Sumotosima: A Framework and Dataset for Classifying and Summarizing Otoscopic Images

Eram Anwarul Khan[1*], **Anas Anwarul Haq Khan**[2,3†]
[1]Bombay Hospital, Department of ENT
[2]IIT Bombay, Department of Computer Science and Engineering
[3]School of Computer Engineering, KIIT, Bhubaneswar
{eramkhan1607, anas290816007}@gmail.com

Abstract

Otoscopy is a diagnostic procedure to examine the ear canal and eardrum using an otoscope. It identifies conditions like infections, foreign bodies, ear drum perforations and ear abnormalities. We propose a novel resource efficient deep learning and transformer based framework, **Sumotosima** (**Sum**marizer for **otos**copic **ima**ges), an end-to-end pipeline for classification followed by summarization. Our framework works on combination of triplet and cross-entropy losses. Additionally, we use Knowledge Enhanced Multimodal BART whose input is fused textual and image embedding. The objective is to provide summaries that are well-suited for patients, ensuring clarity and efficiency in understanding otoscopic images. Given the lack of existing datasets, we have curated our own **OCASD** (**O**toscopic **C**lassification **A**nd **S**ummary **D**ataset), which includes 500 images with 5 unique categories annotated with their class and summaries by Otorhinolaryngologists. Sumotosima achieved a result of **98.03**%, which is **7.00**%, **3.10**%, **3.01**% higher than K-Nearest Neighbors, Random Forest and Support Vector Machines, respectively, in classification tasks. For summarization, Sumotosima outperformed GPT-4o and LLaVA by **88.53**% and **107.57**% in ROUGE-L scores, respectively. We have made our code and dataset publicly available at [1]

1 Introduction

Otoscopy is a medical procedure using an otoscope to visually inspect the ear canal and eardrum. It is performed by healthcare professionals such as Otorhinolaryngologists (Ear Nose Throat specialists). This examination helps identify conditions like ear infections, blockages, eardrum perforations, and other abnormalities. Otoscopy is crucial for detection in treatment of ear-related issues, ensuring better auditory health and preventing complications that can arise from untreated conditions. Our objective is to develop efficient summaries for otoscopic ear images that are both clear and well-defined for patients. A common complaint from patients is the lack of interaction and explanations from healthcare professionals, often due to the high workload and limited availability of medical staff, which limits the time doctors can spend with each patient. By improving the clarity of information through AI-generated summaries, we aim to enhance patient understanding and interaction with their healthcare providers. These summaries identify the disease if any and describe various conditions, including areas of redness, infection spots, ear drum perforations, etc. Further details are provided in Section 3. To facilitate our research, we identified lack of availability in open-source otoscopic image datasets, and none of them accompanying summaries or captions. To address this, we curated our dataset comprising 500 images equally distributed into 5 categories : 'Acute Otitis Media', 'Cerumen Impaction', 'Chronic Otitis Media', 'Myringosclerosis' and 'Normal' from sources such as (POLAT, 2021; Viscaino et al., 2020), and publicly available images on Google. It was was then filtered to remove noisy and redundant data that could introduce bias into the supervised learning results. Additionally, the images were summarized by an Otorhinolaryngologist to ensure the dataset's robustness. Further details and Annotation guidelines are provided in Section 3.1. Our framework is divided into two stages: classification followed by summarry generation. We observed that separating the classification task by applying a combination of triplet loss (Schroff et al., 2015) and cross-entropy loss and passing resultant information alongside the prompt to a Multimodal BART (Xing et al., 2021) yielded superior results. The main contributions of our proposed research are as follows:

[*]MBBS, ENT (Mumbai), MRCS ENT (Ireland)
[†]The authors are jointly the first authors.
[1]https://github.com/anas2908/Sumotosima

1. To best of our knowledge, **Sumotosima** is the first work towards understanding and summarization of otoscopic images.

2. We have curated first and the largest multimodal dataset till date, comprising 500 instances, belonging to 5 unique categories. Each instance in the dataset is accompanied by a Gold Standard summary created by Otorhinolaryngologist itself.

3. In lieu of resource-intensive models, we propose Sumotosima, an approach that involves SoTA classification followed by a summary generation through text and image embedding fusion before passing it to Multimodal BART.

4. The proposed method achieved the best classification results of **98.03**%, with an improvement of **3.01**% compared to previously utilized traditional machine learning approaches. For the summarization task, it demonstrated an improvement of **88.53**% compared to GPT-4o

2 Related Works

The classification of otoscopic images using traditional machine learning architectures has been extensively studied. Notable and recent research by (Başaran et al., 2020b; Goshtasbi, 2020; Crowson et al., 2021) has demonstrated the potential of machine learning for diagnosing ear conditions using otoscopic images. In recent years, Convolutional Neural Networks (CNNs) have become a popular approach for analyzing otoscopic images. These networks have shown better accuracy in detecting common ear diseases such as otitis media, tympanic membrane perforations, and cerumen impaction, as highlighted by studies from (Wu et al., 2021; Başaran et al., 2020a; Tsutsumi et al., 2021). Moreover, both Machine Learning (ML) and Deep Learning (DL) algorithms have been employed to classify various ear conditions and segment areas of interest in otoscopic images, aiding in more precise diagnosis and treatment planning, as demonstrated by (Pham et al., 2021a,b). Additionally, the integration of Artificial Intelligence (AI) with smartphone-connected otoscopes has facilitated remote diagnosis by analyzing images captured using smartphone attachments, providing real-time feedback to healthcare providers and patients, as shown by (Cortés Fuentes et al., 2024). Despite significant advancements in classification and segmentation tasks of otoscopic images using AI, there remains a lack of open-source datasets in this field. Furthermore, no substantial work has been done on summarizing otoscopic findings for better patient understanding. Addressing these gaps could significantly enhance the application of AI in otoscopy, which is the focus of our current research.

3 Dataset

Our dataset, **OCASD** (**O**toscopic **C**lassification **A**nd **S**ummary **D**ataset), was curated in response to the scarcity of available datasets and is derived from previously existing open-source datasets (POLAT, 2021; Viscaino et al., 2020). Instances in (Viscaino et al., 2020) were found to be extensively redundant for the categories 'Cerumen Impaction,' 'Chronic Otitis Media,' 'Myringosclerosis,' and 'Normal.' To address this, we manually removed redundant or nearly identical instances, and an Otorhinolaryngologist handpicked 100 images for each of the four categories, ensuring they were the most informative and contributive. A similar approach was followed for (POLAT, 2021) for category 'Acute Otitis Media' however, due to a lack of informative images, some were additionally sourced from the internet. The distribution of the curated collection is shown in Table 2. Another relatively large dataset (Dubois et al., 2024), containing 11 classes and more than 45,000 images, is not accessible (refer Table 1). After curation, OCASD comprises a total of 500 images across 5 unique categories. Each image is annotated with a summary by an Otorhinolaryngologist the textual statistics are mention in Table 3, following the steps outlined in Section 3.1

3.1 Annotation Steps and Guidelines

Objective : To create clear, concise, and patient-friendly summaries of otoscopic images, identifying the disease (if any) and describing various conditions such as redness, ear wax, infection spots, etc.

3.1.1 General Guidelines

Clarity and Simplicity

- Use simple and non-technical language that patients can easily understand.
- Avoid medical jargon; if medical terms must be used, provide a brief explanation.

Empathy and Reassurance

Dataset	Availability	Total Images	Unique Images	Classes	Annotated Summary
(Dubois et al., 2024)	Closed Source	45,606	-	11	×
(POLAT, 2021)	Open Source	282	×	7	×
(Viscaino et al., 2020)	Open Source	880	×	4	×
OCASD	Open Source	500	✓	5	✓

Table 1: Comparison of otoscopic image datasets based on their availability, total number of images, unique images, number of classes, and whether they include annotated summaries.

Categories	Source	Intial Instances	Instances after Filtering	Instances Included
CI	(Viscaino et al., 2020)	220	134	100
COM		220	174	100
MS		220	141	100
N		220	166	100
AOM	(POLAT, 2021)	83	74	69
	Internet	147	42	31

Table 2: Dataset filtering process showing the initial instances, instances after filtering, and the final instances included for each category from various sources. Categories include Cerumen Impaction (CI), Chronic Otitis Media (COM), Myringosclerosis (MS), Normal (N), and Acute Otitis Media (AOM).

- Write in a reassuring and empathetic tone.
- Aim to alleviate any anxiety by explaining conditions in a calm and positive manner.

Consistency

- Follow a consistent structure for all summaries to ensure uniformity and ease of understanding.

Accuracy

- Ensure that all information is accurate and reflects the observed condition.

3.1.2 Steps for Creating Summaries

Identify the Category

- Confirm the known category of the image (e.g., Cerumen Impaction, Chronic Otitis Media, Myringosclerosis, Normal, Acute Otitis Media).

Observe and Describe Key Features

- Note any visible signs of disease or abnormalities such as redness, swelling, infection spots, foreign body, wax buildup, perforations, or other notable conditions.
- Describe these features in simple terms.

Summarize the Condition

- Begin with a brief statement identifying the disease (if any) or stating that the ear appears normal.
- Example: "This Otoscopic image shows signs of Chronic Otitis Media, which is an infection of the middle ear."

Explain Symptoms and Impact

- Describe how the observed condition might affect the patient, focusing on symptoms they might experience.
- Example: "You may experience symptoms like ear pain, hearing loss, or discharge from the ear."

Review the summary for clarity, accuracy, and tone. Make revisions as necessary to ensure the summary is patient-friendly and informative.

4 Methodology

4.1 Classification

We followed a two-step architecture for our classification task. First, for each image (Anchor), we selected two additional images: one from the same category (Positive) and one from a different category (Negative). This setup facilitates use of triplet loss (Schroff et al., 2015) alongside cross-entropy loss for classification.

The images were resized to 226×226 and passed through ResNeXt-18 (He et al., 2016) to obtain vector features of $\mathbb{R}^{1\times128}_{\text{anc}}$, $\mathbb{R}^{1\times128}_{\text{pos}}$ and $\mathbb{R}^{1\times128}_{\text{neg}}$

Categories	Total Length	Average Words	Vocabulary Richness	Medical Jargon Density
Cerumen Impaction	5487	54.87	426	45
Chronic Otitis Media	4407	44.07	384	73
Myringosclerosis	3583	35.83	233	39
Normal	5759	57.59	368	68
Acute Otitis Media	5862	58.62	370	56

Table 3: Summary of text analysis across different otoscopic categories, including total length, average words, vocabulary richness, and medical jargon density. Efforts have been made to present information in a patient-friendly manner by minimizing medical jargon, ensuring clarity and accessibility for all readers.

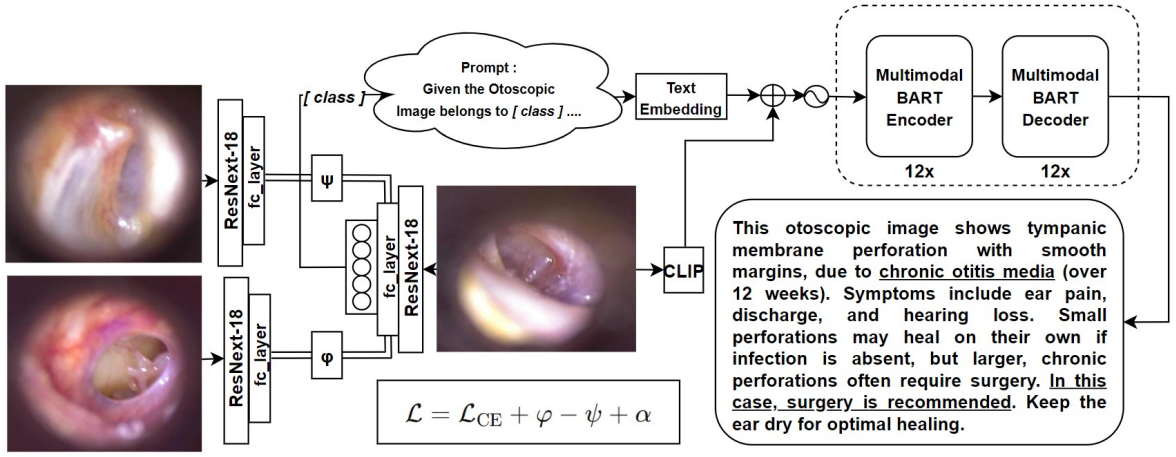

Figure 1: Overview of the end-to-end framework, **Sumotosima** (**Sum**marizer for **otos**copic **ima**ges), designed for classifying and generating summaries of otoscopic images.

for Anchor, Positive, and Negative images respectively. The Euclidean distance between $\mathbb{R}^{1\times128}_{\text{anc}}$ and $\mathbb{R}^{1\times128}_{\text{pos}}$ is denoted by φ, and the distance between $\mathbb{R}^{1\times128}_{\text{anc}}$ and $\mathbb{R}^{1\times128}_{\text{neg}}$ is denoted by ψ. The objective is to minimize φ and maximize ψ, with a margin α to enforce a minimum distance between positive and negative pairs, thus enhancing the discriminative capability of the learned embeddings. α is set to the default value (Schroff et al., 2015).

$$a_i = \mathbb{R}^{1\times128}_{\text{anc}}, \quad p_i = \mathbb{R}^{1\times128}_{\text{pos}}, \quad n_i = \mathbb{R}^{1\times128}_{\text{neg}} \tag{1}$$

$$\varphi = \|a_i - p_i\|_2^2, \quad \psi = \|a_i - n_i\|_2^2 \tag{2}$$

The triplet loss is given by:

$$L_{\text{triplet}} = \sum_{i=1}^{N} [\varphi - \psi + \alpha]_+ \tag{3}$$

The cross-entropy loss is defined as:

$$L_{\text{CE}} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} y_{i,c} \log(\hat{y}_{i,c}) \tag{4}$$

- N is the total number of image instances.
- C is the number of classes (set to 5).
- $y_{i,c}$ is the binary indicator (0 or 1) if class label c is the correct classification for instance i.
- $\hat{y}_{i,c}$ is the predicted probability of instance i being class c.

The total loss while training the model is:

$$L = L_{\text{triplet}} + L_{\text{CE}} \tag{5}$$

The loss backpropagates and adjusts the fully connected layer of the model, denoted as *fc_layer* in Figure 1, thereby refining the vector representation of size $\mathbb{R}^{1\times128}$ in each epoch. After classification, the class is included as part of the prompt for robust results.

4.2 Generation

The anchor image undergoes preprocessing through transformations, including resizing, center cropping, and normalization. The processed image is

then fed into a fine-tuned CLIP (Radford et al., 2021), yielding a dense vector $\mathbf{v}_{\text{CLIP}} \in \mathbb{R}^{1\times 512}$. The resulting embeddings $\mathbf{v}_{\text{CLIP}}$ are combined with the text embeddings $\mathbf{e}_{\text{prompt}} \in \mathbb{R}^d$, which are obtained by passing the class-enriched prompt $\mathbf{p}_c$ through the KM-BART encoder. Although CLIP has not been extensively trained on medical image datasets, it has effectively learned spatial, textual, and color characteristics of images. After fine-tuning on OCASD, CLIP can identify fine details such as the granularity of middle ear parts, the level of redness, potential irritability, and the shape of tympanic membranes. This capability is crucial for determining the criticality of the image in relation to the class specified by the prompt. The fused embeddings $\mathbf{v}_{\text{fused}} = \mathbf{v}_{\text{CLIP}} + \mathbf{e}_{\text{prompt}}$ are then augmented with positional encodings as described in (Vaswani et al., 2017), before being passed to the subsequent KM-BART encoder. Let k denote the position index and i the dimension index. The updated positional encoding is given by:

$$\mathbf{v}'_{\text{fused}} = \mathbf{v}_{\text{fused}} + \mathbf{r}_{k/i}, \tag{6}$$

where $\mathbf{r}_{k/i}$ represents the positional encoding adjustment based on the relative position k and dimension i.

5 Evaluation Metrics

For Classification task we employed Precision, Recall and F1-score.

Precision: The proportion of true positive predictions among all positive predictions. It measures the accuracy of positive predictions.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

Recall: The proportion of true positive predictions among all actual positives. It measures how well the model identifies all relevant cases.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

F1-score: The harmonic mean of Precision and Recall. It provides a single metric that balances both precision and recall.

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

For Generated Summary, we evaluated our model using Automatic metrics and Human Evaluation. For Automatic Evaluation, we used BLEU (Papineni et al., 2002) for translation quality, BERTScore (Zhang et al., 2019) for fluency and coherence and ROUGE (Lin, 2004) for summarization quality. Human Evaluation included Patient Friendliness and Faithfulness. Faithfulness was annotated by our in-house Otorhinolaryngologist as the percentage of error-free samples. For assessing Patient Friendliness, we employed seven annotators with diverse backgrounds: three with some knowledge of medical terminology from science background, and four working professionals—two men and two women—from commerce and arts. Each annotator rated 100 summaries on a three-point scale: **1 for "no use,"** indicating the summary provided no valuable information; **2 for "somewhat useful,"** meaning it included general information about the disease class, such as symptoms like redness, irritability, or ear pain, along with trivial suggestions; and **3 for "very useful,"** where the summary addressed the specific scenario, detailing whether surgery is required or identifying the exact part of the ear affected, offering non-trivial insights that require image analysis. This approach ensured a comprehensive evaluation of patient friendliness from both semi-informed and layperson perspectives.

6 Experiments and Results

6.1 Classification

For classification, we compared machine learning approaches such as KNN (Guo et al., 2003), Random Forest (Breiman, 2001), and Support Vector Machine. Support Vector Machine (Cortes and Vapnik, 1995) was found to perform best among the machine learning algorithms. We then applied the ResNeXt-18 architecture for the deep learning classification task and experimented with cross-entropy and triplet loss functions. We observed that the combination of cross-entropy and triplet loss functions performed best. To handle any bias or error due to the scarcity of the dataset, we performed a **5-fold cross-validation** approach. The average results observed showed a **3.15%** gain over the best performing machine learning approach, a **1.34%** gain over the deep learning architecture with a single cross-entropy loss, and a **1.55%** gain over the deep learning architecture with a single triplet loss. Further refer Table 4

Models	P	R	F1
K Nearest Neighbour	0.911	0.910	0.910
Random Forest	0.954	0.950	0.951
Support Vector Machine	0.953	0.950	0.950
ResNeXt18 (L_{CE})	0.971	0.967	0.968
ResNeXt18 (L_{trp})	0.954	0.978	0.966
ResNeXt18 ($L^{\text{fold 1}}_{\text{CE + trp}}$)	0.973	0.963	0.965
ResNeXt18 ($L^{\text{fold 2}}_{\text{CE + trp}}$)	0.976	0.975	0.975
ResNeXt18 ($L^{\text{fold 3}}_{\text{CE + trp}}$)	1	1	1
ResNeXt18 ($L^{\text{fold 4}}_{\text{CE + trp}}$)	0.988	0.988	0.988
ResNeXt18 ($L^{\text{fold 5}}_{\text{CE + trp}}$)	0.978	0.975	0.975
ResNeXt18 ($L^{\text{avg}}_{\text{CE + trp}}$)	**0.983**	**0.980**	**0.981**

Table 4: Performance comparison of different models on the classification task using Precision (P), Recall (R), and F1-score ($F1$). The best results are achieved by ResNeXt18 using a combination of cross-entropy and triplet loss, with the average results across five folds highlighted. The z-score is approximately 3.81, and the two-tailed p-value is approximately 0.00014. This very small p-value suggests that the difference between the proportions 0.983 (ours) and 0.953 (SVM) is statistically significant.

6.2 Generation

For generation, we compared our model with the open-source vision-language model LLaVA (Liu et al., 2023) and the closed-source GPT-4o with the given prompt : *Given an otoscopic image, generate a summary ensuring clarity and simplicity by using simple and non-technical language that patients can easily understand, avoiding medical jargon unless a brief explanation is provided. When creating summaries, first identify the category of the image (e.g., Cerumen Impaction, Chronic Otitis Media, Myringosclerosis, Normal, Acute Otitis Media). Observe and describe any visible signs of disease or abnormalities such as redness, swelling, infection spots, wax buildup, perforations, or other notable conditions in simple terms. Begin the summary with a brief statement identifying the disease (if any) or stating that the ear appears normal. For example, "This otoscopic image shows signs of Chronic Otitis Media, which is an infection of the middle ear." Explain how the observed condition might affect the patient, focusing on symptoms they might experience, such as, "You may experience symptoms like ear pain, hearing loss, or discharge from the ear." Finally, review the summary for clarity, accuracy, and tone, making revisions as necessary to ensure it is patient-friendly and informative. Everything must be in one paragraph.* For the prompt to Sumotosima, we already predefined the class found from the classification model. We also experimented with the following variations of Sumotosima:

- $Sumotosima_{\mathbf{1}}$: This variant uses traditional BART (Lewis et al., 2019) and not KM-BART. The image information is first converted to captions using a pretrained CLIP model and is passed as input along with the prompt.
- $Sumotosima_{\mathbf{2}}$: This variant uses traditional BART (Lewis et al., 2019) and not KM-BART. The image information is extracted as dense vectors using a pretrained CLIP model and is passed as input along with the prompt.
- $Sumotosima_{\mathbf{3}}$: This variant uses traditional BART (Lewis et al., 2019) and not KM-BART. The image information is first converted to captions using a Finetuned CLIP model on OCASD.
- $Sumotosima_{\mathbf{4}}$: This variant uses traditional BART (Lewis et al., 2019) and not KM-BART. The image information is extracted as dense vectors using a Finetuned CLIP model on OCASD.
- $Sumotosima_{\mathbf{5}}$: This variant uses KM-BART. The image information is extracted as dense vectors using a Finetuned CLIP model on OCASD.

We found $Sumotosima_5$ to give the best results among all the ablation studies and vision-language models used as baselines. Specifically, $Sumotosima_5$ surpassed GPT-4o by **88.53%** and LLaVA by **107.57%** in ROUGE scores. Given that Sumotosima is fine-tuned, it may produce summaries that perform better on automatic evaluation metrics like BLEU and ROUGE. This questions the reliability of these scores. To address this issue, we incorporated BERT-F1 as part of our evaluation framework. $Sumotosima_5$ also surpassed GPT-4o by **17.65%** and LLaVA by **23.09%** in BERT-F1 scores. Further refer Table 5

We also conducted human evaluation as mentioned in Section 5. We found $Sumotosima_5$ to be **19%** more faithful in the generation of summaries and **33.33%** more patient-friendly than GPT-4o. Among all Human Evaluation Metrics, LLaVA was the worst performing. Further refer Table 6

Model	BLEU	BERT-F1	ROUGE-L
GPT-4o	0.0348	0.442	0.185
LLaVA	0.029	0.423	0.168
$Sumotosima_1$	0.063	0.285	0.228
$Sumotosima_2$	0.071	0.296	0.184
$Sumotosima_3$	0.092	0.318	0.253
$Sumotosima_4$	0.101	0.347	0.281
$Sumotosima_5$	**0.128**	**0.521**	**0.349**

Table 5: Comparison of different models for summarization tasks using BLEU, BERT-F1, and ROUGE-L metrics. $Sumotosima_5$ achieved the best performance across all metrics. The Z-score is approximately 8.29, and the two-tailed p-value is approximately 2.22×10^{-16}. This extremely small p-value indicates a statistically significant difference between GPT-4o and $Sumotosima_5$.

Model	Patient Friendliness	Faithfulness
GPT-4o	1.8	73.0%
LLaVA	1.2	38.5%
$Sumotosima_5$	**2.4**	**92.0%**

Table 6: Human Evaluation of models on patient friendliness and faithfulness. Sumotosima$_5$ outperformed GPT-4o and LLaVA in both criteria.

6.3 Experimental Setup

We performed a grid search for hyperparameters using the Adam optimizer and an Nvidia A100 GPU for a total of 3 hours. Refer to Table 7.

Parameters/Resources	Classification	Generation
Training Split	70%	60%
Validation Split	15%	20%
Test Split	15%	20%
Epochs	100	50
Batch Size	32	8
Learning Rate	1e-3	3e-5
GPU	4.5GB	18GB

Table 7: Training parameters and resource allocation for classification and generation tasks, including dataset splits, epochs, batch size, learning rate, and GPU usage.

7 Conclusion

In this work, we introduced a novel pipeline for classifying and generating summaries for otoscopic images of the middle ear, with the objective of developing summaries that are both well-defined and patient-friendly, addressing the challenge of insufficient explanations from medical professionals due to their hectic schedules and limited time per patient. Unlike previous approaches that relied on traditional machine learning algorithms or straightforward deep learning architectures, our model utilizes a combination of triplet loss and cross-entropy loss, built upon the ResNeXt-18 architecture, achieving a classification accuracy of **98.03%**, surpassing all baselines on our OCASD dataset, which comprises 500 images across 5 different classes. As this is the first work on summarization for otoscopic images, we addressed the lack of annotated summary datasets by creating summaries for all 500 images, combining class-enriched prompts with image embeddings obtained from a fine-tuned CLIP model, and feeding these into a Multimodal BART model for summary generation. Our framework, Sumotosima, significantly outperformed GPT-4o and LLaVA in summarization, with improvements of **88.53%** and **107.57%** in ROUGE scores, respectively. Looking ahead, we plan to extend this work by incorporating additional metadata, such as patient age, gender, medical history, and educational background, to generate more robust and patient-specific summaries.

8 Limitations and Ethical Considerations

One limitation of our summarization process for otoscopic images is the small size of the dataset, which may affect the robustness and generalizability of the findings. Although we adopted cross-validation techniques to mitigate this issue, the limited dataset size remains a significant constraint. Additionally, there is a lack of consistency in the data sources, with some categories derived from one source and others from different sources, potentially impacting the uniformity of the information. On the ethical side, the summaries were created by an Otorhinolaryngologist, ensuring high quality and reliability in the interpretations. This approach prioritizes the accuracy and trustworthiness of the summaries, addressing potential concerns related to the validity of the data.

References

Erdal Başaran, Zafer Cömert, and Yüksel Çelik. 2020a. Convolutional neural network approach for automatic tympanic membrane detection and classification. *Biomedical Signal Processing and Control*, 56:101734.

Erdal Başaran, Zafer Cömert, Yüksel Çelik, Ümit Budak, and Abdulkasdir Şengür. 2020b. Otitis media diagnosis model for tympanic membrane images processed in two-stage processing blocks. In *Modelling*

and Analysis of Active Biopotential Signals in Healthcare, Volume 2, pages 14–1. IOP Publishing Bristol, UK.

Leo Breiman. 2001. Random forests. *Machine learning*, 45:5–32.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20:273–297.

Ignacio A Cortés Fuentes, Úrsula Zelada B, and Pedro Zitko M. 2024. Smartphone-based video-otoscopy performed by non-ent physicians as a complementary tool to conventional otoscopy for tympanic membrane diagnosis. *Speech, Language and Hearing*, 27(1):14–19.

Matthew G Crowson, Christopher J Hartnick, Gillian R Diercks, Thomas Q Gallagher, Mary S Fracchia, Jennifer Setlur, and Michael S Cohen. 2021. Machine learning for accurate intraoperative pediatric middle ear effusion diagnosis. *Pediatrics*, 147(4).

Constance Dubois, David Eigen, François Simon, Vincent Couloigner, Michael Gormish, Martin Chalumeau, Laurent Schmoll, and Jérémie F Cohen. 2024. Development and validation of a smartphone-based deep-learning-enabled system to detect middle-ear conditions in otoscopic images. *npj Digital Medicine*, 7(1):162.

Khodayar Goshtasbi. 2020. *Machine Learning Models to Predict Diagnosis and Surgical Outcomes in Otolaryngology*. University of California, Irvine.

Gongde Guo, Hui Wang, David Bell, Yaxin Bi, and Kieran Greer. 2003. Knn model-based approach in classification. In *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3-7, 2003. Proceedings*, pages 986–996. Springer.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023. Improved baselines with visual instruction tuning.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Van-Truong Pham, Thi-Thao Tran, Pa-Chun Wang, Po-Yu Chen, and Men-Tzung Lo. 2021a. Ear-unet: A deep learning-based approach for segmentation of tympanic membranes from otoscopic images. *Artificial Intelligence in Medicine*, 115:102065.

Van-Truong Pham, Thi-Thao Tran, Pa-Chun Wang, and Men-Tzung Lo. 2021b. Tympanic membrane segmentation in otoscopic images based on fully convolutional network with active contour loss. *Signal, Image and Video Processing*, 15(3):519–527.

KEMAL POLAT. 2021. eardrum.zip.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823.

Kotaro Tsutsumi, Khodayar Goshtasbi, Adwight Risbud, Pooya Khosravi, Jonathan C Pang, Harrison W Lin, Hamid R Djalilian, and Mehdi Abouzari. 2021. A web-based deep learning model for automated diagnosis of otoscopic images. *Otology & Neurotology*, 42(9):e1382–e1388.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Michelle Viscaino, Juan C Maass, Paul H Delano, Mariela Torrente, Carlos Stott, and Fernando Auat Cheein. 2020. Computer-aided diagnosis of external and middle ear conditions: A machine learning approach. *Plos one*, 15(3):e0229226.

Zebin Wu, Zheqi Lin, Lan Li, Hongguang Pan, Guowei Chen, Yuqing Fu, and Qianhui Qiu. 2021. Deep learning for classification of pediatric otitis media. *The Laryngoscope*, 131(7):E2344–E2351.

Yiran Xing, Zai Shi, Zhao Meng, Gerhard Lakemeyer, Yunpu Ma, and Roger Wattenhofer. 2021. Km-bart: Knowledge enhanced multimodal bart for visual commonsense generation. *arXiv preprint arXiv:2101.00419*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Frequently Asked Questions (FAQs)

1. **Dataset Size and Diversity: Considering the limited size of the curated OCASD dataset, do you have any plans to expand it? Additionally, how do you ensure diversity?**
   Ans. The OCASD dataset consists of 500 unique images, distributed across 5 distinct classes. We are currently in discussions with various medical institutions to expand the dataset size by sixfold. Additionally, we plan to introduce new classes while ensuring the dataset remains as balanced as possible and maintains diversity.

2. **Just curious, if you have tested the model on external datasets or different types of medical images to assess generalizability?**
   Ans. The OCASD dataset comprises unique images collected from three diverse sources, including internet-sourced images verified by an otorhinolaryngologist. This ensures the model has exposure to a variety of otoscopic image types, reducing bias toward any specific otoscopic tool. However, the model has not been tested on other medical image types (e.g., Chest X-rays, Brain CT scans), as it is explicitly designed for otoscopic images. Using it on unrelated medical images could lead to hallucinations or random outputs. In the future, we aim to enhance the model's discriminative capabilities to differentiate otoscopic images from other types of medical images.

3. **How reliable are the classification and summarization results, given that they were done by a single expert, and how likely is it that two medical experts would fully agree on the diagnostic or treatment procedures in the same scenario?**
   Ans. The OCASD images were sourced from three pre-existing classification tasks with predetermined classes. Our in-house otorhinolaryngologist meticulously filtered out images that did not meet quality standards. She also extended the classifications into summaries, completing this task independently due to the limited availability of medical experts. We plan to make the dataset publicly available, inviting qualified experts to contribute to its further development and expansion.

4. **Does the author have plans to expand this work to Indian languages?**
   Ans. Our current priority is to expand the dataset. One of the key objectives of our work is to develop a model capable of generating patient-friendly otoscopic summaries, and addressing language barriers could add significant value. Achieving this will require collaboration with qualified medical experts from different regions of India. We plan to make the dataset publicly available to encourage contributions from experts who wish to further enhance it.

5. **Would the authors consider fine-tuning with open-source LLAVA, and how would the performance of the model compare to the current fine-tuned model and zero-shot performance of publicly available multimodal LLMs?**
   Ans. We have compared the results of zero-shot LLAVA and GPT-4o with our model and observed the superior performance of $Sumotosima$. $Sumotosima$ was designed in a resource-efficient manner, incorporating language models like KM-BART and CLIP. In the future, as we expand the dataset, we will have ample resources to fine-tune VLMS like LLaVA, which has billions of parameters. Currently, with only 500 image-text pairs, fine-tuning such models may not be optimal at this stage.

OCASD Example

Our dataset, **OCASD** (**O**toscopic **C**lassification **A**nd **S**ummary **D**ataset), is derived from previously existing open-source datasets. Instances were found to be extensively redundant for the categories 'Cerumen Impaction,' 'Chronic Otitis Media,' 'Myringosclerosis,' 'Normal,' and 'Acute Otitis Media.' To address this, we manually removed redundant or nearly identical instances, and an Otorhinolaryngologist handpicked 100 images for each of the five categories. After curation, OCASD comprises a total of 500 images across 5 unique categories. Each image is annotated with a summary by an Otorhinolaryngologist. A few instances of all five classes can be found below in the Figures 1, 2, 3, 4, 5 .

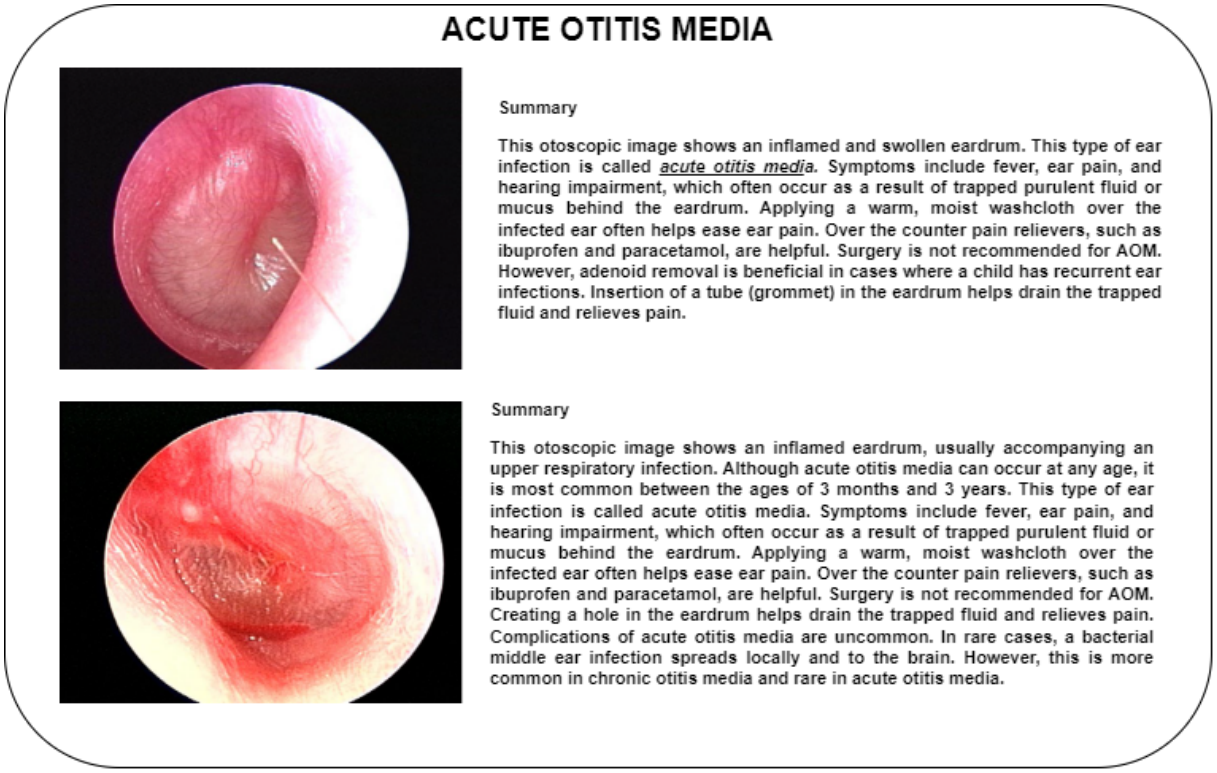

Figure 1: Example of otoscopic image illustrating **Acute Otitis Media (AOM)**. This image showcases inflammation and features typical of AOM, including redness and possible bulging of the eardrum.

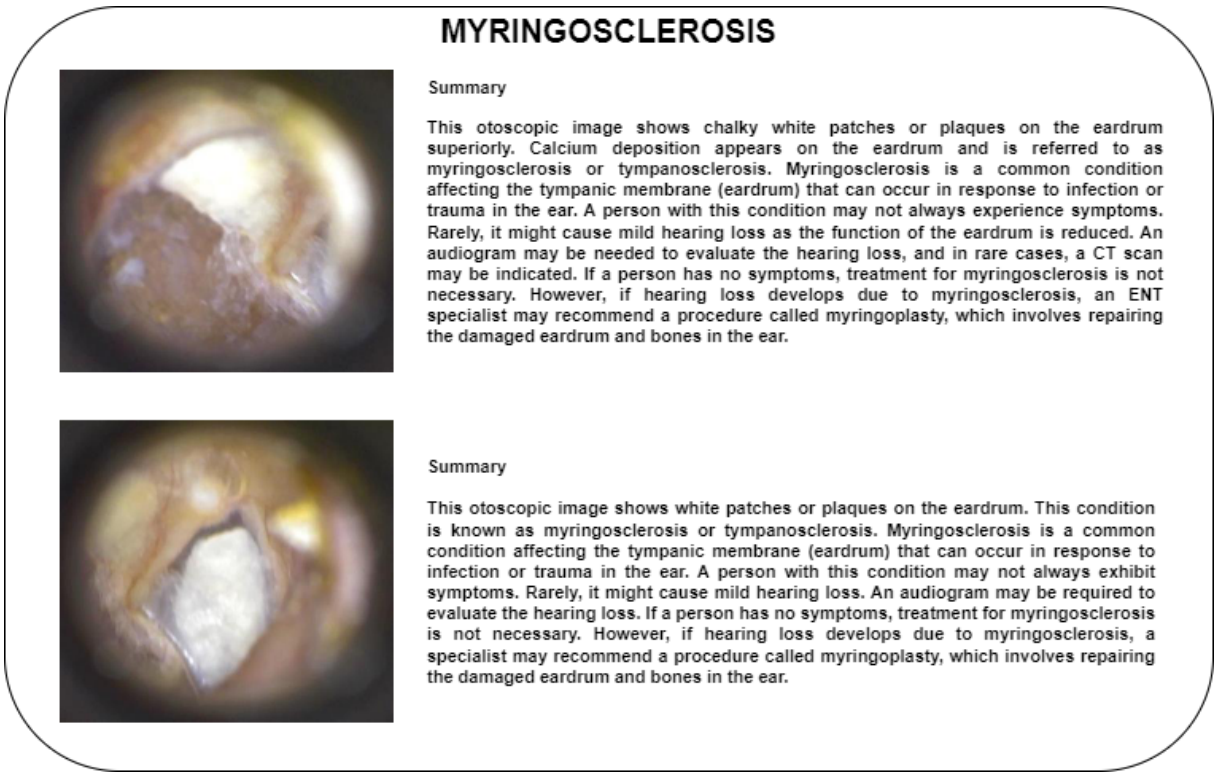

Figure 2: Example of otoscopic image depicting **Myringosclerosis**. White patches or plaques on the eardrum are evident, often resulting from calcium deposition due to infection or trauma.

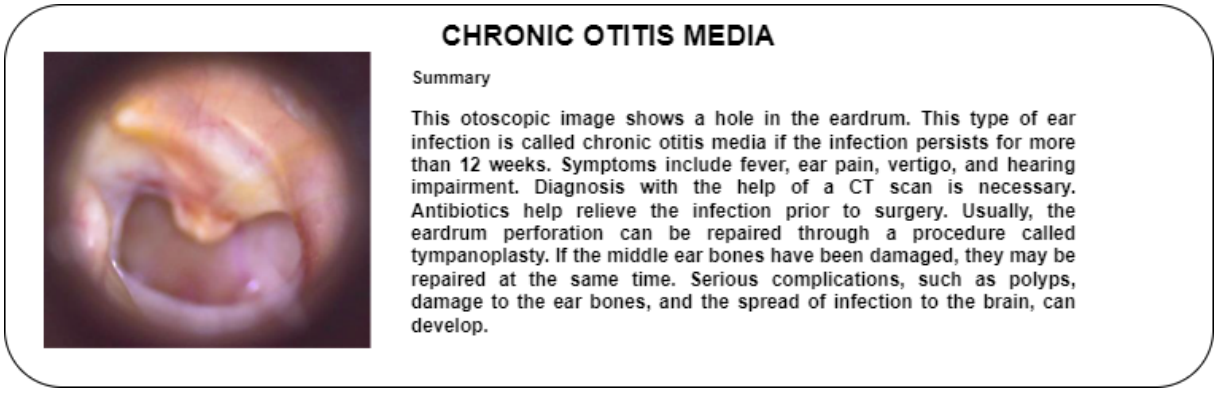

Figure 3: Example of otoscopic image showing **Chronic Otitis Media (COM)**. Visible perforation and chronic inflammation of the eardrum are characteristic of this condition.

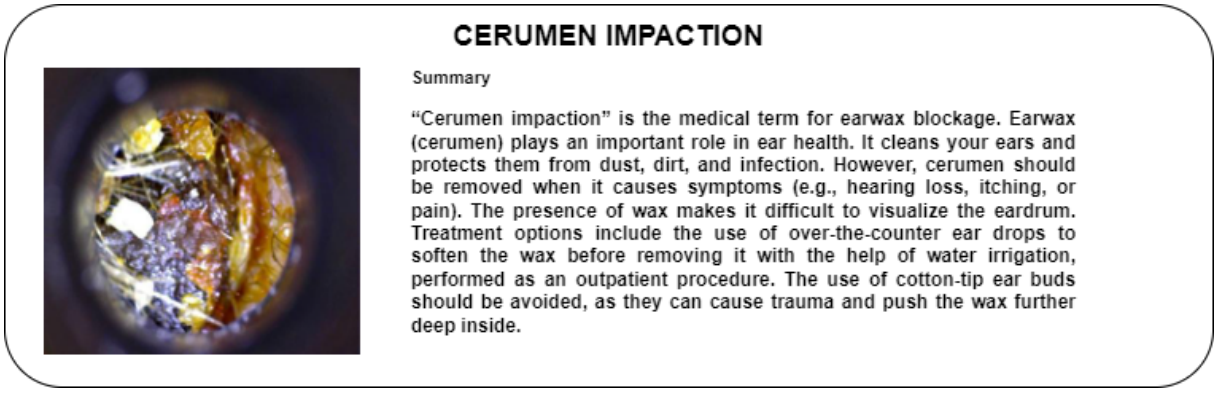

Figure 4: Example of otoscopic image indicating **Cerumen Impaction (CI)**. Excessive accumulation of earwax obscures the eardrum and external auditory canal.

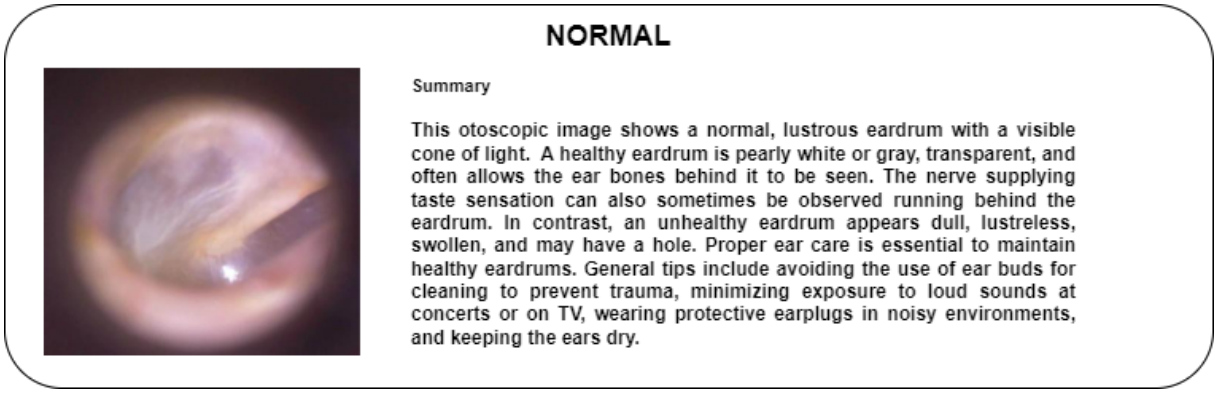

Figure 5: Example of otoscopic image demonstrating a **Normal Eardrum**. A healthy, lustrous, and translucent eardrum is visible, often showing underlying structures like ear bones.