Peer Review File

Reviewer A

Comment 1: Overall a well written paper with exciting potential.
Reply 1: Thank you

Comment 2: Minor typo in line 40 crossed out
Reply 2: Typo corrected

Comment 3: AI accuracy is reliant on large image datasets. Would a larger sample of abnormal images have increased the overall accuracy of AI recognition? If so please comment on this in the paper. Missing approximately 1 in 5 abnormal ears is problematic, please comment on how this can be reduced in future AI recognition of ear disease?

Reply 3: In response to reviewer comments we have added the following section to the Discussion section of the manuscript.

> The accuracy of the model is likely to improve with a larger training set that include images from beyond the subject service. The amount of data needed to train AI models is dependent on the complexity of the diagnosis. Tasks that are easy to solve for a human reader requires less training data than the detection of subtle or uncommon pathologies. [20]

Reviewer B

Comment 4: Overall – This paper represents a nice study assessing the early utility of AI in the diagnosis of Telehealth screening

Reply 4: Thank you

Comment 5: Could the authors elaborate on the types of diagnoses present within the training / validation dataset (6818 images) ? For example, were the distribution of abnormal diagnoses similarly to the test set ?

Reply 5: Diseases were approximately represented by the same proportion in the training, test, and validation groups. We have now added this detail to the manuscript.

Comment 6: Could the authors elaborate on the disease condition entitled " Abnormal without specific disease condition label " – this seems to represent about 20% of the conditions examined

Reply 6: We have added details about the source of these labels, which additionally explains why some abnormal images do not have a specific label attached. The manuscript now reads:

> Workflows of the screening service meant that images could be labelled with a specific disease condition or alternatively, categorised as abnormal (without specific disease condition).

Comment 7: Line 23 – remove duplicate "children"

Reply 7: Done

Comment 8: Line 78 – suggest "tympanometry" instead of tympanography

Reply 8: Corrected as per reviewer's suggestions