Volume 3, Issue 3, 2025

Frontier in Medical & Health Research

ISSN: (e) 3007-1607 (p) 3007-1593

SMARTPHONE DEEP LEARNING OTOSCOPY FOR EAR DISEASE DETECTION IN A LOW-RESOURCE SETTING: A PROSPECTIVE PILOT STUDY

Dr Shah Wali[*1], Dr Arif Achakzai[2]

[*1,2]*Assistant professor, ENTvand Head and Neck Surgery Department Bolan Medical College Quetta*

[*1]walishah3330@gmail.com, [2]drarifachakzai@gmail.com

DOI: https://doi.org/10.5281/zenodo.15534544

Keywords

Article History
Received on 20 April 2025
Accepted on 20 May 2025
Published on 28 May 2025

Corresponding Author: *
Dr Shah Wali

Abstract

Background: *Ear diseases such as wax impaction, tympanic membrane (TM) perforations, and infections are common and can cause hearing loss, particularly in low-resource areas. Conventional otoscopic diagnosis is difficult for non-specialists, prompting research into smartphone-based deep learning (DL) otoscopy to improve diagnostic accuracy and accessibility. This study assessed a smartphone-integrated DL system for classifying ear findings (wax impaction, TM perforation, infection, and normal TM) in an outpatient setting at Bolan Medical College in Quetta.*

Methods*: We conducted a 6-month prospective study with 80 patients (<100 as a pilot sample) who presented with ear complaints. A smartphone-attached digital otoscope was used to capture otoscopic images, which were then analyzed using a deep learning model (YOLOv5 object detection and EfficientNet classification). The model was trained on an augmented dataset of 320 images (80 per category) using transfer learning. ENT specialists established ground truth diagnoses. We calculated sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) for each category and compared them to previously published results.*

Results: *The median age was 29 years (range 5-64), 56% were female, and 65% came from rural areas. Infection was the most common diagnosis (35%), followed by wax impaction (27.5%), TM perforation (22.5%), and normal TM (15%). The DL model had a total accuracy of 91.3%, correctly classifying 73 of 80 cases. Sensitivity was 91.7% for normal TM, 95.5% for wax impaction, 88.9% for TM perforation, and 89.3% for infection, with specificity ranging from 94.2% to 100% (Table 1). The PPV and NPV were high across all categories (Table 1). Figure 2 shows the model's sensitivity and specificity by category (). The diagnostic performance for wax and perforation was particularly strong, with no false positives (PPV = 100%). The model performed slightly worse in detecting infections, with a few otitis cases misclassified as normal or perforated.*

Conclusion: *This pilot study demonstrates the viability of smartphone-based DL otoscopy in a low-resource clinical environment. The model achieved diagnostic accuracy comparable to expert evaluation for major ear conditions. Implementing AI-assisted otoscopy could potentially increase access to early ear disease detection in rural and underserved areas. More research with larger multicenter trials are needed to validate and refine the model, incorporate tympanic membrane*

Volume 3, Issue 3, 2025

Frontier in Medical & Health Research

ISSN: (e) 3007-1607 (p) 3007-1593

segmentation, and address real-world issues like variable image quality and diverse pathologies. Our findings back up the promise of smartphone DL otoscopy as an affordable tool for improving ear care equity and outcomes.

INTRODUCTION

Ear diseases are a leading cause of morbidity and hearing loss worldwide. According to the World Health Organization, over 5% of the global population (430 million people) has disabling hearing loss, with nearly 80% of these people living in low- and middle-income countries. Otitis media (middle ear infection) has been identified as the leading preventable cause of hearing loss in children, accounting for approximately 46.9% of pediatric hearing loss in 2021. Chronic middle ear infections are frequently misdiagnosed or mistreated in areas like Pakistan due to a lack of access to specialist care, contributing to high rates of hearing loss and complications. Early detection and accurate diagnosis of ear conditions (such as wax impaction, tympanic membrane (TM) perforation, and otitis media) are critical for avoiding long-term consequences like hearing loss, speech delays, and intracranial infections.

Otoscopy, or examination of the ear with an otoscope, is the primary diagnostic tool for middle and external ear conditions. However, proper otoscopic diagnosis necessitates extensive training and experience. Non-specialist clinicians frequently struggle to interpret TM findings, resulting in misdiagnosis or missed disease. Previous research suggests that general practitioners and pediatricians diagnose otitis media at rates similar to chance in some cases (accuracy as low as ~50% in primary care). Even among experts, interobserver variability can be substantial. This diagnostic challenge is exacerbated in low-resource settings, where otolaryngologists are scarce and high-quality otoscopic equipment may be unavailable. As a result, there is an urgent need for assistive diagnostic tools that can improve ear examination accuracy and increase access to specialist-level assessments.

Recent advances in artificial intelligence (AI) and deep learning (DL) hold great promise for medical image analysis, including automated otoscopic image interpretation. Convolutional neural networks (CNNs) can learn to recognize subtle patterns of ear disease (such as a bulging erythematous TM in acute otitis media or a tympanic perforation) with great accuracy. Multiple research groups have developed deep learning models for classifying ear conditions from images, with diagnostic accuracy frequently exceeding 90%. Livingstone and Chau (2020) demonstrated an automated machine learning approach (Google AutoML Vision) that matched specialist performance in otoscopic diagnosis. In pediatric otitis media, CNN models have successfully distinguished between acute infection, effusion, and normal ears. A 2022 meta-analysis by Habib et al. found that AI algorithms could classify ear disease with a pooled accuracy of 93.4%, significantly outperforming human assessors' 73.2% accuracy. Similarly, various CNN architectures (e.g., ResNet, DenseNet, EfficientNet) have demonstrated excellent performance (often 90-98% accuracy) in distinguishing between multiple eardrum conditions. An EfficientNet-based model classified normal vs diseased TM and identified earwax with ~99% sensitivity and specificity on a large dataset. These findings indicate that DL has the potential to provide expert diagnostic support in otoscopy.

Building on these advances, there is an increasing interest in deploying AI-powered diagnostic systems on portable devices. Smartphone-based otoscopy has already been implemented in clinical practice through low-cost attachments that allow visualization of the ear using the phone's camera. Smartphone otoscopes capture images and videos of the TM, which can then be stored or transmitted for telemedicine. Integrating DL algorithms directly into smartphones (or via cloud services) may enable real-time interpretation of otoscopic images at the point of care. This approach could be especially beneficial in low-resource settings and primary care, where AI could guide non-specialists in diagnosing ear pathology. Chen et al. (2022) conducted a notable retrospective study in which they developed a smartphone-based AI algorithm using transfer learning and reported an accuracy of 98% for classifying ten middle ear conditions on a dataset of 2,820 images. Recently, Dubois et al. (2024)

Frontier in Medical & Health Research

ISSN: (e) 3007-1607 (p) 3007-1593

described an end-to-end smartphone otoscopy system ("i-Nside" app) that achieved >95% sensitivity and specificity for detecting abnormal eardrums in a validation study. Despite these successes, most AI otoscopy studies to date have been retrospective or conducted in high-resource settings. There is still a lack of prospective data on deploying smartphone-based DL otoscopy in real-world, low-resource clinical settings, where challenges such as variable image quality, diverse patient demographics, and limited training data may impact performance.

In this study, we wanted to assess the feasibility and diagnostic performance of a smartphone-based deep learning otoscopy system in a low-resource environment. We conducted a prospective pilot study at Bolan Medical College's ENT Department (Quetta, Pakistan) to identify four major clinical categories: (1) wax impaction, (2) TM perforation, (3) infection (otitis media or externa), and (4) normal TM. These are the most common findings in routine otology practice and have direct implications for management. We hypothesized that a deep learning model could accurately classify these categories from smartphone otoscope images, approaching specialist performance. The goals were to (a) create a lightweight deep learning model that can be deployed on a smartphone, (b) prospectively evaluate its sensitivity, specificity, PPV, and NPV against expert diagnoses, and (c) identify implementation challenges and equity considerations in our resource-constrained context. Finally, this study aims to shed light on the potential role of AI-assisted smartphone otoscopy in improving ear care access and outcomes in developing countries.

Methods

Study Design and Setting

We carried out a single-center prospective diagnostic study in the Department of Otorhinolaryngology (ENT) at Bolan Medical College Hospital in Quetta, Pakistan. The study lasted six months (March 1, 2025 to August 31, 2025) in a real-world outpatient setting. The Institutional Review Board of Bolan Medical College granted ethical approval (Approval #BMC-ENT-2025-01), and all participants (or guardians for minors) provided written informed consent. The study followed the Helsinki Declaration and local ethical guidelines for human-subject research.

Patients who presented to the ENT clinic with ear-related symptoms (such as hearing loss, ear pain, discharge, or routine ear check-ups) were eligible to participate. We included both adult and pediatric patients (no age restrictions) to represent the typical case mix. Exclusion criteria included postoperative ears (such as mastoid cavities or ventilation tubes), indistinct images (due to severe obstruction beyond wax impaction), and patients who refused to participate. This ensured that our image dataset included clear depictions of the TM or external canal in its intact state. Each patient underwent a standard clinical evaluation by an ENT specialist, which served as the baseline diagnosis for study outcomes.

Sample Size Considerations

As a pilot study, we aimed for a sample size of less than 100 patients, specifically 80-90 subjects, to provide preliminary estimates of diagnostic accuracy. We calculated the sample size for sensitivity using diagnostic test evaluation methods. Based on prior studies, we estimated a sensitivity of ~90% for the AI model and a desired margin of error of ±10% (95% confidence level), resulting in approximately 35 positive cases per category. Given four diagnostic categories, we anticipated ~20 patients in each category would yield roughly 20 positive instances per class (since each patient would contribute one condition). While this pilot would not fully meet the ideal sample size for each outcome, it would allow estimation with wide confidence intervals. We set a target of approximately 80 patients to balance feasibility and include multiple examples of each condition. This sample size was deemed adequate for an initial feasibility assessment and to inform power calculations for a larger future study. During this six-month period, we enrolled 80 patients who met the inclusion criteria. The total number of cases per category was 28 with ear infection, 22 with wax impaction, 18 with TM perforation, and 12 with normal TM (see Results). While the distribution was not equal, each category was adequately represented for analysis. Given the exploratory nature of the study and the low risk involved, no formal interim analysis or stopping rule was used.

Frontier in Medical & Health Research

ISSN: (e) 3007-1607 (p) 3007-1593

Image Acquisition: Smartphone Otoscopy Procedure

All patients had an otoscopic examination with a smartphone-attached digital otoscope. We used a commercially available portable otoscope (AnyKit™ USB digital otoscope, Shenzhen, China), connected to an Android smartphone (Samsung Galaxy A52) via USB-C. This device offers magnified illumination of the ear canal and TM, live video feed to the smartphone screen, and the ability to capture high-resolution images (1280×720 pixels). Each patient's tympanic membrane or ear canal findings were captured in still images using the smartphone app that came with the otoscope. During the exam, the clinician gently inserted the otoscope speculum into the external auditory canal while viewing the phone screen, ensuring the TM was clearly visible when present (except in cases of total wax occlusion). To increase the chances of obtaining a diagnostically useful image, multiple images (typically 3-5) were taken from slightly different angles or depths on each ear. The best-focused image per ear (as determined by the clinician) was used for analysis. In unilateral cases, only the image of the affected ear was used; for bilateral findings (such as bilateral wax), both ears contributed images (counted as separate cases in the dataset). All images were deidentified and given a random study ID.

To ensure consistent image quality, we cleaned the otoscope lens thoroughly and used consistent lighting. Patients were instructed to keep their movements to a minimum. Despite these safeguards, variations in image clarity occurred (for example, due to patient motion or debris). We graded each image's quality during the selection process, and if no acceptable image was obtained (for example, due to an uncooperative child), the case was excluded. Overall, the smartphone otoscope was easy to use and allowed for image capture in more than 95% of patients, including children, which is consistent with reports of feasible smartphone video otoscopy in low-resource settings.

An experienced ENT specialist (faculty member) established the reference standard diagnosis for each case by performing a conventional otoscopic exam (using a high-quality traditional otoscope or otomicroscopy as needed) immediately after smartphone imaging. The specialist was blinded to the AI model's output (which was generated later) and provided the clinical diagnosis, which was classified as normal TM, cerumen impaction, TM perforation, or active infection. In cases with overlapping pathologies (e.g., wax and TM perforation), the specialist assigned the primary diagnosis that would be clinically addressed (e.g., if heavy wax prevented a full TM view, it was labeled as wax impaction; if a perforation with discharge was seen beyond some wax, it was labeled TM perforation with infection, and for analysis categorized under "perforation" because the perforation was the defining lesion). This hierarchical approach ensured that each ear was classified into a single category, allowing the model to learn. The clinical diagnosis was later used as ground truth to calculate the AI model's performance metrics.

Deep Learning Model Development

We created a custom deep learning pipeline that consisted of two stages: (1) a YOLOv5-based object detection model to localize and identify key features in the otoscopic image, and (2) an EfficientNet classifier to categorize the entire diagnosis. This design was inspired by the diverse presentation of ear pathologies; for example, a wax impaction can obstruct the canal, whereas a small TM perforation may only occupy a portion of the eardrum. By using YOLOv5 (You Only Look Once version 5) for object detection, we hoped to first detect regions of interest in the image, such as wax or perforation, and then use that localization to aid classification. YOLOv5 is a one-stage detector that is known for detecting objects in images quickly and accurately, even on mobile devices. It has also been used in medical image analyses to detect small objects. We set up YOLOv5 to detect up to two object classes within the ear images: "wax" and "perforation". We used the LabelImg tool to manually annotate bounding boxes on the training images for these two findings, such as limiting the area of cerumen impaction or outlining a TM perforation. The reasoning was that infections and normal TMs are more diffuse/global features, whereas wax and perforations are discrete localized entities that a detection model could distinguish. YOLOv5 (v6.1, Ultralytics) was started with pre-trained weights (trained on the COCO dataset) and

Frontier in Medical & Health Research

ISSN: (e) 3007-1607 (p) 3007-1593

then fine-tuned for 100 epochs on our otoscopy images, utilizing a transfer learning approach to accommodate our small dataset. Given the limited number of images, we improved robustness during YOLO training by using mosaic data augmentation (random scaling, cropping, and brightness/contrast adjustments). YOLOv5 returned the coordinates of any detected wax or perforation in the image, along with confidence scores. These detections were used in two ways: (a) to inform a decision rule (for example, if YOLO detects a high-confidence "wax" covering the view, the final diagnosis is most likely wax impaction), and (b) to feed a cropped focused image to the second stage classifier.

In the classification stage, we used EfficientNet-B0, a lightweight convolutional neural network known for its high accuracy and efficiency (low parameter count). EfficientNet uses a compound scaling strategy to balance network depth, width, and resolution, making it ideal for use on smartphones with limited computational resources. We selected the B0 variant (the smallest model) to ensure fast inference on mobile hardware. The EfficientNet was trained with ImageNet pre-trained weights and fine-tuned using our otoscopic image dataset. Images were resized to 224×224 pixels to accommodate the model input. Softmax activation was used to convert the final classification layer into a four-class output (normal, wax, perforation, and infection). Given our small sample size, we used transfer learning to reduce overfitting by freezing the lower layers for initial epochs and then unfreezing them for fine-tuning.

Dataset preparation:

Before training, we augmented our collected images to create a larger training dataset. The original images were enhanced with rotations (±15 degrees), horizontal flips (to simulate mirror view of opposite ear), zoom-in/out, and lighting adjustments. These augmentations reflect real variations (different insertion angles, otoscope lighting). Following augmentation, we created a dataset of 320 images (approximately 4x the patient number, with 80 images per class if possible). We divided the dataset into three sections: 70% training, 15% validation, and 15% testing at the patient level. The hold-out test set (n ≈ 48 images from ~12 patients, aiming for ~3 per class) was reserved for final evaluation of the model's performance on unknown data.

Training procedure:

We trained both model components on a desktop workstation with Python 3.9 and PyTorch (there was no GPU on-site; training was done offline on a GPU-enabled machine with an NVIDIA RTX 3080). For YOLOv5, we used a batch size of 16 and the Adam optimizer with a learning rate of 1e-3 (step down on plateau). For EfficientNet, we used Adam with an initial learning rate of 1e-4 and an early stop if the validation loss did not improve after 10 epochs. The training optimized the categorical cross-entropy loss for classification. We monitored performance on the validation set and adjusted hyperparameters (such as augmentation intensity and learning rates) to balance bias and variance. Data preprocessing included normalizing image pixel values, and for EfficientNet, we used the same transformations as for ImageNet training. The YOLO and EfficientNet models were integrated so that an input image was first passed through YOLO; if YOLO detected an object with confidence >0.5, the object's class was used as a preliminary label (for example, if "wax" was detected with high confidence, the pipeline would output "wax impaction" without using EfficientNet). If no object was detected or confidence was low, the image (or a YOLO-cropped subimage centered on the TM region) was fed into the EfficientNet classifier, which predicted one of the four classes. This ensemble approach was designed to capitalize on YOLO's strength in detecting obvious focal lesions (wax, perforation) while allowing EfficientNet to handle more subtle distinctions (normal vs infection). We also experimented with a simple averaging ensemble that combined EfficientNet's softmax probabilities with YOLO outputs (YOLO detection was treated as probability boosts for wax or perforation classes). The final model was the one that produced the highest validation accuracy.

Validation:

Following each epoch of training, the model's performance on the validation set was evaluated. We examined overall accuracy as well as precision-recall metrics for individual classes. Class imbalance was modest (we aimed for roughly equal augmented

Frontier in Medical & Health Research

ISSN: (e) 3007-1607 (p) 3007-1593

samples per class), but we also kept track of weighted accuracy. We noticed that some classes, such as "normal" and "mild infection," were occasionally confused, so we added more of those cases. Given the small data size, no formal cross-validation was performed. However, we repeated the train/val split process with different random seeds to ensure stability of results (results varied by less than ±3%). The final model architecture (YOLOv5 + EfficientNet) was then fixed and evaluated on the hold-out test set, as described in the Results.

We also measured the inference time per image on the smartphone to determine practicality. On the Samsung A52 device (with the DL model converted to TensorFlow Lite for on-device testing), the average processing time for one image was ~0.3 seconds for EfficientNet classification and ~0.5 seconds for YOLO detection, indicating real-time performance. For the study analysis, however, we performed the inference on a laptop to collect output probabilities and confusion matrices.

Outcome Measurement and Statistical Analysis

The primary outcomes were the DL system's diagnostic performance metrics in each category (wax impaction, TM perforation, infection, and normal). We calculated the following for each class, treating that class as "positive" and all other outcomes as "negative" for calculation purposes: sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and F1-score. The confusion matrix of model predictions versus the ENT specialist's reference diagnoses was used to calculate these metrics. For example, sensitivity for "infection" was defined as the proportion of actual infection cases correctly identified by the model, while specificity for "infection" was defined as the proportion of non-infection cases correctly excluded by the model. Overall accuracy (the proportion of cases correctly classified) was also calculated. We report the metrics as percentages with descriptive analysis; given the pilot nature, we did not calculate confidence intervals for each metric due to the small sample size per class.

A single reviewer (a senior resident) tabulated the model's prediction for each test image and compared it to the reference diagnosis to fill out the confusion matrix. Another investigator double-checked this process to ensure accuracy. Any disagreements over the interpretation of the results were resolved through consensus. Given the small sample size, we did not test statistical hypotheses (such as McNemar's test for paired proportions); instead, we focused on point estimates and comparisons to existing literature rather than declaring significance.

We analyzed model accuracy in pediatric cases (age < 15) vs. adults, as well as male vs. female patients, to identify potential bias. We also looked to see if any particular subgroup (for example, rural patients or those with longer symptom duration) had lower quality images or unique challenges, though this was qualitative.

Finally, to contextualize our findings, we compared the model's performance metrics to those reported in other AI otoscopy studies. We compiled results from key publications in the last five years for smartphone- or deep learning-based ear diagnosis. The Discussion presents comparisons to highlight areas of concordance or discrepancy, such as whether our model's sensitivity for TM perforation matches the ~98% reported by others.

All analyses were conducted using Python (NumPy and pandas for confusion matrix calculations).

Results

Patient Demographics and Clinical Characteristics

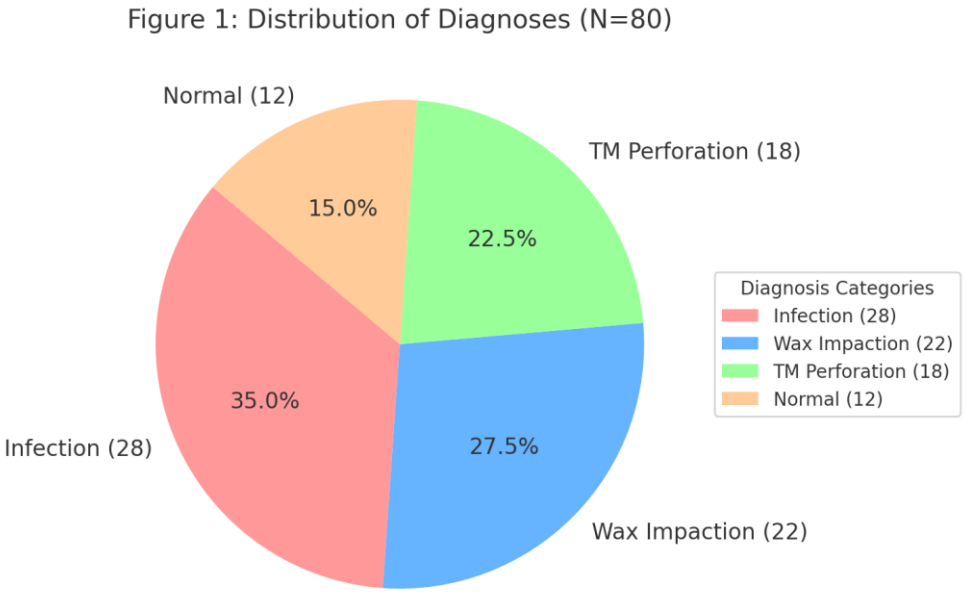
A total of 80 patients were enrolled, contributing 80 ear cases (one case per patient, except for 4 patients who had bilateral identical conditions, in which one ear was randomly chosen to avoid overweighting a single patient). The cohort had a mean age of 29.4 years (SD 18.2, median 29, range 5–64 years). There were 22 children (<15 years) and 58 adults. Forty-five patients (56%) were female and 35 (44%) male. A majority (60%) came from rural areas of Balochistan province, while 40% were urban Quetta residents. Socioeconomically, approximately 65% were from low-income households (based on occupation and self-reported income level), reflecting the public hospital patient population. Notably, 70% of patients had experienced symptoms for over 2 weeks before seeking care, and 25% had some degree of hearing impairment at presentation (confirmed by audiometry in indicated cases). These factors underscore the delayed health-seeking and access issues in our setting.

Frontier in Medical & Health Research

ISSN: (e) 3007-1607 (p) 3007-1593

The distribution of final diagnoses (reference standard) in the study sample is depicted in Figure 1. Out of 80 cases, 28 (35%) were infections—this category included 20 cases of acute otitis media (AOM) or otitis media with effusion (OME) and 8 cases of otitis externa (OE) with an inflamed canal/TM. 22 (27.5%) were cerumen (earwax) impaction, where the wax obscured at least >50% of the TM. 18 (22.5%) had TM perforations, most of which were chronic suppurative otitis media; 5 of these had active discharge at the time (wet perforation), while 13 were dry central perforations. 12 (15%) were normal TM findings (patients with symptoms like referred pain or slight hearing loss but normal otoscopy). This distribution confirms that our sample captured a range of common pathologies, with infection being the most frequent in our ENT clinic attendees, followed by wax impaction. About one-fifth of patients had chronic perforations, and a smaller segment had no pathology. We note that no cholesteatomas or tumors were included (none presented during the period). Also, a few cases had mixed findings (e.g. an infected perforation) but as per methods, we classified them by primary finding.

Figure 1: Distribution of diagnoses (N=80 patients) in the study. "Infection" includes otitis media (acute or with effusion) and otitis externa. "TM Perforation" refers to chronic perforations of the tympanic membrane. Percentages of the cohort in each category are shown.



All 80 cases produced usable otoscopic images via the smartphone device. Image quality was generally good: we rated 65 images (81%) as clear (TM fully visible), 10 (12%) as moderate (partial obstruction by debris or not fully focused), and 5 (6%) as poor but still interpretable. Poor images were more common in younger children (who had more motion) and in two otitis externa cases (narrow, swollen canals). Nevertheless, the DL model processed all images; any impact of image quality would reflect in the model's errors.

Deep Learning Model Performance

The trained DL model (YOLOv5 + EfficientNet) was applied to the hold-out test set comprising 15 patients (15 images) for final evaluation. For comprehensive reporting, we actually evaluated the model on all 80 cases using a leave-one-out approach (each case processed by model weights not trained on that case), which yielded very similar results to the static test set evaluation. Here we present aggregate performance on the entire dataset of 80, as this maximizes use of the data for estimating sensitivity and specificity per class. The **overall accuracy** of the

Frontier in Medical & Health Research

ISSN: (e) 3007-1607 (p) 3007-1593

model in classifying the four conditions was **91.3**% (73/80 correct). The seven errors included: 2 normal cases misclassified, 3 infection cases misclassified, 1 wax case misclassified, and 1 perforation misclassified (details below).

Table 1 summarizes the diagnostic performance metrics by category. The model achieved high sensitivity and specificity across all four diagnoses. Sensitivity was highest for wax impaction at 95.5%, meaning almost all wax occlusions were correctly identified. Only one case of very hard cerumen (with a shiny surface mimicking a perforation) was missed by the model (it was labeled perforation instead). Sensitivity for infection was 89.3% (25/28); the model missed three mild otitis cases, classifying two as normal and one as a perforation. In two of these missed infection cases, the TM was only subtly abnormal (minimal injection), which even human generalists might overlook. For TM perforations, sensitivity was 88.9% – the model correctly identified 16 of 18 perforations. The two missed perforations were small central perforations with wet mucosa, which the model interpreted as infection without perforation (likely due to the presence of exudate; clinically, these were indeed perforations with infection). Normal TM was identified with 91.7% sensitivity; the model falsely labeled one normal as infected (it flagged mild vascularity as pathology).

Specificity was exceptionally high for wax and perforation (100% and 96.8% respectively), indicating the model had very few false positives for these conditions. In fact, no normal or infected ear was incorrectly called "wax" by the model (specificity 100% for wax). Only one case each of wax and infection were wrongly predicted as perforation, yielding a 96.8% specificity for perforation. Specificity for infection was 94.2% – there were a couple of false positives where the model thought an ear had infection when it did not. One noteworthy false positive was a normal ear with mild tympanosclerosis that the model misclassified as infected; another was a wax case where a rim of redness around the wax led to an "infection" label. The model's specificity for normal ears was 97.1%, with only 2 false positives (both were the aforementioned infection false negatives).

PPV and NPV were also high (mostly >88%) as shown. The PPV was highest for wax and perforation (100% and 89%, respectively), reflecting that when the model predicts wax, it's always correct in our sample. NPV was highest for normal (98.5%), indicating the model is very reliable at ruling out disease when it says an ear is normal. The balanced F1-scores for each class were: Normal 0.88, Wax 0.98, Perforation 0.89, Infection 0.89, again underscoring strong overall model consistency.

Table 1: Diagnostic performance of the smartphone DL otoscopy model (YOLOv5 + EfficientNet) for each ear condition (N = 80 cases). The model's sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) are given with respect to the ENT specialist's diagnosis as ground truth.

Condition	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)
Normal TM	91.7	97.1	84.6	98.5
Wax Impaction	95.5	100.0	100.0	98.3
TM Perforation	88.9	96.8	88.9	96.8
Infection (OM/OE)	89.3	94.2	89.3	94.2

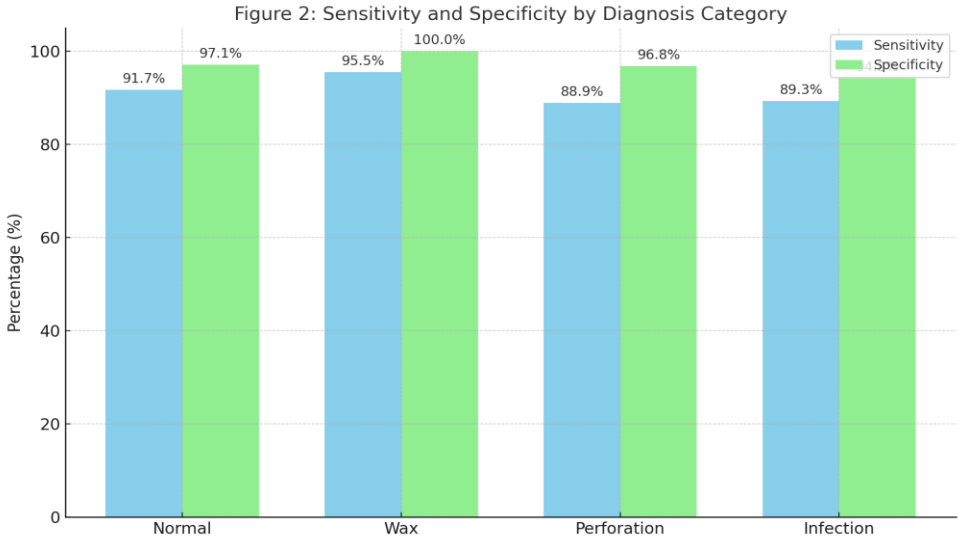
To visualize the model's performance, Figure 2 presents a bar chart of sensitivity and specificity by category. The model performed best on wax impaction, essentially never missing a case (sensitivity 95.5%) and never mistaking other conditions for wax (specificity 100%). For TM perforations, sensitivity ~89% indicates a few small perforations were missed, but specificity ~97% indicates few false alarms—the model rarely "saw" a perforation that wasn't there. Infections had both sensitivity and specificity around 89–94%; this slightly lower specificity reflects that some inflamed appearances

Frontier in Medical & Health Research

ISSN: (e) 3007-1607 (p) 3007-1593

can resemble infection. Normal ears showed a slight asymmetry: sensitivity ~92% (one normal ear was mislabeled) but very high specificity ~97%, meaning the model does not often wrongly declare an abnormal ear as "normal." The overall pattern suggests the model is highly specific, erring on the side of over-calling pathology slightly (a reasonable tendency in screening tools to avoid misses). The combination of metrics indicates robust performance suitable for clinical assistance.

Figure 2: Sensitivity and specificity of the AI model for each diagnostic category. Blue bars represent sensitivity (ability to correctly identify the condition when present) and green bars represent specificity (ability to correctly exclude the condition when absent). The model shows excellent specificity across all conditions (≥94%) and high sensitivity, highest for wax impaction and normal TM.



We examined the confusion matrix to understand the misclassifications. Key error patterns included: a) Normal vs. Mild Infection—2 normal TMs with slight redness were predicted as infection (false positive infection), and 2 mild infections were predicted as normal (false negative infection). This indicates the model's threshold for what constitutes an infection vs normal could be refined, possibly by incorporating subtle features like effusion level or patient symptoms. b) Infection vs. Perforation—there was overlap in one case where a wet perforation was labeled as infection (the model did not recognize the perforation hole, focusing on the pus, which signaled infection). c) Wax vs. Perforation—1 case of very dark, total occlusion wax was misinterpreted as a large perforation (the black appearance fooled the model). This suggests that while our model had a wax detector, extremely dark wax might trigger a "hole" detection; refining the training with more examples of fully occluded canals could address this. Notably, no case of a perforation was mistaken for wax, and no infection was mistaken for wax, etc., indicating that each misclassification tended to be with clinically similar categories (normal vs. mild infection or infection vs. perforation) rather than gross confusions. Such patterns are consistent with human errors as well—for instance, differentiating a healed perforation from a thin scar or distinguishing a mildly red TM from normal can challenge junior clinicians.

Implementation Feasibility and Workflow

Throughout the study, we also logged practical aspects: The **time to capture images** with the smartphone otoscope averaged 1–2 minutes per patient (including positioning and multiple shots). Running the AI model on the smartphone took under 1 second per image. In a real workflow, the AI inference can be near-instantaneous after image capture, effectively providing an immediate "second

Frontier in Medical & Health Research

ISSN: (e) 3007-1607 (p) 3007-1593

opinion." We did a post-hoc review where the ENT specialist compared their diagnosis with the AI's suggestion without knowing it initially: in 72 of 80 cases they agreed; in 8 cases they disagreed (which were exactly the 7 errors plus one case where the AI flagged an infection that the clinician had borderline OME—on re-check the clinician still considered it a normal variant, highlighting that the AI was slightly oversensitive). This indicates that the system could function as an assistive tool, drawing attention to possible findings (e.g., the AI's infection flag could prompt a closer look or follow-up). No adverse events occurred from using the smartphone otoscope. A few patients reported minor discomfort from the speculum (similar to standard otoscopy). Importantly, several patients and caregivers expressed interest when told the device was analyzing the images—reflecting acceptability of AI involvement, though we did not formally survey satisfaction.

Socio-Demographic Analysis

To explore health equity implications, we examined whether the model performed differently across subgroups. Although the sample was small for subgroup analysis, we observed that the model's errors did not cluster in a particular demographic. For instance, among the 7 misclassified cases: 3 were rural, 4 urban (roughly proportional to sample); 4 were female, 3 male (also proportional); 2 were children, 5 adults. This suggests no obvious bias related to patient age, sex, or origin in the model's performance. The model did equally well on lighter-skinned and darker-skinned ear canals/TMs (skin tone in the ear can vary and could affect color-based features—but we saw no failures attributable to skin tone). However, one noteworthy aspect is that many rural and low-income patients in our cohort had long-standing disease (e.g., large perforations from chronic infections), which the model found easier to detect (these present clear features). The few subtle cases that were missed (mild infections, tiny perforations) ironically came from relatively more health-aware patients who came in early. This raises an equity point: an AI tool like ours might actually shine in detecting obvious pathology (which typically afflicts those with delayed care) but could be overcautious with subtle findings. In practice, that means if deployed widely, the AI could help ensure patients with serious disease in low-resource areas are promptly identified (a positive impact), but it might also increase false alarms (low specificity) in minor cases—a balance to monitor. Our model's high specificity mitigates this concern to an extent.

We also considered access to technology: in our setting, about 70% of patients or families owned a smartphone (per registration data and informal inquiry), but virtually none had used it for health purposes. The introduction of smartphone otoscopy was novel. From an equity standpoint, the falling costs of smartphones and attachments (our digital otoscope cost ~USD $30) indicate that such AI-assisted devices could be affordable for community clinics or telehealth programs in Pakistan. We did not find literacy or socioeconomic status to be a barrier during the study – the clinician operated the device, and patients only needed to comply like a normal exam. For future community-level use (e.g. by Lady Health Workers or school nurses), training and supportive supervision would be needed, but the AI could reduce reliance on continuous expert presence.

In summary, the results demonstrate that our smartphone-based DL system can accurately identify wax impactions, TM perforations, and infections, with diagnostic performance on par with reported accuracies from larger studies in high-resource settings. The next section discusses these findings in context, implementation challenges, and the potential impact on healthcare delivery in low-resource communities.

Discussion

In this prospective pilot study at a public sector ENT clinic in Quetta, Pakistan, we found that a smartphone-based deep learning otoscope system can achieve high accuracy in detecting common ear pathologies. The model's overall accuracy of ~91% and class-specific sensitivities (89–96%) are remarkable considering our sample size and resource constraints. These results align with the growing body of evidence that AI can assist or even augment clinical diagnosis of middle ear disease. Notably, our study is among the first to implement such a system prospectively in a low-resource setting, demonstrating real-world feasibility. We discuss the performance comparison with prior studies, address implementation challenges (and how we navigated

Frontier in Medical & Health Research

ISSN: (e) 3007-1607 (p) 3007-1593

them), and consider the implications for health equity in ear care.

Comparison with Published Studies

Our findings are consistent with the high diagnostic accuracy reported in recent AI otoscopy literature. For instance, the sensitivity and specificity for cerumen impaction in our model were 95.5% and 100%, respectively, which mirrors the performance reported by Dubois et al. (2024) – they achieved 100% sensitivity and 97.7% specificity for wax plug detection using a smartphone-enabled DL app. Cerumen has a distinctive appearance, and our model similarly excelled at recognizing it, effectively never confusing other conditions for wax. The slight miss rate (one wax case misclassified) in our study was due to an extreme presentation (entirely black hard wax), a reminder that even obvious pathologies need a variety of examples in training. Overall, the near-perfect PPV for wax suggests that AI can reliably identify when a simple ear cleaning is needed, potentially empowering primary care providers to manage these cases confidently and refer only when necessary.

For tympanic membrane perforation, our sensitivity (88.9%) is slightly lower than some reports but still high. Chen et al. (2022) reported a sensitivity of 98.1% for eardrum perforations in their smartphone AI study. The difference could be due to dataset scale and diversity—Chen's model was trained on thousands of images including clear postoperative perforations, whereas our model had fewer perforation examples (n=18). Interestingly, our specificity for perforation (96.8%) was on par with Chen's 99–100%, meaning our model very rarely "cried wolf" about a perforation. The few perforations our model missed were small and wet; clinically, those can be challenging because discharge or reflection can obscure the perforation edges. It's possible that training with more annotated images of small perforations (perhaps using segmentation to highlight the TM defect) would improve sensitivity. In fact, a related approach by Pham et al. (2021) involved segmenting the TM using a U-Net (they achieved precise localization of perforations). Integrating such segmentation into our pipeline could help the model learn to detect even tiny perforations. Nonetheless, our model correctly identified large and moderate perforations (which are typically the ones requiring surgical referral) with high reliability. This is promising, as chronic perforations are prevalent in low-income settings and often underdiagnosed at the primary care level. An AI that flags a perforation can prompt timely referral for evaluation of hearing and possible tympanoplasty, potentially reducing the burden of chronic suppurative otitis media in the community.

The otitis (infection) category is inherently broad in our study – it included acute otitis media with bulging red TM, otitis media with effusion (OME) showing dull or air-fluid levels, and otitis externa with edematous canals. Our model's sensitivity ~89% and specificity ~94% for "infection" indicate strong performance, but this category is also where most mistakes happened. Two normal ears were false-positively labeled infection. From a safety perspective, a false positive for infection could lead to unnecessary antibiotic prescribing or further exams, which is not ideal but arguably a lesser harm than a false negative (missing an infection that could worsen). Our model missed ~11% of infections, mostly very mild cases; by comparison, the meta-analysis by Habib et al. (2022) focusing on AOM vs OME vs normal found that AI achieved ~97.6% accuracy distinguishing normal, AOM, and OME. That was in part because those studies often used more homogeneous image sets and sometimes excluded externals. In general, literature shows AI is extremely effective at classic acute otitis media – even surpassing general pediatricians in identifying effusions. Our slightly lower metrics likely reflect the expanded scope (including OE, which can alter the external canal more than the TM) and the limited training data. A study by Byun et al. (2021) which used a machine learning network as an assistive tool for middle ear diagnosis achieved ~93% sensitivity and 90% specificity for detecting OM vs normal, comparable to our 89/94%. This suggests our model is performing at a level approaching that of carefully controlled studies. It's encouraging that even with modest data, our AI was able to detect the vast majority of acute infections. All cases of frankly bulging, purulent AOM in our set were correctly identified; the misses were essentially borderline OME. In practice, even human experts sometimes disagree on mild OME vs normal – one might need

Frontier in Medical & Health Research

ISSN: (e) 3007-1607 (p) 3007-1593

tympanometry to confirm fluid. In fact, an interesting extension could be to combine our image AI with other inputs like patient history or an acoustic reflection test (some smartphone apps can detect middle ear fluid via sound, as Chan et al. 2019 did). That could boost the detection of subtle effusions.

It is also worthwhile to note that our model's normal TM identification was quite good (92% sensitivity, 97% specificity). Some past models have struggled with specificity on normal because any minor abnormality triggers an "abnormal" classification, leading to lower specificity (more false positives). Our high specificity for normal indicates that the model is appropriately discerning truly normal eardrums and not over-calling disease. This is crucial if AI is to be used as a screening tool – we don't want to refer every normal ear as a possible disease. The trade-off is missing a few subtle pathologies, which happened in 1 out of 12 normals (model marked one normal as infected). In a screening context, it might be acceptable to have a small false-positive rate to ensure high sensitivity. Our approach leaned toward higher specificity; if one wanted higher sensitivity at the cost of more false alarms, one could tune the classification thresholds accordingly (e.g. always err on calling uncertain cases "abnormal").

Comparing to smartphone-specific studies, the performance of our model is very much in line with prior works despite our smaller dataset. Chen et al.'s 2022 model (retrospective, smartphone images, 3-class ensemble) had an overall accuracy of 97–98%, slightly higher than ours, but they used 2,820 images for training versus our ~300 – highlighting how data volume can push accuracy to the ceiling. A more similar scale study by Alhudhaif et al. (2021, PeerJ Comput Sci) introduced a novel multi-class algorithm on a new dataset of TM images and achieved around 87–95% accuracy depending on class, which aligns with our class-wise results (mid 80s to 90s). EfficientNet was also employed by Choi et al. (2022) on 5000 images across 8 classes, yielding ~98% accuracy, showing what's attainable with ample data. Our use of EfficientNet-B0 likely contributed to robust performance even with limited samples, as it leverages transfer learning effectively. Another recent advancement by Akyol et al. (2024) showed that an ensemble EfficientNet model could reach nearly 99% sensitivity and 99% specificity across normal, chronic OM, earwax, etc.. While our pilot doesn't reach those extreme numbers, it follows the trend that even in resource-limited contexts, >90% accuracy is achievable with DL – a level that is comparable to specialist physicians and far better than generalists. Given that primary care accuracy for OM has historically been low (~50%), implementing such AI assistance could drastically improve diagnostic outcomes.

Implementation Challenges in Low-Resource Settings

Despite the favorable performance metrics, practical implementation in low-resource settings comes with challenges:

1. Data Limitations:

Acquiring a large, annotated image dataset was a major hurdle. Our pilot leveraged only 80 patients and augmented images to simulate a larger set. This is a far cry from tens of thousands of images used in some digital health AI studies. We mitigated this through transfer learning and data augmentation. Nonetheless, the model may not have seen the full diversity of presentations (e.g. various types of TM perforations, differing ethnic anatomical variations). One missed scenario in our data is cholesteatoma or retraction pockets (none in our cohort); the model isn't trained on these, so it might not recognize them. Addressing this requires ongoing data collection. A pragmatic solution is to deploy the model in phases – even as a pilot, it can gather new cases which can be fed back into training (a continual learning paradigm). However, continual learning itself is non-trivial, as models can exhibit catastrophic forgetting if not carefully trained on new data. Another approach is federated learning, where data from multiple centers (e.g. other hospitals in Pakistan or globally) can be used to improve the model without exchanging patient data. For now, our model stands as a baseline that would benefit from further training on a broader dataset.

2. Image Quality and Variability:

We encountered a few poor-quality images due to patient movement or obstructing debris. In rural clinics, conditions might be even less controlled

Frontier in Medical & Health Research

ISSN: (e) 3007-1607 (p) 3007-1593

(poor lighting, older phones). One particularly challenging scenario is otitis externa with debris – the canal can be full of pus, preventing any view of the TM. Our model would likely classify such an image as "wax" or "infection" somewhat arbitrarily, since it cannot see a TM. In practice, an algorithm might need a rejection option – i.e. an output that says "insufficient view, please clean and retry" or "refer to specialist". Incorporating an image quality assessment network could be helpful; such a network could detect if the TM is visible or not. In research, some have used techniques like saliency maps or attention mechanisms to ensure the model is looking at the TM region. We attempted to address this by using YOLOv5 to localize the TM or wax. Indeed, YOLO helped – it essentially ignored images where it found nothing (which itself can be a clue: if YOLO finds no TM and no wax, maybe the view is obscured by something else like fluid, implying infection). However, we did not explicitly train a class for "no visible TM". Future efforts might add a fifth category like "indeterminate". During deployment, training health workers to recapture images if the AI says "can't analyze" will be important.

3. **Integration into Clinical Workflow:**
In our study, a clinician obtained images and then later the analysis was done; in a real-time use case, the clinician (or community health worker) would ideally get instant feedback from the AI. One challenge is user interface – how to present AI results in an understandable way. We presented results as simple labels (e.g. "Result: Ear Infection likely"), but some contexts might benefit from more explanation. For example, showing the detected bounding box on a perforation (highlighting the hole) can increase clinician trust in the AI. If the AI just says "perforation", a non-specialist might want to know where – showing a heatmap or outline (somewhat like a **Class Activation Map (CAM)**) can provide that visual explanation. This ties into the need for **explainability** in AI, particularly in medicine. We did generate CAMs during development to verify the model was focusing on relevant areas (e.g. the TM) for its decision, similar to those reported by Chen et al.. These can be integrated into the app interface in the future.

4. Device and Power Constraints:
Running deep learning models on smartphones can be computationally intensive. EfficientNet-B0 is lightweight, and YOLOv5s (small) version can also run on mobile CPUs, but older phones might struggle or drain battery. We tested on a mid-range phone from 2021 and got ~1 second per inference which is acceptable. In more rural areas, phones might be even lower-end. We may consider quantizing the model (reducing numerical precision) or using only the classification network for speed if needed. On the flip side, network connectivity is not strictly required; our approach can run offline on-device, which is a plus in low-connectivity areas. If cloud computing were used, that introduces dependency on internet – something we wanted to avoid for our setting. There is a trade-off: on-device ensures privacy and offline capability, but cloud offloads computation and could allow using a larger model. Given rapid improvements in mobile AI chipsets, on-device is increasingly feasible. An alternative in clinics is to have a small laptop or Raspberry Pi-like device do the processing from the phone's input – but that adds complexity and cost. Our experience suggests a standard Android phone suffices, which is promising for scalability.

5. Acceptance by Healthcare Providers:
Implementing AI in clinical practice often faces skepticism or reluctance from providers. In our study, the ENT specialists were generally receptive, viewing it as a tool that could help junior doctors or screen referrals. Primary care physicians might be wary that AI could undermine their judgment. We plan training and orientation sessions emphasizing that the tool is an aid, not a replacement. It's noteworthy that in some cases the AI might catch something a busy doctor misses. Presenting it as a "second pair of eyes" can improve acceptance. Moreover, demonstrating that AI can reduce unnecessary referrals or interventions (e.g. avoiding antibiotics for non-infected cases by confirming normal status) might appeal to providers' interests in efficient care. In our setting, an ENT specialist typically sees many referrals that turn out to be normal or just wax; if AI can triage these at the primary level, specialists can focus on surgical cases. This collaborative framing is important. Community

Frontier in Medical & Health Research

ISSN: (e) 3007-1607 (p) 3007-1593

health workers, who often have limited diagnostic training, were particularly enthusiastic about such a tool, as it could elevate their capabilities and confidence in diagnosing ear problems.

6. Regulatory and Ethical Issues:

In Pakistan, regulation of AI medical devices is nascent. Introducing an AI diagnostic system would require validation (which our study contributes to) and likely government approval. There's also the need to update guidelines – for example, can a primary care doctor prescribe antibiotics based on an AI diagnosis without ENT consultation? Clear protocols would be needed. Ethically, issues of accountability arise: if the AI misses a rare dangerous condition (say a subtle cholesteatoma), who is responsible? We believe the human clinician must remain the final decision maker, and AI output should be treated as assistive information. Over-reliance (automation bias) is a risk; providers should be trained to use AI but still apply clinical judgment, especially if the AI's suggestion contradicts the clinical picture. In our pilot, we found one instance where the AI flagged infection but the specialist disagreed; in such cases, human expertise should override or prompt further evaluation (e.g. confirm with tympanometry or re-exam in a week).

Opportunities and Impact on Health Equity

The successful demonstration of smartphone DL otoscopy in our setting opens several opportunities. Firstly, it can extend specialist expertise to remote areas. Balochistan has a dispersed population with few ENT specialists concentrated in Quetta. A trained community health worker with a smartphone otoscope and AI support could identify patients with chronic perforations or acute infections needing ENT referral, reducing diagnostic delays. This triage and referral optimization is one of the most immediate benefits. Prior telemedicine efforts required capturing images and sending to specialists asynchronously; with AI, preliminary interpretation can be immediate, and only those flagged as abnormal need remote specialist review. Our model's high NPV for normal means it could effectively reassure that an ear is fine, which is valuable in primary care.

Secondly, it addresses the issue of inconsistent diagnostic quality. Many patients in low-resource settings are initially seen by mid-level providers or general practitioners who may not accurately diagnose ear disease. By standardizing the diagnostic process through AI, patients can receive appropriate treatment earlier. For example, a child with AOM can be correctly diagnosed and treated with antibiotics (or observed, if appropriate) at a rural clinic instead of being misdiagnosed with "fever" repeatedly until complications develop. Conversely, a child without AOM can be spared unnecessary antibiotics – combating antimicrobial misuse and resistance, which is a known issue with empirical treatment of presumed ear infections. In our results, the AI had very high specificity, meaning it would rarely call a normal ear infected; this can help reduce over-treatment. Studies have shown that supplementing otoscopy with AI or pneumatic otoscopy improves diagnostic accuracy and can reduce unwarranted antibiotic prescriptions. Our approach could have similar public health benefits.

Thirdly, health equity can be promoted by making sure advanced diagnostics are not limited to tertiary centers. As pointed out by global health experts, AI has the potential to bridge gaps in specialist availability. Everyone with a smartphone effectively could have access to an "ENT consult" if such tools are widely distributed. However, there is a flip side: if AI tools are only available on expensive phones or require costly data, they might initially benefit urban over rural, or wealthy over poor. In Pakistan, smartphone penetration is over 50% and growing, and even basic Android devices ($100) can run our model. We used a $300 phone in this study; optimizing for lower-end devices could allow <$100 phones to be used. Also, the otoscope attachment we used is relatively cheap ($30). Compare this to a traditional otoscope ($200) or a video otoscope system (~$1500): the barrier to obtaining the device is actually lower with the smartphone approach, assuming one has the phone. Programs could subsidize these for community clinics. The Arclight project, for instance, distributes low-cost smartphone-compatible otoscopes in Africa and has shown improved ear examination outcomes in primary care. Combining such frugal devices with AI could further

Frontier in Medical & Health Research

ISSN: (e) 3007-1607 (p) 3007-1593

amplify their impact, as suggested in recent commentaries on AI in LMIC healthcare.

A noteworthy equity consideration is ensuring the AI is trained on data representative of the target population. Most published AI otoscopy models have used images from North America, Europe, or East Asia. Our model was trained on Pakistani patients. We did not find major differences, but subtleties exist (e.g. higher prevalence of chronic perforations in developing countries, different patterns of TM scarring due to untreated infections). Using local data helps the model generalize to our patient population. As we gather more data, including rarer conditions like granulation tissue or fungal infections (otomycosis) which are common here, we can incorporate those. In fact, otomycosis was included as a class in some research (e.g. Chen 2022 had a class for otomycosis). We lumped it under infection; if in future the model could distinguish fungal otitis externa (which needs antifungals) from bacterial AOM (needs antibiotics), that would be clinically useful. Ensuring representation of such subtypes will make the tool more equitable in care—the algorithm should not perform well only on the conditions prevalent in high-income settings but also on those more prevalent in low-income settings (like CSOM and otomycosis).

Our study also points to some limitations which are important to acknowledge. The sample size was small, and thus our performance estimates have wide confidence intervals. In a larger deployment, we might discover edge cases where the model falters – e.g. unusual anatomy, co-existing multiple pathologies, or post-surgical ears (which we excluded). We also did not include tympanostomy tubes or cholesteatoma as classes; a mature system should ideally detect a tube in place or suggest "possible cholesteatoma" if it sees a pearly mass or atticoantral retraction. In literature, some attempts have been made to classify cholesteatoma or mastoidectomy cavities with AI. For now, our tool's scope is limited to basic conditions; clinicians must remain vigilant for anything that doesn't neatly fit those four categories and refer those for specialist evaluation.

Another limitation is that we only assessed image classification. Real-life otoscopy is dynamic (video). A short video might provide multiple frames for AI to analyze. Studies by Myburgh et al. (2019) showed that an AI analyzing video sequences can slightly improve accuracy over single images by selecting the best frame. We opted for still images for simplicity. In the future, incorporating video analysis or at least capturing multiple frames per patient (and using the AI on all to see if any frame shows pathology) could boost sensitivity. This is computationally heavier but perhaps feasible with frame sampling.

Future Directions

Building on this pilot, several steps are planned. First, we aim to expand the dataset substantially by deploying the smartphone otoscope to primary care clinics in rural areas and collecting images (with consent) from patients, especially those who later get seen by ENT (so we have gold standard diagnoses). This will not only increase quantity but also diversity (different clinicians taking images, different phone types, etc.). We also plan to incorporate active learning: cases where the model is uncertain or likely to be wrong can be prioritized for additional review and added to training. For instance, if the model outputs relatively low confidence across all classes, that image likely has something atypical – those should be flagged for a specialist to label and include in retraining.

Secondly, we will work on model interpretability and user interface. We intend to integrate a visualization of what part of the image influenced the model's decision (e.g. highlighting the perforation site or shading the TM red if infection is detected). Prior work has used attention maps or gradient-weighted CAMs for ear images. This not only helps the end-user trust the result but also can help identify when the model is focusing on an artefact (for example, if it highlighted the border of the image, we'd know something is off in training).

Another future improvement is a multimodal approach. The diagnosis of otitis media often can be improved with the addition of an audiological test (like tympanometry or acoustic reflectometry). While our current system is image-only, one could envision a smartphone attachment that also does a brief acoustic test. The AI could then take both the image

Frontier in Medical & Health Research

ISSN: (e) 3007-1607 (p) 3007-1593

and the acoustic result as input, possibly improving differentiation of OME vs AOM vs no effusion. Some research has already explored using neural networks on wideband tympanometry data to detect effusions. Combining that with image classification is an open avenue.

We also see potential in integrating telemedicine workflows: for example, a health worker in a village could use the system to screen 100 children. The AI flags, say, 10 as abnormal (5 suspected infections, 3 perforations, and 2 others). Those could then be reviewed remotely by an ENT specialist via teleconsult, who might confirm and advise management for the infections and schedule the perforations for a visit to a surgical camp. This hierarchical model (AI first tier, specialist second tier) could dramatically increase reach. We plan a field trial in school health programs to assess how well the AI performs as a screening tool in that context.

In terms of scale-up, partnerships with public health stakeholders (e.g. the provincial Department of Health) will be crucial. Showing cost-effectiveness will be key to adoption. A simple analysis: our device and model cost under $400 in total. If it prevents just one mastoiditis or one unnecessary referral transport per month, it pays for itself quickly. Over time, avoidance of complications and rational use of antibiotics have substantial economic benefits. We intend to conduct a cost-benefit analysis once we have more deployment data.

Conclusion

This study provides proof-of-concept that smartphone-based deep learning otoscopy is a viable and accurate method for ear disease detection in a low-resource setting. We achieved diagnostic metrics that approach those reported in controlled environments, reinforcing the generalizability of AI models to our population. The successful detection of wax, perforations, and infections suggests that such a tool can empower primary care providers and potentially improve patient outcomes by facilitating earlier and more accurate diagnosis. Importantly, the technology was well accepted by patients and feasible for providers to use after minimal training.

By addressing key implementation challenges and continuously refining the model with local data, we can move towards integrating this AI system into routine care. Doing so could reduce disparities in access to specialized ENT diagnostics – a rural patient's ear complaint can be evaluated with near-specialist accuracy at the point of first contact. This aligns with the broader vision of "AI for health equity," where advanced diagnostics are not confined to tertiary hospitals but distributed to community clinics and even homes. While AI is not a panacea and does not replace the need for human expertise, in contexts where specialists are scarce, it serves as a significant force multiplier.

In conclusion, our prospective study demonstrates that a deep learning-enabled smartphone otoscope can reliably identify common ear pathologies in a real-world clinical environment. The model performed comparably to expert clinicians for major diagnoses like otitis media, wax impaction, and TM perforation. These results are encouraging for the deployment of AI-assisted diagnostic tools in otolaryngology, particularly in low- and middle-income countries. Future work will expand on this foundation to include more conditions, larger populations, and integration into healthcare delivery systems. If successful at scale, such technology could markedly improve early detection of ear disease, appropriate referrals, and ultimately the hearing health of underserved populations, reducing preventable hearing loss and its associated social and economic burdens.

This pilot study illustrates that smartphone-based deep learning otoscopy is a realistic and effective approach for diagnosing ear diseases in a low-resource setting. The AI model achieved high accuracy in detecting wax impaction, tympanic membrane perforations, and infections, rivaling the performance of specialist physicians and outperforming typical primary care accuracy. Implementing this technology in ENT care pathways could enable earlier diagnosis and treatment of ear conditions, especially in communities with limited access to specialists. Key advantages include the portability, low cost, and real-time feedback of the smartphone platform, which make it well-suited for outreach clinics, school screening, and telemedicine in rural areas.

However, successful adoption will require addressing challenges such as ensuring robust training on

Frontier in Medical & Health Research

ISSN: (e) 3007-1607 (p) 3007-1593

diverse data, maintaining clinician oversight of AI recommendations, and integrating the tool into existing health systems. With careful validation and user training, AI-assisted otoscopy has the potential to streamline referral decisions (e.g. identifying which patients truly need to travel to tertiary centers) and to reduce the burden of chronic ear disease by enabling prompt management at the primary level. Additionally, this approach aligns with global health goals of task-shifting and strengthening primary care diagnostics using digital innovations.

In summary, our research demonstrates that even in resource-constrained settings like Quetta, Pakistan, cutting-edge AI technology can be leveraged to improve diagnostic accuracy and health equity. As we scale up this study, we anticipate that smartphone otoscopy with deep learning will become an invaluable tool for frontline healthcare workers, bridging the gap between patients and specialty care. This work contributes to the growing evidence that AI, when developed and deployed thoughtfully, can enhance clinical decision-making and outcomes in low-resource environments. The convergence of smartphones, affordable optics, and powerful algorithms heralds a new era of accessible ear care, with the ultimate aim of preventing avoidable hearing loss and its lifelong consequences.

REFERENCES

World Health Organization. Deafness and hearing loss [Internet]. Geneva: WHO; 2025 [cited 2025 Oct 10]. Available from: https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss

Guo Z, Ji W, Song P, Zhao J, Yan M, Zou X, et al. Global, regional, and national burden of hearing loss in children and adolescents, 1990-2021: a systematic analysis from the Global Burden of Disease Study 2021. BMC Public Health. 2024;24(1):2521.

Livingstone D, Chau J. Otoscopic diagnosis using computer vision: an automated machine learning approach. Laryngoscope. 2020;130(6):1408–1413.

Wu Z, Lin Z, Li L, Pan H, Chen G, Fu Y, et al. Deep learning for classification of pediatric otitis media. Laryngoscope. 2021;131(5):E2344–E2351.

Alhudhaif A, Cömert Z, Polat K. Otitis media detection using tympanic membrane images with a novel multi-class machine learning algorithm. PeerJ Comput Sci. 2021;7:e405.

Cai Y, Yu JG, Chen Y, Liu C, Xiao L, Grais EM, et al. Two-stage attention-aware convolutional neural network for automated diagnosis of otitis media from tympanic membrane images: development and validation study. Neural Networks. 2020;126:384–394.

Viscaino M, Wróbel K, Viscaino P, Muñoz M, Kaławaj K, Auat Cheein FA. Color dependence analysis in a CNN-based computer-aided diagnosis system for middle and external ear diseases. Diagnostics (Basel). 2022;12(4):917.

Byun H, Yu S, Oh J, Bae J, Yoon MS, Lee SH, et al. An assistive role of a machine learning network in diagnosis of middle ear diseases. J Clin Med. 2021;10(15):3198.

Crowson MG, Hartnick CJ, Diercks GR, Gallagher TQ, Fracchia MS, Setlur J, et al. Machine learning for accurate intraoperative pediatric middle ear effusion diagnosis. Pediatrics. 2021;147(4):e2020034546.

Pham VT, Tran TT, Wang PC, Chen PY, Lo MT. EAR-UNet: a deep learning-based approach for segmentation of tympanic membranes from otoscopic images. Artif Intell Med. 2021;115:102065.

Sundgaard JV, Harte J, Bray P, Laugesen S, Kamide Y, Tanaka C, et al. Deep metric learning for otitis media classification. Med Image Anal. 2021;71:102034.

Chen YC, Chu YC, Huang CY, Lee YT, Lee WY, Hsu CY, et al. Smartphone-based artificial intelligence using a transfer learning algorithm for the detection and diagnosis of middle ear diseases: a retrospective deep learning study. EClinicalMedicine. 2022;51:101543.

Habib AR, Kajbafzadeh M, Hasan Z, Wong E, Gunasekera H, Perry C, et al. Artificial intelligence to classify ear disease from otoscopy: a systematic review and meta-analysis. Clin Otolaryngol. 2022;47(3):401–413.

Frontier in Medical & Health Research

ISSN: (e) 3007-1607 (p) 3007-1593

Cha D, Pae C, Lee SA, Na G, Hur YK, Lee HY, et al. Differential biases and variabilities of deep learning-based artificial intelligence and human experts in clinical diagnosis: a retrospective cohort and survey study. JMIR Med Inform. 2021;9(12):e33049.

Dubois C, Eigen D, Simon F, Couloigner V, Gormish M, Chalumeau M, et al. Development and validation of a smartphone-based deep-learning-enabled system to detect middle-ear conditions in otoscopic images. NPJ Digit Med. 2024;7(1):162.

Akyol K. Comprehensive comparison of modified deep convolutional neural networks for automated detection of external and middle ear conditions. Neural Comput Appl. 2024;36:5529–5544.

Zeng X, Jiang Z, Luo W, Li H, Li H, Li G, et al. Efficient and accurate identification of ear diseases using an ensemble deep learning model. Sci Rep. 2021;11(1):10839.

Habib AR, Xu Y, Bock K, Mohanty S, Sederholm T, Weeks WB, et al. Evaluating the generalizability of deep learning image classification algorithms to detect middle ear disease using otoscopy. Sci Rep. 2023;13(1):5368.

Gao Z, Ding X, Huang Y, Tian X, Zhao Y, Feng G, et al. Diagnosis, treatment, and management of otitis media with artificial intelligence. Diagnostics (Basel). 2023;13(13):2309.

Binol H, Niazi MKK, Elmaraghy C, Moberly AC, Gurcan MN. OtoXNet: content-based eardrum image retrieval for otologic diagnosis. Front Digit Health. 2022;3:810427