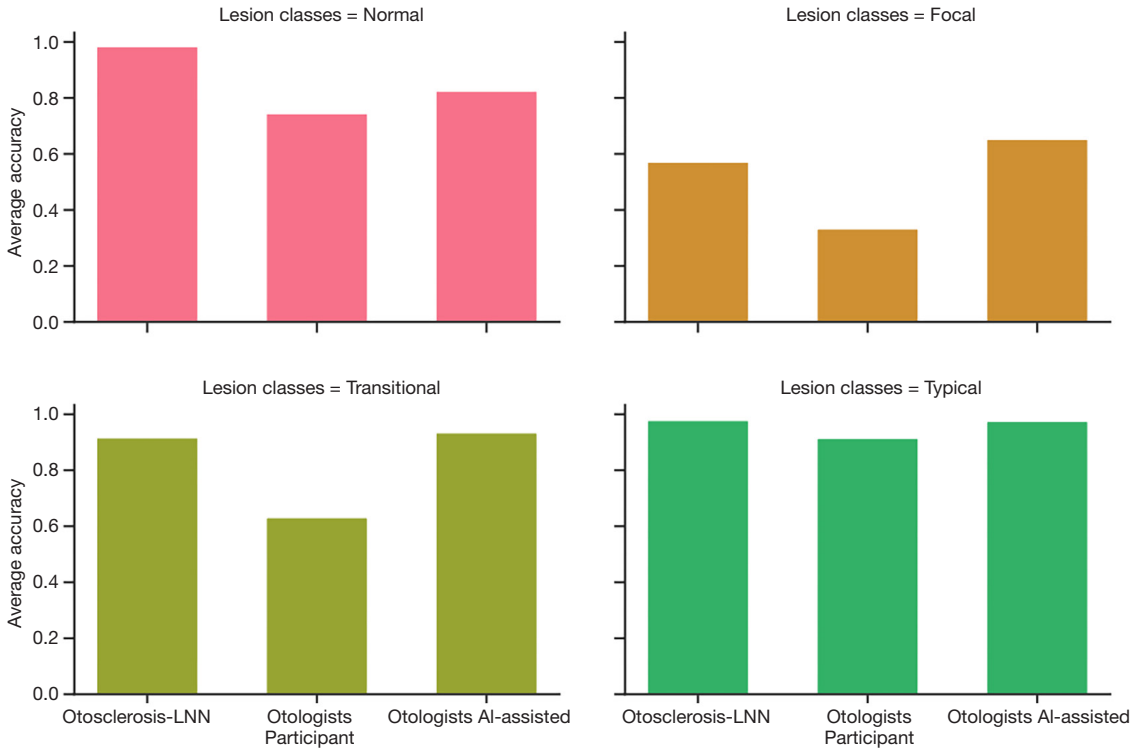


Figure 8 Demonstration of the average diagnostic accuracy of the otosclerosis-LNN model, all otologists, and their diagnostic performance assisted by the model. LNN, Logical Neural Network.

Explanation of the detection network in the otosclerosis-LNN model

The interpretable artificial intelligence in our study means that the proposed diagnostic model not only provides diagnostic results, but also provides diagnostic basis for doctors. The diagnosis basis we provide includes three aspect. The neural responses learned by the detection network are visualized, showing a pleasing phenomenon that the areas highlighted by the detection network are exactly the fissula ante fenestram areas carefully labelled by experienced otologists. The learned lesion representation is visualized on 2D plane by using PCA to reduce its dimension. The visualization result shows that the normal and fenestral otosclerosis ear samples are well separated by the diagnostic model. Most importantly, compared to previous deep learning-based diagnosis works that only output a diagnostic result with a probability, our diagnostic model not only outputs the diagnostic result of normal or fenestral OS but also outlines the possible lesion regions in CT slices. This greatly increases its applicability in clinical practice and helps doctors understand the diagnosis basis of the model.

To understand why the otosclerosis-LNN model works, we visualized the neural responses learned by the detection network in the model. To visualize this, we performed a forward inference of the model for inputting HRCT slices, then randomly selected six feature maps outputted by the convolutional layers before the final fully connected layer. This experiment was conducted on an independent test set and the results (see *Figure 9*) demonstrated the areas focused by the detection network were the fissula ante fenestram areas carefully marked by experienced otologists (*Figure 9C,D*). This explains to a certain extent why the model could accurately diagnose lesions around the stapes footplate. In addition, after using PCA to reduce the feature dimension of the classification layer and visualizing it on a 2D plane, we observed that normal and fenestral OS ear samples were well divided (see *Figure 9B*).

Discussion

Several AI-based CAD approaches have been developed to assist doctors in recent years, such as predicting the

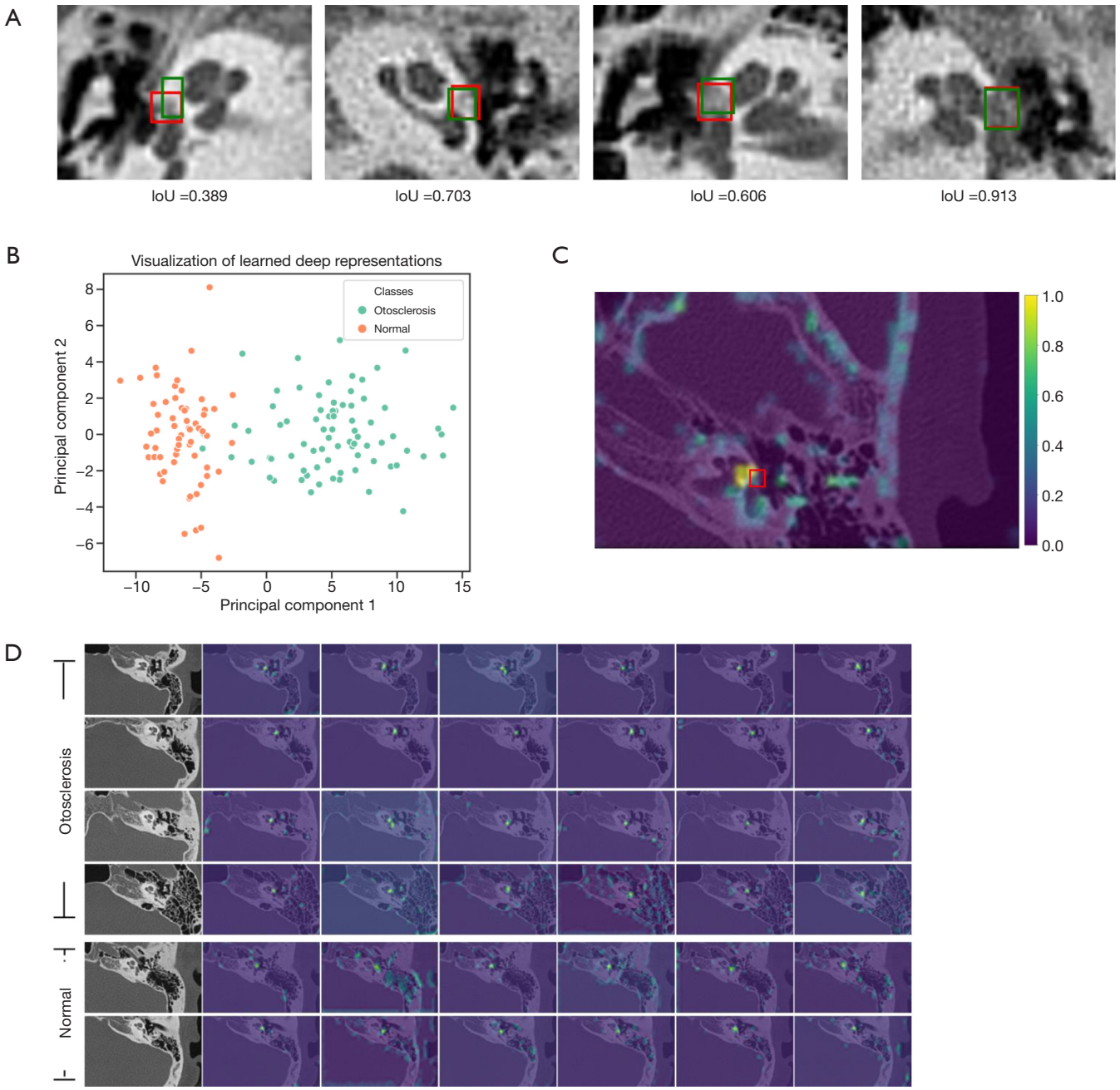


Figure 9 Visualizations of the attention areas of the detection network in our otosclerosis-LNN model. (A) Visualization of otosclerosis detection results. The green box is drawn by the clinician, and the red box is the result of artificial intelligence output. The green box and red box represent the true label and the model prediction, respectively. The IoU value at the bottom is a commonly used metric to measure the detection performance and indicates the coincidence degree between the true box and the prediction result, where higher is better. (B) Visualization of learned deep representations before the output layer. (C) Comparison of attention areas with the annotated red box. (D) Six heat maps of left ears (four abnormal and two normal) generated by the detection network of our proposed otosclerosis-LNN model. These heat maps demonstrate which areas in CT slices are focused by the detection network. IoU, Intersection over Union; LNN, Logical Neural Network.

prognosis of lung cancer patients (18,29), detecting abnormality in lower extremity radiographs (21,30), delineating all organs at risk (19), and predicting drug-protein interaction (19). While compared with these rapidly developing medical directions, the research on otology diagnosis based on AI is still in its infancy, we took a substantial step forward by proposing a novel explainable OS localization deep neural network for the detection and triage of ears based on temporal bone HRCT. The proposed otosclerosis-LNN model, involving supervised training on a collected dataset containing 134,574 CT slices from 1,331 patients, demonstrated a favorable performance for automated OS localization in three-dimensional temporal bone CT scans. The model achieved an AUC of 99.5% (per-ear-sensitivity of 96.4%, per-ear-specificity of 98.9%) on the collected test set containing 31,774 CT slices from 144 patients, which is an encouraging result in view of the high variability in input CT slices such as the diversity of patients, and the presence of various, and unlabeled other abnormalities.

This study is the first to apply deep learning technique to extract the region of interest from whole-volume high-resolution CT scans of temporal bones for diagnosing fenestral OS. Previous diagnostic works based on deep learning paid less attention to otological diseases and more to the skin, breast, lung, liver, retina, esophagus, and colon (18,19,21,29,31-38). Some diagnostic works for ear disease achieved promising results such as the diagnosis of COM based on temporal bone CT scans using deep learning (22), diagnosis of secretory otitis media with machine learning algorithm (22), diagnosis of ear diseases based on otoscope images using ensemble deep models (39), and prediction of hearing and speech perception in children with cochlear implants using AI technology (40). However, there are no studies devoted to diagnosing fenestral OS lesions, which is the most likely ear disease to be misdiagnosed and has the most obvious therapeutic effect. In addition, because the model can automatically extract all stapes footplate regions in CT scans, the potential application of the otosclerosis-LNN model can be used not only for the diagnosis of fenestral OS, but for the diagnosis of other otological diseases in the region of the stapes footplate.

We conducted a detailed analysis of the diagnosis results of otolaryngologists in the clinical diagnostic study and observed that there was a significant trend of concentrated diagnosis for some ear samples through statistics. Based on the statistical results, we further subdivided stapedial OS into focal, transitional, and typical types and found that the otosclerosis-LNN model also showed similar diagnostic results in these three OS types, evidenced by the diagnostic accuracy gradually decreasing from the typical type to the focal type. This result indirectly demonstrated the rationality of the fine-grained classification for fenestral OS. Our newly defined OS types can provide a guide for fenestral OS diagnosis in clinical practice.

In addition, our otosclerosis-LNN model employs an end-to-process strategy instead of end-to-end deep neural networks widely used in existing works. CT scans-based on end-to-end diagnostic approaches (20,23,29,36,41-43) can output desired results by directly inputting the original CT slices. Researchers only need to focus on how to design an appropriate deep network structure and define corresponding objective loss to optimize network parameters. The advantage of the end-to-end model is that all processing works can be transferred to the model and the disease diagnosis model can be trained by using a large amount of training data and time. Its disadvantage is that little human prior knowledge is incorporated into the diagnosis process. When dealing with the problem of the serious imbalance of categories, the end-to-end model needs to design a more complex network to deal with the categories with fewer samples, and the model is often difficult to converge. In contrast, the end-to-process strategy employed in our model is a combination of a heuristic image processing algorithm and deep learning technique, which allows a deep learning-based detection network to make full use of its advantages in dealing with the category imbalance problem. The superior performance of the otosclerosis-LNN model, including high sensitivity, high specificity, and low computational time, demonstrates that the end-to-process strategy is a reasonable choice for the OS localization task. It is hoped that this strategy can provide a good reference for other diagnostic work in model design, especially when dealing with extremely imbalanced categories.

Compared to previous deep learning-based diagnosis works (13,17,18,30,36,39,40,44,45), which only output a diagnostic result with a probability, our otosclerosis-LNN model not only outputs the diagnostic result of normal or fenestral OS but also outlines the possible lesion regions in CT slices. This greatly increases its applicability in clinical practice and helps doctors understand the diagnosis basis of the model. For example, for abnormal examinations, the model can provide valuable diagnostic evidence for otologists, which encourages them to quickly check the abnormal areas on the CT slices suggested by the model.

The model can quickly identify normal examinations with a high confidence score of thresholds, allowing otologists to allocate more time and energy to abnormal and complex cases.

The proposed otosclerosis-LNN model evaluated on the collected test set performs comparably or favorably compared to chief physicians (~20 years) and other otologists. The test results of the seven otologists invited into the comparison study showed that in addition to the two chief physicians, the diagnostic sensitivity and specificity of other doctors have much room for improvement. This result also reveals an important message that the diagnostic conclusions of doctors may not be as accurate as they think, especially for junior and middle-level doctors. Therefore, it is of great significance to build clinically applicable CAD systems for assisting doctors by using deep neural networks which imbue the rich clinical experience and knowledge of senior doctors. Such a diagnostic system is particularly important for areas with scarce medical resources. To demonstrate the effectiveness of the CAD system, we conducted an experiment using it to assist doctors in diagnosis. The experimental results demonstrated most otologists, including chief physicians, achieved significant improvement in terms of sensitivity, specificity, and accuracy.

Our study has some limitations. Firstly, we have evaluated the otosclerosis-LNN model on the collected test set from two different institutions, showing encouraging results. However, the performance of the model may vary in different hospitals, ways of operating, patient differences, and imaging equipment. In addition, the model outputs the binary classification results of normal or fenestral OS based on the automatically detected fissula ante fenestram space in three-dimensional HRCT images and further studies are needed to diagnose other diseases in the stapes footplate region. Finally, otologists often combine various patient information such as CT scans, clinical symptoms, medical records, and audiological examination in clinical diagnosis. Therefore, a CAD model based on deep learning can further consider combining multi-modal information and learn how doctors synthesize this to make a final diagnosis. Despite these limitations, the achieved diagnostic performance of our otosclerosis-LNN model is encouraging in that it is either comparable or better than that of senior otologists. Furthermore, otologists can significantly improve their diagnostic level when assisted by the otosclerosis-LNN model. The diagnosis of ear diseases based on deep learning techniques is rare, and far behind other fields. It is now time to advance this field and design AI diagnosis models to help otologists perform routine clinical diagnosis, thus effectively reducing their diagnostic burden and avoiding "harmful" misdiagnosis.

In conclusion, we have presented a deep learning-based OS diagnostic model that can automatically localize OS lesion regions in three-dimensional HRCT scans. Though, our artificial intelligence technology has not been used in clinic for the time being, it shows superiority compared with the test results of doctors and plays an auxiliary role in existing cases. With further verification by randomized controlled trials, the model may eventually enter into clinical application and be used to assist otologists in automatically and quickly diagnosing fenestral OS.

Acknowledgments

Funding: This work was supported by the National Natural Science Foundation of China (No. 61772137, 61902076, 81670281, 81570919), the big data and artificial intelligence project (2020DSJ07), the Shanghai Municipal Commission of Science and Technology Research Project (18140900304, 19140900902), and the clinical research and transformation incubation project (SZA2020004).

Footnote

Reporting Checklist: The authors have completed the STARD reporting checklist. Available at http://dx.doi.org/10.21037/atm-21-1171

Data Sharing Statement: Available at http://dx.doi.org/10.21037/atm-21-1171

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at http://dx.doi.org/10.21037/atm-21-1171). Dr. BY reports funding from the big data and artificial intelligence project (2020DSJ07), the Shanghai Municipal Commission of Science and Technology Research Project (18140900304, 19140900902), and the clinical research and transformation incubation project (SZA2020004). The other authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was

conducted in accordance with the Declaration of Helsinki (as revised in 2013). The study was approved by institutional committee of Eye and ENT Hospital, Fudan University (No. 2020005) and informed consent was taken from all the patients.

References

1. Deafness and hearing loss: World Health Organization; 2020, updated March 1. Available online: https://www.who.int/zh/news-room/fact-sheets/detail/deafness-and-hearing-loss
2. Altmann F, Glasgold A, Macduff JP. The Incidence of Otosclerosis as Related to Race and Sex. Ann Otol Rhinol Laryngol 1967;76:377-92.
3. Morrison AW. Genetic factors in otosclerosis. Ann R Coll Surg Engl 1967;41:202.
4. Declau F, Van Spaendonck M, Timmermans J, et al. Prevalence of histologic otosclerosis: an unbiased temporal bone study in Caucasians. Otosclerosis and Stapes Surgery. 65: Karger Publishers; 2007:6-16.
5. Ealy M, Smith RJH. Otosclerosis. Adv Otorhinolaryngol 2011;70:122-9.
6. Ealy M, Smith RJ. The genetics of otosclerosis. Hear Res 2010;266:70-4.
7. Lee TC, Aviv R, Chen J, et al. CT grading of otosclerosis. Am J Neuroradiol 2009;30:1435-9.
8. Stankovic KM, McKenna MJ. Current research in otosclerosis. Curr Opin Otolaryngol Head Neck Surg 2006;14:347-51.
9. Chole RA, McKenna M. Pathophysiology of otosclerosis. Otol Neurotol 2001;22:249-57.
10. Vicente AO, Yamashita HK, Albernaz PL, et al. Computed tomography in the diagnosis of otosclerosis. Otolaryngol Head Neck Surg 2006;134:685-92.
11. Cureoglu S, Baylan MY, Paparella MM. Cochlear otosclerosis. Curr Opin Otolaryngol Head Neck Surg 2010;18:357.
12. Lagleyre S, Sorrentino T, Calmels MN, et al. Reliability of high-resolution CT scan in diagnosis of otosclerosis. Otol Neurotol 2009;30:1152-9.
13. Naumann IC, Porcellini B, Fisch U. Otosclerosis: incidence of positive findings on high-resolution computed tomography and their correlation to audiological test data. Ann Otol Rhinol Laryngol 2005;114:709-16.
14. Huang TS. editor. A retrospective study of 37 cases. The Function and Mechanics of Normal, Diseased and Reconstructed Middle Ears: Proceedings of the Second International Symposium on Middle-Ear Mechanics in Research and Otosurgery, Held in Boston, MA, USA, October 21st-24th, 1999; 2000: Kugler Publications.
15. Rusk N. Deep learning. Nat Method 2016;13:35.
16. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436-44.
17. Mei X, Lee HC, Diao KY, et al. Artificial intelligence-enabled rapid diagnosis of patients with COVID-19. Nat Med 2020;26:1224-8.
18. Mukherjee P, Zhou M, Lee E, et al. A Shallow Convolutional Neural Network Predicts Prognosis of Lung Cancer Patients in Multi-Institutional CT-Image Data. Nat Mach Intell 2020;2:274-82.
19. Tang H, Chen X, Liu Y, et al. Clinically applicable deep learning framework for organs at risk delineation in CT images. Nat Mach Intell 2019;1:480-91.
20. Shan H, Padole A, Homayounieh F, et al. Competitive performance of a modularized deep neural network compared to commercial algorithms for low-dose CT image reconstruction. Nat Mach Intell 2019;1:269-76.
21. Hwang EJ, Park S, Jin KN, et al. Development and validation of a deep learning-based automated detection algorithm for major thoracic diseases on chest radiographs. JAMA Netw Open 2019;2:e191095.
22. Wang YM, Li Y, Cheng YS, et al. Deep learning in automated region proposal and diagnosis of chronic otitis media based on computed tomography. Ear Hear 2020;41:669-77.
23. Ren S, He K, Girshick R, et al. Faster r-CNN: Towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell 2017;39:1137-49.
24. Kim J, Kwon Lee J, Mu Lee K. editors. Accurate image super-resolution using very deep convolutional networks. Proceedings of the IEEE conference on computer vision and pattern recognition; 2016.
25. Lin TY, Goyal P, Girshick R, et al. editors. Focal loss for dense object detection. Proceedings of the IEEE international conference on computer vision; 2017.

26. Kingma DP, Ba J. Adam: A method for stochastic optimization. International Conference on Learning Representations, 2015.
27. Chaudhary A, Chouhan KS, Gajrani J, et al. Deep Learning With PyTorch. Machine Learning and Deep Learning in Real-Time Applications: IGI Global; 2020:61-95.
28. Buda M, Maki A, Mazurowski MA. A systematic study of the class imbalance problem in convolutional neural networks. Neural Netw 2018;106:249-59.
29. Ardila D, Kiraly AP, Bharadwaj S, et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nat Med 2019;25:954-61.
30. Varma M, Lu M, Gardner R, et al. Automated abnormality detection in lower extremity radiographs using deep learning. Nat Mach Intell 2019;1:578-83.
31. Chen PC, Gadepalli K, MacDonald R, et al. An augmented reality microscope with real-time artificial intelligence integration for cancer diagnosis. Nat Med 2019;25:1453-7.
32. De Fauw J, Ledsam JR, Romera-Paredes B, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. Nat Med 2018;24:1342-50.
33. Kather JN, Pearson AT, Halama N, et al. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. Nat Med 2019;25:1054-6.
34. Skrede OJ, De Raedt S, Kleppe A, et al. Deep learning for prediction of colorectal cancer outcome: a discovery and validation study. Lancet 2020;395:350-60.
35. Wu N, Phang J, Park J, et al. Deep neural networks improve radiologists' performance in breast cancer screening. IEEE Trans Med Imaging 2020;39:1184-94.
36. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542:115-8.
37. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 2016;316:2402-10.
38. Shen L, Zhao W, Xing L. Patient-specific reconstruction of volumetric computed tomography images from a single projection view via deep learning. Nat Biomed Eng 2019;3:880-8.
39. Cha D, Pae C, Seong SB, et al. Automated diagnosis of ear disease using ensemble deep learning with a big otoendoscopy image database. EBioMedicine 2019;45:606-14.
40. Papsin BC, Gordon KA. Cochlear implants for children with severe-to-profound hearing loss. New Eng J Med 2007;357:2380-7.
41. Zheng S, Li Y, Chen S, et al. Predicting drug–protein interaction using quasi-visual question answering system. Nat Mach Intell 2020;2:134-40.
42. Zhao Y, Chen Y, Bindel D. Towards unbiased end-to-end network diagnosis. IEEE/ACM Transactions on Networking 2009;17:1724-37.
43. Silver D, Hasselt H, Hessel M, et al. editors. The predictron: End-to-end learning and planning. International Conference on Machine Learning; 2017: PMLR.
44. Zheng X, Yao Z, Huang Y, et al. Deep learning radiomics can predict axillary lymph node status in early-stage breast cancer. Nat Commun 2020;11:1236.
45. Quesnel AM, Moonis G, Appel J, et al. Correlation of computed tomography with histopathology in otosclerosis. Otol Neurotol 2013;34:22.

(English Language Editor: B. Draper)

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

Supplementary

Table S1 Basic characteristics of enrolled participants in the training set

Dataset	Male	Female	Total
Training Set (discharge diagnosed as otosclerosis)			
Number of patients(retrieved)	298	696	994
Age (mean ± SD)	39.11±10.97	41.49±10.62	40.76±10.78
Number of CT scans	238	556	794
Labels assigned to ears (stapedial)	288	702	990
Total number of CT slices	-	-	101,446
Cropped ear slices	-	-	2,458
Training Set (discharge diagnosed as external auditory canal tumor/new organism)			
Number of patients(chosen)	151	149	300
Age (mean ± SD)	48.89±19.82	44.95±18.52	46.93±19.29
Number of CT scans	151	149	300
Labels assigned to ears	173	168	341
Total number of CT slices	-	-	33,128
Cropped ear slices	-	-	994

Table S2 Basic characteristics of enrolled participants in the retrospective clinical test set

Dataset	Male	Female	Total
Testing Set (bilateral stapedial otosclerosis)			
Number of patients	13	29	42
Age (mean ± SD)	32±14.31	40.71±10.42	38.14±12.36
Number of CT scans	13	29	42
Labels assigned to ears	26	58	84
Total number of CT slices	-	-	9,160
Testing Set (unilateral stapedial otosclerosis)			
Number of patients	0	2	2
Number of CT scans	0	2	2
Labels assigned to ears	0	4	4
Testing Set (bilateral normal)			
Number of patients	37	63	100
Age (mean ± SD)	39.11±14.37	48.71±15.58	44.98±15.91
Number of CT scans	37	63	100
Labels assigned to ears	74	126	200
Total number of CT slices	-	-	22614

http://dx.doi.org/10.21037/atm-21-1171

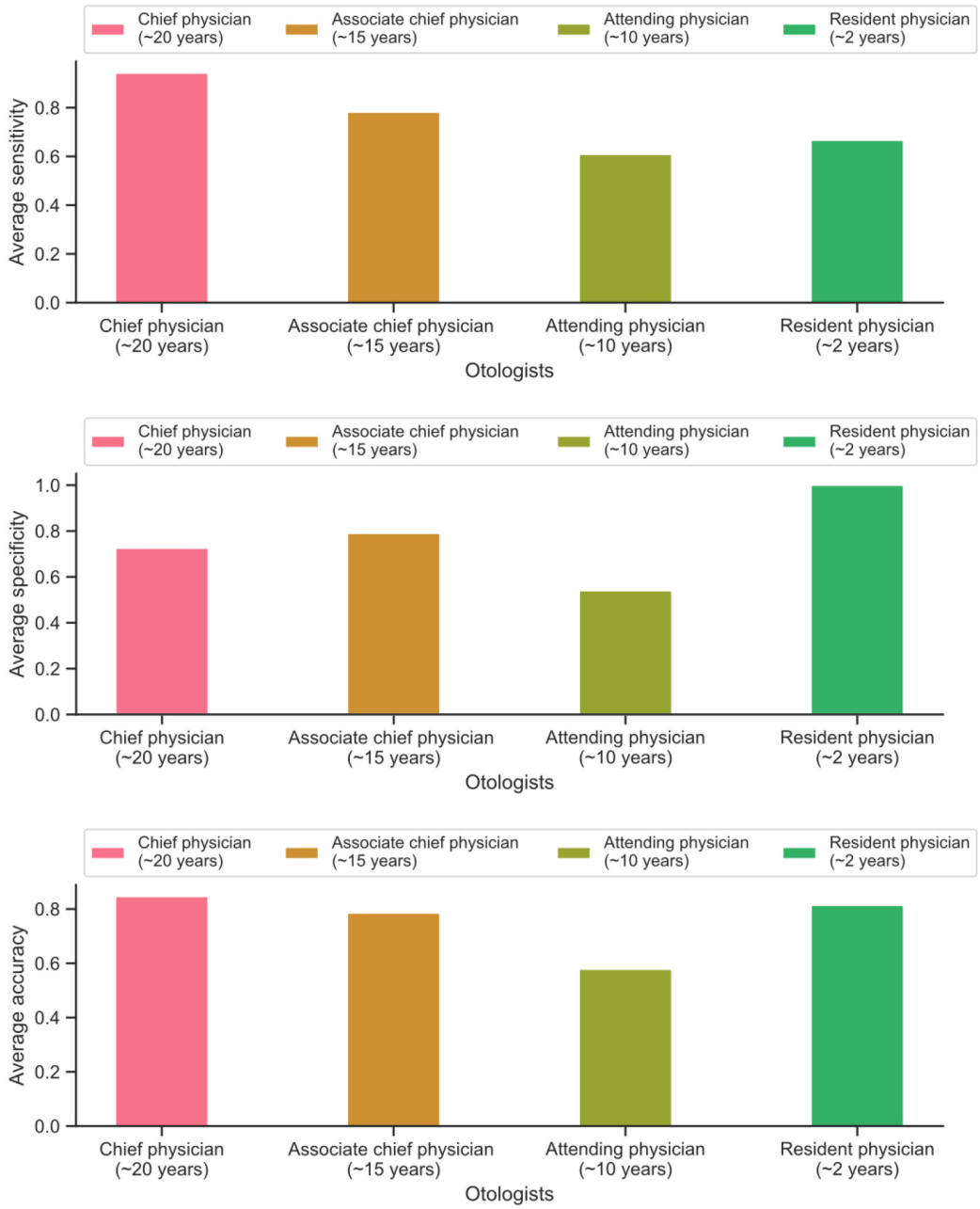


Figure S1 Demonstration of the study results of the ability of otolaryngologists to clinically diagnose otosclerosis. We showed the comparison of average diagnostic performance between chief physician, associate chief physician, attending physician, and resident physician in terms of sensitivity, specificity, and accuracy.

http://dx.doi.org/10.21037/atm-21-1171

Table S3 Comparison of our otosclerosis-LNN model with seven otologists in terms of sensitivity and specificity on the prospectively collected clinical test set, which contains a total of 140 ears (otosclerosis =78, normal =62)

Otologist	Diagnosis time (seconds per ear)	Total number of ears, n=140 (otosclerosis =78, normal =62)					
		True positives	False negatives	True negatives	False positives	Sensitivity	Specificity
Otosclerosis-LNN	0.06	75	3	62	0	96.15%	100%
Chief physician A (~20 years)	34.3	76	2	28	34	97.44%	45%
Chief physician B (~20 years)	32.2	71	7	62	0	91.03%	100%
Associate chief physician A (~15 years)	30	50	28	57	5	64.10%	92%
Associate chief physician B (~15 years)	42.9	72	6	41	21	92.31%	66%
Associate chief physician C (~10 years)	77.1	42	36	28	34	53.85%	45%
Attending physician (~10 years)	20	53	25	39	23	67.95%	63%
Resident (~2 years)	30	52	26	62	0	66.67%	100%

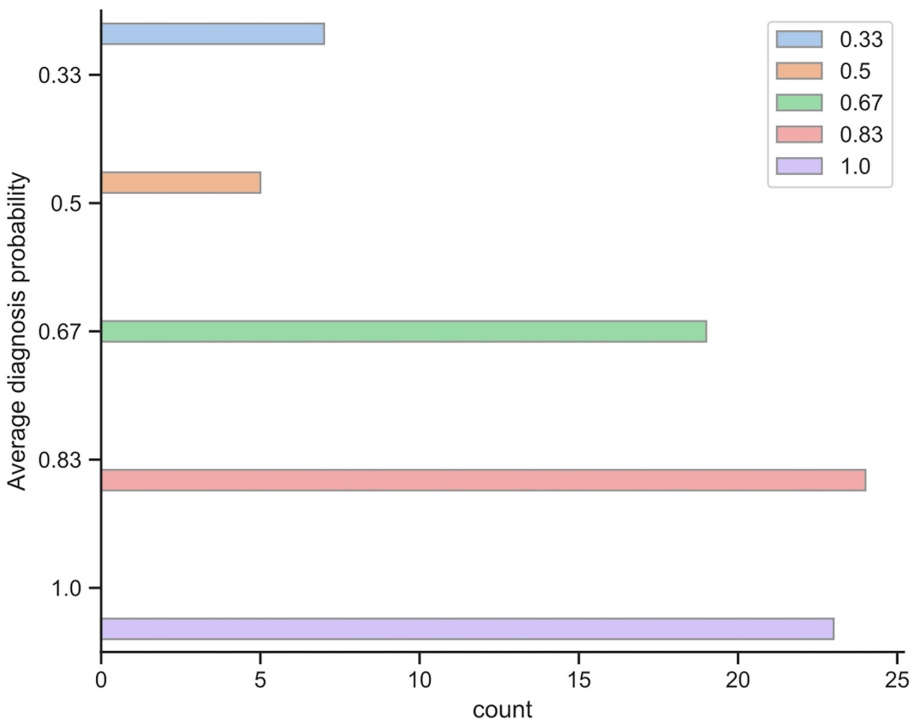


Figure S2 Histogram of the average diagnosis accuracy of otolaryngologists of each ear in the clinical test set.

http://dx.doi.org/10.21037/atm-21-1171

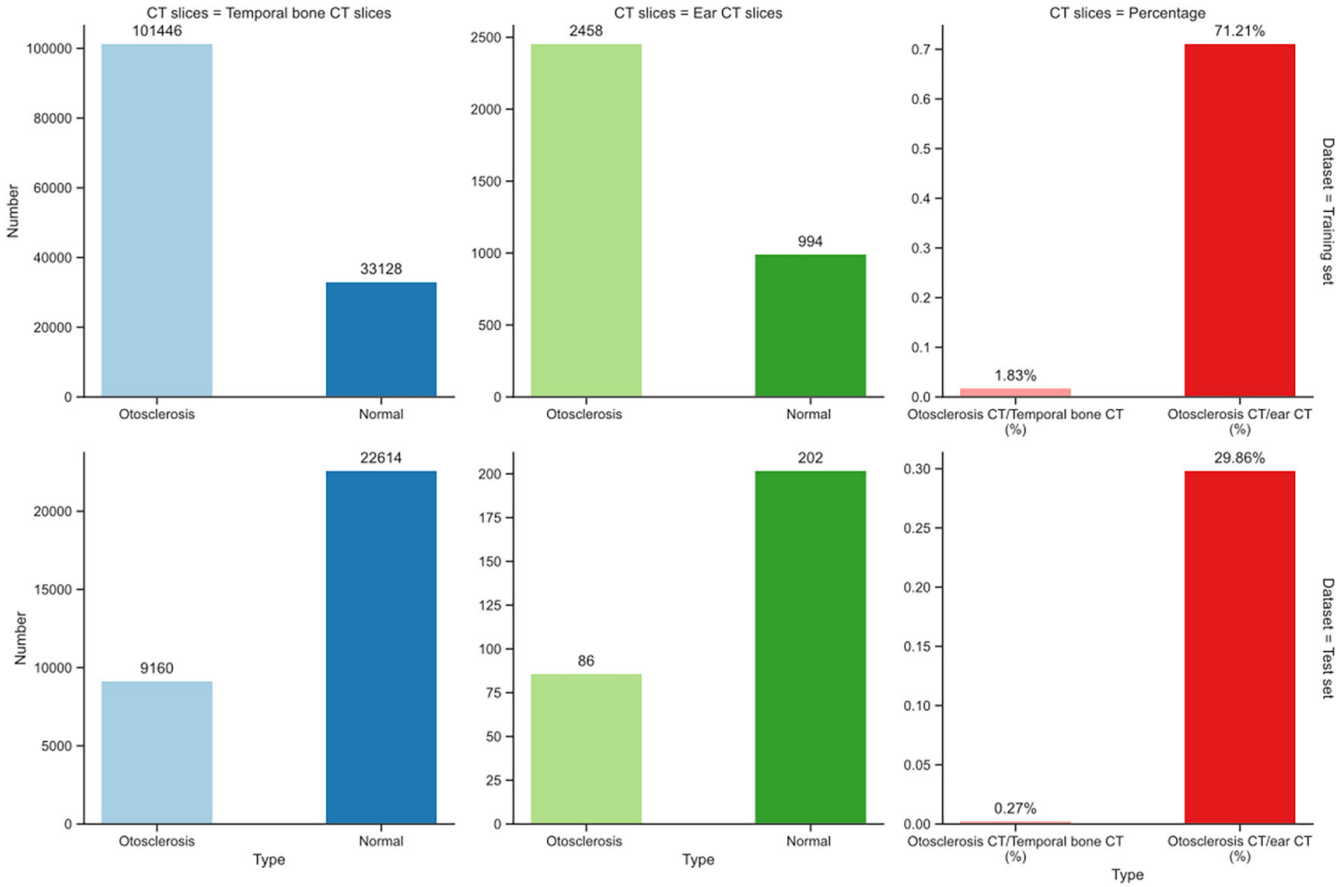


Figure S3 Demonstration of the number of the temporal bone CT slices and ear CT slices in the training and test sets. The rightmost figure shows the ratio of the number of CT slices with otosclerosis to the total number of temporal bone CT slices and the number of ear CT slices, respectively and the proportions are 1.83% and 71.21%, respectively. Therefore, our otosclerosis-LNN model employs an end-to-process strategy instead of end-to-end deep neural networks widely used in existing works to train the deep learning-based otosclerosis detection network. The end-to-process strategy allows us to train the deep learning model using ear CT slices instead of whole temporal bone CT slices, which avoids the extreme class imbalance problem.

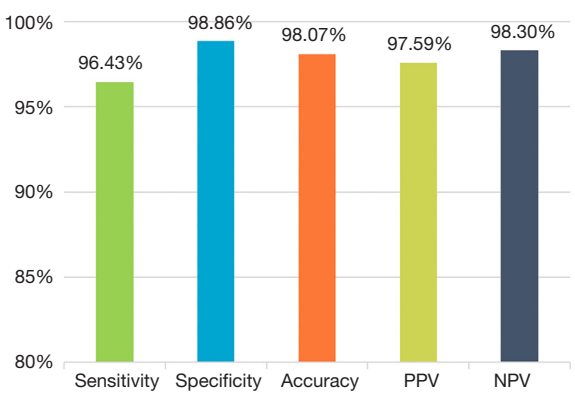


Figure S4 Demonstration of the diagnostic performance of our otosclerosis-LNN model on the external test set.

http://dx.doi.org/10.21037/atm-21-1171

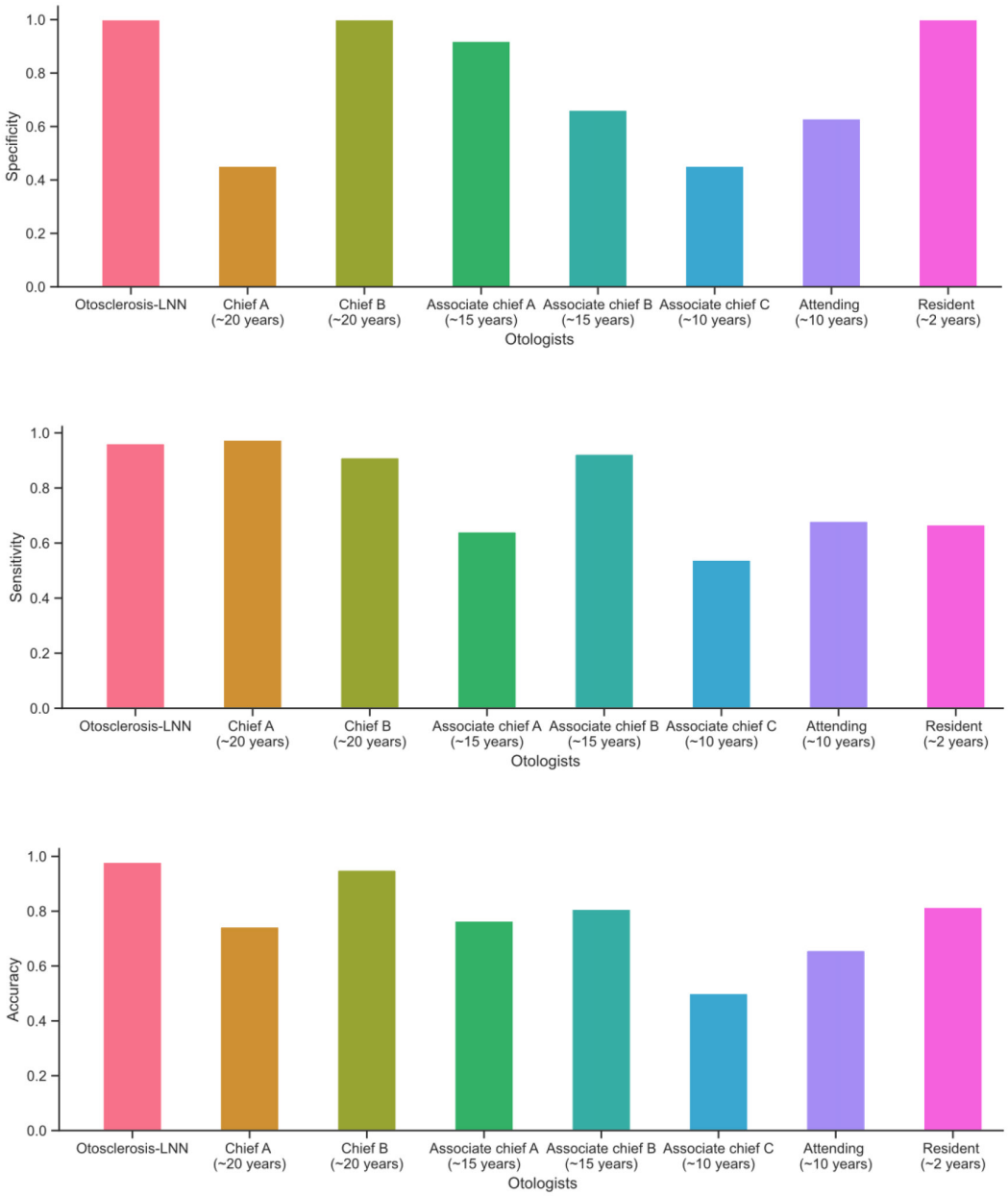


Figure S5 Comparison of our otosclerosis-LNN model with otologists. The otosclerosis-LNN model demonstrated higher screening performance than both two chief physicians (~20 years) enrolled in the comparison study on the prospectively collected test set. The otosclerosis-LNN model achieved sensitivity and specificity of 96.15% and 100%, respectively (Table S3), and the average sensitivity and specificity of the two chief physicians (~20 years) attending the comparison study are 94.20% and 72.55%, respectively. These two important indicators are lower than the otosclerosis-LNN model. In addition, two associate chief physicians achieved average sensitivity and specificity of 70.07% and 67.70%, respectively, which is lower than that of two chief physicians. The attending physician (~10 years) achieved sensitivity and specificity of 67.90% and 62.90%, respectively and the resident (~2 years) achieved sensitivity and specificity of 66.70% and 100%, respectively. Overall, compared with other doctors, the chief physicians demonstrated a higher diagnostic level in the test set, but still lower than the otosclerosis-LNN model.

http://dx.doi.org/10.21037/atm-21-1171

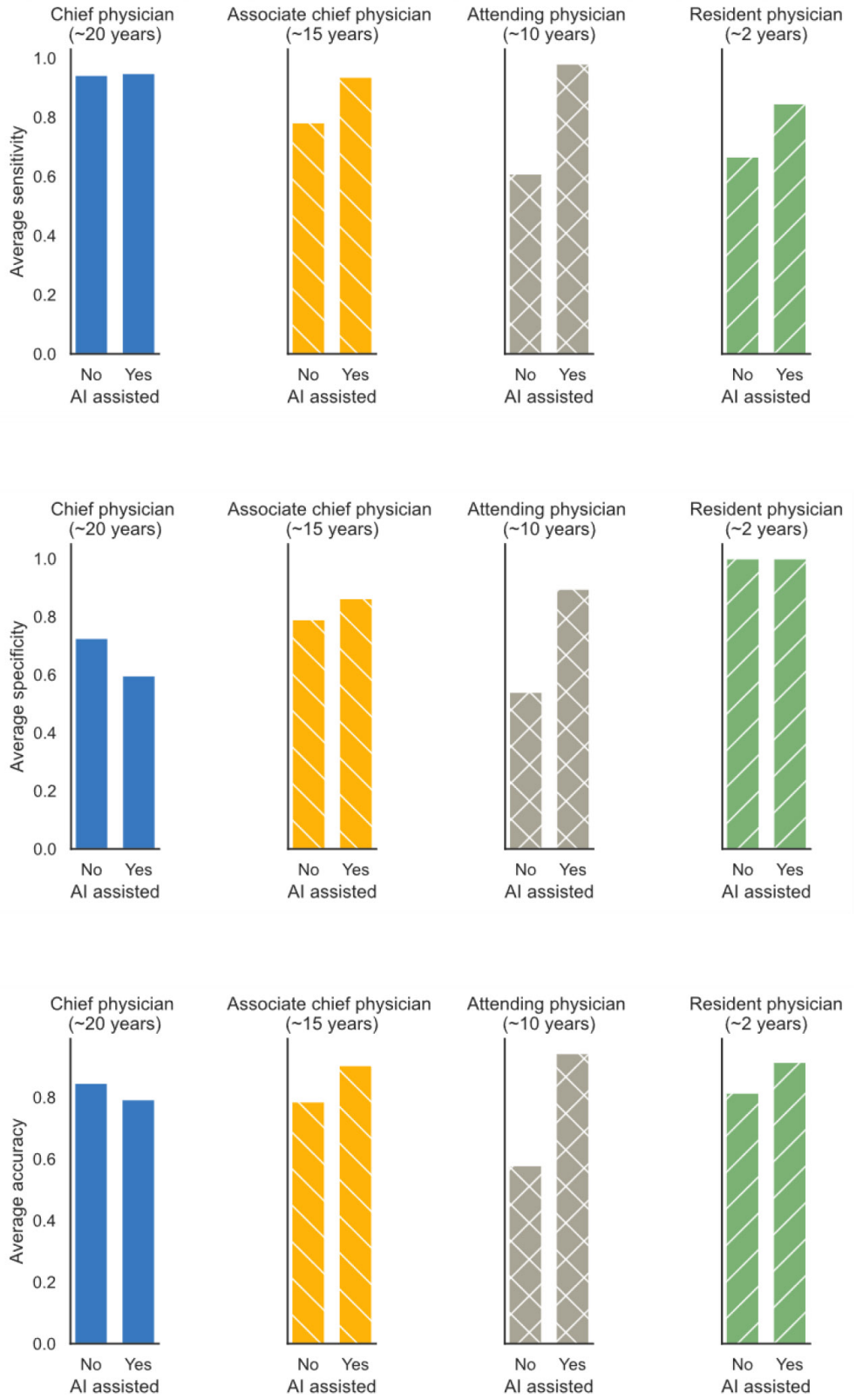


Figure S6 Demonstration of the diagnostic performance of otologists with the assistance of our otosclerosis-LNN model.

http://dx.doi.org/10.21037/atm-21-1171

Table S4 Using the otosclerosis-LNN model to assist otolaryngologists in the diagnosis of stapedial otosclerosis in the temporal bone high-resolution CT images

Otologists	Total number of ears, n=140 (otosclerosis =78, normal =62)					
	True positives	False negatives	True negatives	False positives	Sensitivity	Specificity
Otosclerosis-LNN	75	3	62	0	96.15%	100.00%
Chief physician A (~20 years)	78	0	12	50	100.00%	19.35%
Chief physician B (~20 years)	70	8	62	0	89.74%	100.00%
Associate chief physician A (~15 years)	70	8	58	4	89.74%	93.55%
Associate chief physician B (~15 years)	76	2	49	13	97.44%	79.03%
Associate chief physician C (~10 years)	76	2	55	7	97.44%	88.71%
Attending physician (~10 years)	77	1	56	6	98.72%	90.32%
Resident (~2 years)	66	12	62	0	84.62%	100.00%

http://dx.doi.org/10.21037/atm-21-1171