Towards Generalizable Image Classification Models for Detecting Middle Ear Diseases

Manfred Lindmark[1], Thorbjörn Lundberg[2], Mimmi Werner[3], Paolo Soda[1], Christer Grönlund[1], Fredrik Öhberg[1]

[1] Dept. of Diagnostics and Intervention, Biomedical Engineering, Umeå University, Umeå, Sweden, [2] Department of Public Health and Clinical Medicine, Unit of Family Medicine, Umeå University, Umeå, Sweden, [3] Department of Clinical Sciences, Otorhinolaryngology, Umeå University, Umeå, Sweden

Background

Diagnosing ear diseases from images of the tympanic membrane (TM) is a promising new application of deep learning-based classification models. While a high diagnostic accuracy has been reported by several studies, it has also been shown that these models generalize poorly to external data with different image characteristics. The aim of this study was to evaluate a model's generalizability by training artificial neural networks on images from the first three cohorts and use the remaining two as an external test set.

Materials & Methods

We used five cohorts of images from different countries, captured with different instruments at different times and with different image processing, resulting in a large heterogeneity between them. The network was trained on three open data sets of TM images and tested on two additional sources. In this study we tested several methods to determine the optimal strategy to improve generalizability for classification of normal/abnormal TM images. The investigated methods were image pre-processing, data augmentation, different network architectures (convolutional neural networks and Vision Transformer) and manually re-evaluating the open datasets by correcting misclassifications and removal of low-quality samples.

Results & conclusion

The accuracy of our best model was 82 percent (sensitivity 71%, specificity 89%) on the external cohorts. For the same network trained without augmentation and with non-processed datasets the accuracy was 74 percent (sensitivity 58%, specificity 84%). Our results have shown that network choice, augmentation strategy and data cleaning all individually have a significant impact on model generalizability.