Background: The integration of artificial intelligence (AI), particularly deep learning models, has transformed the landscape of medical technology, especially in the field of diagnosis using imaging and physiological data. In otolaryngology, AI has shown promise in image classification for middle ear diseases. However, existing models often lack patient-specific data and clinical context, limiting their universal applicability. The emergence of GPT-4 Vision (GPT-4V) has enabled a multimodal diagnostic approach, integrating language processing with image analysis.

Objective: In this study, we investigated the effectiveness of GPT-4V in diagnosing middle ear diseases by integrating patient-specific data with otoscopic images of the tympanic membrane.

Methods: The design of this study was divided into two phases: (1) establishing a model with appropriate prompts and (2) validating the ability of the optimal prompt model to classify images. In total, 305 otoscopic images of 4 middle ear diseases (acute otitis media, middle ear cholesteatoma, chronic otitis media, and otitis media with effusion) were obtained from patients who visited Shinshu University or Jichi Medical University between April 2010 and December 2023. The optimized GPT-4V settings were established using prompts and patients’ data, and the model created with the optimal prompt was used to verify the diagnostic accuracy of GPT-4V on 190 images. To compare the diagnostic accuracy of GPT-4V with that of physicians, 30 clinicians completed a web-based questionnaire consisting of 190 images.

Results: The multimodal AI approach achieved an accuracy of 82.1%, which is superior to that of certified pediatricians at 70.6%, but trailing behind that of otolaryngologists at more than 95%. The model’s disease-specific accuracy rates were 89.2% for acute otitis media, 76.5% for chronic otitis media, 79.3% for middle ear cholesteatoma, and 85.7% for otitis media with effusion, which highlights the need for disease-specific optimization. Comparisons with physicians revealed promising results, suggesting the potential of GPT-4V to augment clinical decision-making.

Conclusions: Despite its advantages, challenges such as data privacy and ethical considerations must be addressed. Overall, this study underscores the potential of multimodal AI for enhancing diagnostic accuracy and improving patient care in otolaryngology. Further research is warranted to optimize and validate this approach in diverse clinical settings.

(JMIR AI 2024;3:e58342) doi: 10.2196/58342
KEYWORDS
artificial intelligence; deep learning; machine learning; generative AI; generative; tympanic membrane; middle ear disease; GPT4-Vision; otolaryngology; ears; ear; tympanic; vision; GPT; GPT4V; otoscopic; image; images; imaging; diagnosis; diagnoses; diagnostic; diagnostics; otitis; mobile phone

Introduction
The emergence of artificial intelligence (AI) has altered the landscape of medical technology, particularly in diagnosis, which leverages the identification of features based on imaging and physiological data [1-3]. In the field of otolaryngology, AI and deep learning models are being used for imaging; ongoing efforts focus on classifying diseases based on tympanic membrane images of middle ear disease [4-6]. Technological advancements, including deep learning and transfer learning using pretrained models, have resulted in an accuracy range of 70%-90% in models for analyzing otoscopic images [7]. There have also been advancements in its application, such as implementing smartphone-based point-of-care diagnostics [8]. However, these models rely on trained image data, require large image data sets, and do not consider patient information or clinical context. Consequently, the universality of these models is limited, and their optimal application in clinical practice remains unclear.

Recently, large-scale language-processing models have become available for general use. Further, 1 such model, the GPT-4, has demonstrated specialist-level medical knowledge through its language-processing abilities [9-11]. Since October 2023, GPT-4 Vision (GPT-4V) has gained the ability to evaluate image data, enabling a multimodal diagnostic approach that incorporates both language processing and image analysis [12]. GPT-4V enables the integration of patient information analysis and image-based deep learning models, providing valuable support in diagnosis and treatment, similar to decisions made in a clinical setting [13]. Multimodal AI, which bases diagnosis on multiple pieces of information, has been reported to be more effective than methods that rely on a single type of information. This is demonstrated in various medical applications, including the combination of pathology images with genomic information [14] and their use in liver cancer [15] and cervical cancer [16], where imaging information is integrated. In otorhinolaryngology, there have been few reports; however, efforts to incorporate AI for otoscopic images could further improve the quality of care.

In this study, we aimed to investigate the effectiveness of a multimodal approach using GPT-4V to diagnose middle ear disease. This approach was designed to integrate patient-specific data (age, sex, and chief complaint) with tympanic membrane images to assess the accuracy of the versatile GPT-4V. The model’s accuracy was compared with physicians’ diagnoses to validate its effectiveness in image-based deep learning. The potential future development of the multimodal AI approach for classifying middle ear diseases is also discussed.

Methods
Study Design
GPT-4V has been available as an image recognition model since September 25, 2023. This study’s design was divided into two phases: (1) establishing a model with appropriate prompts and (2) validating the ability of the optimal prompt model to classify images (Figure 1).

Figure 1. Overview of this study. The model was divided into two phases: (1) establishment and (2) tuned model validation.TM: tympanic membrane.
Correct Otoscopic Images and Patient Information
This study included 305 otoscopic images of middle ear disease obtained from patients who visited Shinshu University or Jichi Medical University between April 2010 and December 2023. The endoscope used was an Olympus ENF-VH and ENF-V3 (Olympus), and the video system was an Olympus VISERA ELITE OTV-S190. Further, 1 image was obtained from each patient. We excluded images with poor quality and those in which multiple diseases were suspected. The remaining images were classified into 4 disease categories: acute otitis media (AOM), middle ear cholesteatoma (chole), chronic otitis media (COM), and otitis media with effusion (OME). The final diagnoses were based on the judgment of the otolaryngologists who treated the patients. These images were accompanied by patient-specific information, such as age, sex, and chief complaint (eg, fever, otalgia, otorrhea, ear fullness, deafness, facial palsy, dizziness, and tinnitus). We excluded images taken after otologic surgery. Of note, only 1 image was obtained from each patient.

GPT-4V Settings and Prompt Tuning
The GPT-4V settings were established using prompts reported in previous studies [17,18]. Briefly, conditions and prompts for providing answers were verified using 10 images for each disease. According to a report on prompts [19], image data or patient information were manually input into GPT-4V, and the generated results were evaluated by the physicians (MN and HY).

Accuracy Verification of GPT-4V Using the Optimal Prompt Model
The model with the optimal prompt created was used to verify the diagnostic accuracy of GPT-4V on 190 images (37 in AOM, 53 in chole [6 in congenital, 47 in acquired], 51 in COM, and 49 in OME), which were different from those for tuning prompts. To account for the variability in responses, each administration was performed 3 times, and responses that were answered 2 or more times were considered to be the actual response.

Comparison of AI Accuracy With Physician Accuracy
To compare the diagnostic accuracy of GPT-4V with that of physicians, 30 clinicians completed a web-based questionnaire consisting of 190 images. The web-based survey included tympanic membrane images and patient information (age, sex, and chief complaint) in a 4-choice question format. The respondents included 8 certificated pediatricians, 8 otolaryngology residents, 8 certificated otolaryngologists, and 6 experts in otolaryngology (more than 15 years of experience).

To show the trend in the percentage of correct responses according to the difficulty of the questions, the questions were divided into 3 levels (easy, normal, and hard) according to the overall percentage of correct responses by physicians, and the percentage of correct responses for each level and each question was compared between the GPT-4V and all doctors, otolaryngologists, and pediatricians.

Ethical Considerations
Patient information was anonymized to protect privacy and used only with the approval of the Ethics Committee of the Shinshu University School of Medicine (6088).

Statistical Analysis
Groups were compared by 1-way ANOVA. Subsequently, multiple comparison tests (the Bonferroni method) were used to compare groups. Statistical significance was set at P<.05. A 1-sample proportion test was used to compare the performance of the physician with that of GPT-4V in terms of the correct response rate.

Results
Establishment of Optimal Prompts
In the initial stage, we sought an optimal input method using 10 images for each disease (AOM, chole, COM, or OME; 40 images total). First, we input only images or options; GPT mostly requires clinical information, such as patient history and symptoms, although no response regarding the disease was generated (Figure 1 and Multimedia Appendix 1). Second, the names of the 4 diseases were added as candidate answers, but again, no response regarding the disease was generated. When detailed patient information, such as age, sex, and main symptoms, was inputted, GPT-4V provided answers, indicating that input images with patient data were the optimal prompt for testing the accuracy of GPT-4V.

Accuracy Validation of the Multimodal AI Approach
The performance of the multimodal AI approach in this study for classifying middle ear diseases was validated, with an overall diagnostic accuracy of 82.1% for the GPT-4V-based analysis. Disease-specific accuracy rates were 89.2% for AOM (true positives [TP]=33, false positives [FP]=1, false negatives [FN]=4, precision=0.97, recall=0.89, F1-score=0.93), 76.5% for COM (TP=39, FP=7, FN=12, precision=0.85, recall=0.76, F1-score=0.8), 79.3% for cholesteatoma (TP=42, FP=13, FN=11, precision=0.76, recall=0.79, F1-score=0.78), and 85.7% for OME (TP=42, FP=10, FN=7, precision=0.81, recall=0.86, F1-score=0.83; Figure 2).

These results indicate high discrimination among various disease types; however, there were also some incorrect responses. Representative images of correct and incorrect GPT-4V classifications for each disease are shown in Figure 3.
Comparison of Diagnostic Accuracy by Physicians and GPT-4V

The same images with patients’ information used by GPT-4V were evaluated by pediatricians (n=8), otolaryngology residents (n=8), certificated otolaryngologists (n=8), and experts in otolaryngology (n=6), and the diagnostic accuracy of each group was compared. The mean diagnostic accuracy was 70.6% (SE 4.2%) for pediatricians, 95.5% (SE 1%) for otolaryngology residents, 97.3% (SE 0.8%) for certificated otolaryngologists, and 98.2% (SE 0.4%) for experts in otolaryngology. ANOVA revealed significant differences among the 4 groups (F=13.43, P<.001). In the post hoc comparison, a significant difference was observed between pediatricians and the other 3 groups (P<.001). The GPT-4V correct response rate was 82.1%, surpassing that of pediatricians by 11.5% and trailing behind otolaryngologists by an average of just over 10% (Figure 4).

The accuracy rates for specific diseases were as follows: 92.3% for AOM (pediatricians 80.4%, otolaryngology residents 94.9%, certificated otolaryngologists 97%, and experts in otolaryngology 98.2%), 95.9% for COM (pediatricians 89.5%, otolaryngology residents 94.9%, certificated otolaryngologists 97%, and experts in otolaryngology 98.2%).
otolaryngology residents 96.6%, certificated otolaryngologists 99.8%, and experts in otolaryngology 98.4%), 81.8% for chole (pediatricians 46%, otolaryngology residents 93.2%, certificated otolaryngologists 93.6%, and experts in otolaryngology 98.4%), and 93.7% for OME (pediatricians 81.6%, otolaryngology residents 97.2%, certificated otolaryngologists 99%, and experts in otolaryngology 98%).

In the confusion matrix of all doctors, there was a notable tendency to misclassify chole as OME and AOM as OME. Among pediatricians, there were more errors in classifying chole as AOM or COM (Figure 5).

Figure 4. Result of human validations with doctors of TM images and patients’ data. The graph shows the average correct rate for doctors (pediatricians, otolaryngology residents, certificated otolaryngologists, and experts in otolaryngology), and the dotted line shows the correct answer rate of GPT-4V. GPT-4V: GPT-4 Vision; TM: tympanic membrane. **P value <.01.

Figure 5. Confusion matrix of doctors (pediatricians, otolaryngology residents, certificated otolaryngologists, and experts in otolaryngology) for classifying 4 middle ear diseases. (A) Confusion matrix of all doctors (N=30). The average (percentage of total responses) is shown. (B) Confusion matrix of doctors in each group: pediatricians (n=8), otolaryngology residents (n=8), certificated otolaryngologists (n=8), and experts in otolaryngology (n=6). The averages of each group (percentage of total responses) are shown. AOM: acute otitis media; chole: cholesteatoma; COM: chronic otitis media; OME: otitis media with effusion.

Regarding the difference in the trend of the percentage of correct answers between GPT-4V and physicians according to the difficulty of the questions, even the percentage of correct answers for GPT-4V tended to decrease gradually from 85.7% for easy, 84% for normal, and 71.1% for hard questions (Table 1).

Furthermore, compared with otolaryngologists, GPT-4V had a significantly lower percentage of correct answers for all
questions (99.7% for easy, 97.1% for normal, and 90.8% for hard questions; all \( P<.001 \)). In contrast, the results of the “hard” and “normal” groups were similar. Compared with pediatricians, the GPT-4V outperformed the pediatricians in easy questions with 96.6%, although no statistically significant difference was observed \( (P=.006) \). However, the GPT-4V had a predominantly higher percentage of correct answers for normal (76.3%, \( P=.07 \)) and hard questions (45.4%, \( P<.001 \)).

**Table 1.** Comparison of the scores by GPT-4 Vision (GPT-4V) and human validation with physicians across various difficulty levels \( (N=190) \).

<table>
<thead>
<tr>
<th>Difficulty level</th>
<th>Questions, n (%)</th>
<th>GPT-4V (mean %)</th>
<th>All doctors (mean %)</th>
<th>Otolaryngologists</th>
<th>Pediatricians</th>
</tr>
</thead>
<tbody>
<tr>
<td>Easy ( (&gt;95%) )</td>
<td>77 (40.5)</td>
<td>85.7</td>
<td>97.8 ( (97.4-98.2) )</td>
<td>12.1</td>
<td>&lt;.001 ( a )</td>
</tr>
<tr>
<td>Normal ( (&gt;85%, &lt;95%) )</td>
<td>75 (39.5)</td>
<td>84</td>
<td>90.4 ( (89.7-91.0) )</td>
<td>6.4</td>
<td>.13</td>
</tr>
<tr>
<td>Hard ( (&lt;85%) )</td>
<td>38 (20)</td>
<td>71.1</td>
<td>76.8 ( (73.7-79.8) )</td>
<td>5.7</td>
<td>.44</td>
</tr>
</tbody>
</table>

\( a \)Statistically significant.

**Discussion**

**Principal Results**

In this study, we assessed the accuracy of the GPT-4V multimodal AI approach in classifying middle ear disorders, yielding the following three key findings. First, GPT-4V, a general-purpose model focusing on large-scale language models, achieved approximately 80% accuracy in classifying middle ear disease. The model’s performance, evaluated using images and patient data, was superior to that of nonotolaryngologists, although it was lower than the average accuracy of otolaryngologists. Second, the GPT-4V was able to classify diseases when patient information and disease options were input. Further improvements in accuracy could be achieved with more detailed patient information. Third, accuracy varied by disease, suggesting the potential for optimizing AI usage and improving accuracy by understanding the specificity of GPT-4V in classifying particular diseases.

**Comparison With Prior Work**

The GPT-4V model has undergone training and uses 0-shot learning, which recognizes image features based on natural language to classify diseases based on image information and previously learned disease features \( [20] \). GPT-4V can yield effective results with fewer resources than previous deep learning models, which typically require a large amount of image data, computational resources, time, and parameter adjustments for training. By inputting new information rather than simply classifying image data, it becomes possible to tailor diagnoses and diagnostic aids for each individual. Furthermore, GPT-4V and other large-scale language processing models feature prompt development that is appropriate for its usage purposes, since the accuracy of such models varies depending on the prompt adjustments.

Compared with physicians’ accuracy, the model’s performance in this study was higher than that of a pediatrician but lower than that of an otolaryngologist. In a previous comparison between deep learning and humans, Crowson et al \( [21] \) classified 22 tympanic membrane images and found that the deep learning model achieved an accuracy of 95.5%, compared with an accuracy of 65% for 39 clinicians. Suresh et al \( [22] \) also reported that a machine-learning model created from 1000 images was more effective than pediatricians, with an accuracy rate of 90.6%, surpassing the clinicians’ accuracy of 59.4%. Our results indicated that the model did not reach the proficiency level of otolaryngologists; however, it could be valuable for using tympanic membrane images in medical practice outside of otolaryngology. In particular, GPT-4V judgments predominantly exceeded pediatricians’ correct response rates for questions with normal to hard difficulty, suggesting that the present model may be useful for nonotolaryngologists who have difficulty in making such judgments.

Moreover, previous reports on deep learning classification models have determined the presence or absence of inflammation and exudates based on photographs alone. Further studies are needed to identify the optimal stage in the examination for implementing the image classification model and the subsequent policy decisions that should follow.

GPT-4V allows for the classification of diseases using patient information. While comments about medical or harmful content (with restrictions on medical advice) may result in a lower correct response rate, informative or educational responses are still possible if they are well-informed. Efforts have been made to use large language models (LLMs) to improve the accuracy of prompts. Therefore, it is possible to develop appropriate prompts for medical imaging and middle ear disorders. The accuracy of the LLM is expected to further improve with the development of prompts that are specifically tailored for medical imaging and middle ear disease \( [23,24] \).
Acknowledgments
We are thankful to our colleagues at Shinshu University: Dr Sota Ichimura, Dr Kota Hirose, Dr Mariko Kasuga, Dr Shu Yokota, Dr Kentaro Hori, Dr Arisa Oguchi, Dr Kenjiro Sugiyama, Dr Jun Shinagawa, Dr Yoichiro Iwasa, Dr Keita Tsukada, Dr Tomohiro Oguchi, and Dr Nobuyoshi Suzuki of the Department of Otolaryngology—Head and Neck, and those at Jichi Medical University: Dr Kentaro Hori, Dr Arisa Oguchi, Dr Kenjiro Sugiyama, Dr Jun Shinagawa, Dr Yoichiro Iwasa, Dr Keita Tsukada, Dr Tomohiro Oguchi, Dr Nobuyoshi Suzuki of the Department of Otolaryngology, and Dr Hitoshi Irabu, Dr Kazuhiro Noguchi, Dr Ryo Nakagawa, Dr Narutoshi Yamazaki, Dr Yuji Takaso, Dr Keisuke Koyama, Dr Yukari Nakamura, Dr Chia Sasaki, and Dr Keigo Noda et al.
Nishida for their invaluable cooperation in this study. We thank Editage [30] for English language editing. The authors declare that no financial support was received for the research, authorship, or publication of this paper.

**Authors’ Contributions**

MN handled the conceptualization, data curation, formal analysis, investigation, methodology, project administration, resources, validation, visualization, and writing of the original draft. HY worked on the conceptualization, data curation, formal analysis, investigation, methodology, project administration, resources, validation, and review and editing of the writing. TO, RK, and YU handled the investigation, project administration, and review and editing of the writing. AN, MI, and YT did the supervision and review and editing of the writing.

**Conflicts of Interest**

None declared.

**Multimedia Appendix 1**

Representative image and prompt of this study. (A) Representative image of input and output to GPT-4 Vision. Input can be combined with text and images in input to obtain output. (B) Example of changing the prompt content and an output that asks for patient information. By presenting a concept as ORDER and adding conditions as restriction, appropriate prompts were attempted to be developed. In the output, it is required to input patient information such as age, medical history, and chief complaint. (C) An example of an answer with an optimized prompt. Present the diagnosis, the rationale for the diagnosis, and treatment and prevention methods.

[PDF File (Adobe PDF), 483 KB - Multimedia Appendix 1]

**References**


