Financial disclosures:
TA: None;
GA: None;
USE: Genentech (F), Kodiak (F), Novartis (F), Apellis (F,C), RetInSight (F,P);
HB: Heidelberg Engineering (F), Apellis (F,R),

1079 - C0173

Improving the robustness of deep learning systems for automated AMD screening in retinal OCT

Teresa Araújo[1], Guilherme Aresta[1], Ursula Schmidt-Erfurth[2], Hrvoje Bogunović[1]

[1]Christian Doppler Lab for Artificial Intelligence in Retina, Medical University of Vienna, Vienna, Austria
[2]Department of Ophthalmology and Optometry, Medical University of Vienna, Vienna, Austria

Contact:
teresa.safinisterraaraujo@meduniwien.ac.at
https://optima.meduniwien.ac.at

Purpose

Deep learning (DL) methods for automated retinal OCT screening often provide **overly confident predictions for unrelated pathologies (outliers),** compromising their translation to clinical use. This study aims at **improving the robustness/reliability** of the DL-based diagnostic systems when presented with disease types not included as part of the training.

Methods

Outlier exposure (OE), i.e. training a model with a small number of outlier cases, is explored to improve outlier detection during automated screening for **age-related macular degeneration (AMD)** on retinal OCTs.

We use a multi-center dataset with target classes: non-pathological, intermediate AMD (iAMD), neovascular AMD (nAMD) and geographical atrophy (GA), and outliers: diabetic macular edema (DME), retinal vein occlusion (RVO), and Stargardt disease. We fine-tuned a DL model (EfficientNetV2-B0) for central B-scan classification.

The tested approach is **entropy normalization OE,** i.e. approximating the outlier prediction probabilities to the uniform distribution. As a baseline, the network is trained without OE. Each sample's outlier score was obtained with the following metrics:

- maximum predicted classification probability (MP)
- entropy of the output probabilities
- Cosine distance based on the features of the penultimate network layer

The AMD dataset had **3364 OCTs (2661 patients)** and were split patient-wise into 70% training, 15% validation, and 15% testing; 295 outlier samples were included in the test set (162 DME, 19 Stargardt and 114 RVO), and 500 OCTs were available for OE.

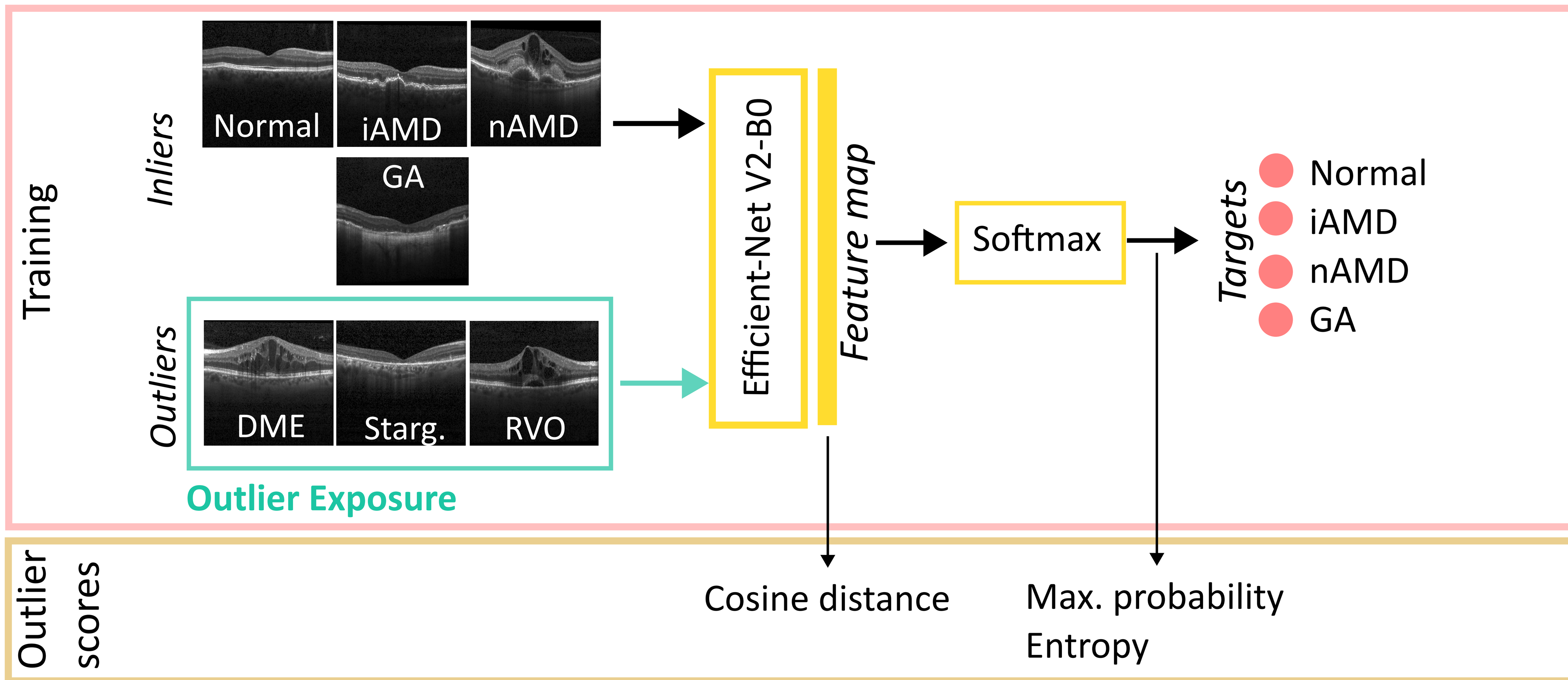

Fig. 1: Method for outlier detection.

Results

Providing a reduced number of outlier cases, increased the outlier detection performance without deteriorating the inlier classification performance: 0.98 macro-average area under the Receiver Operating Characteristic (AUC).

Table 1: AUC for the identification of outliers based on different metrics.

Method / Metric	Maximum probability	Entropy	Cosine distance
Baseline (no OE)	0.69	0.71	0.85
With OE (4 cases per class)	0.78	0.80	0.90

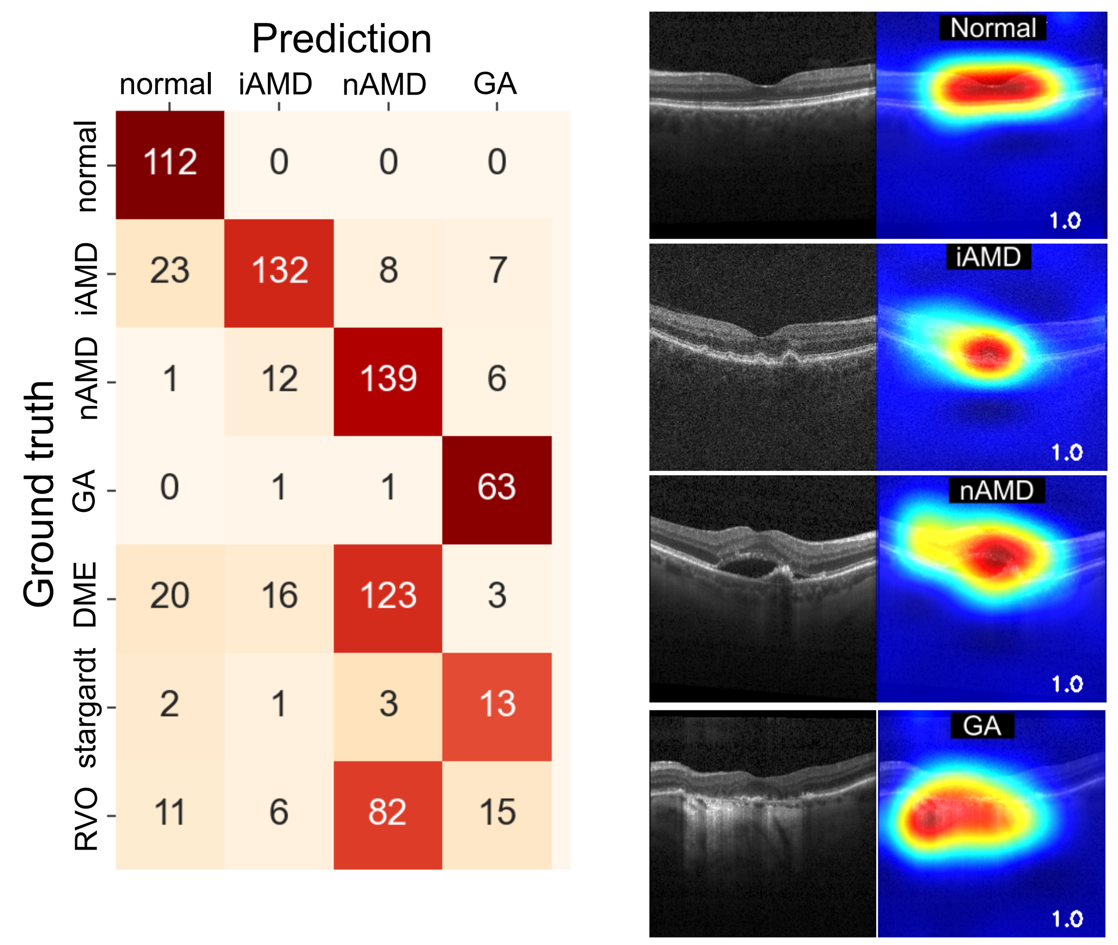

Fig. 2: Classification confusion matrix for AMD screening and the corresponding explanation maps.

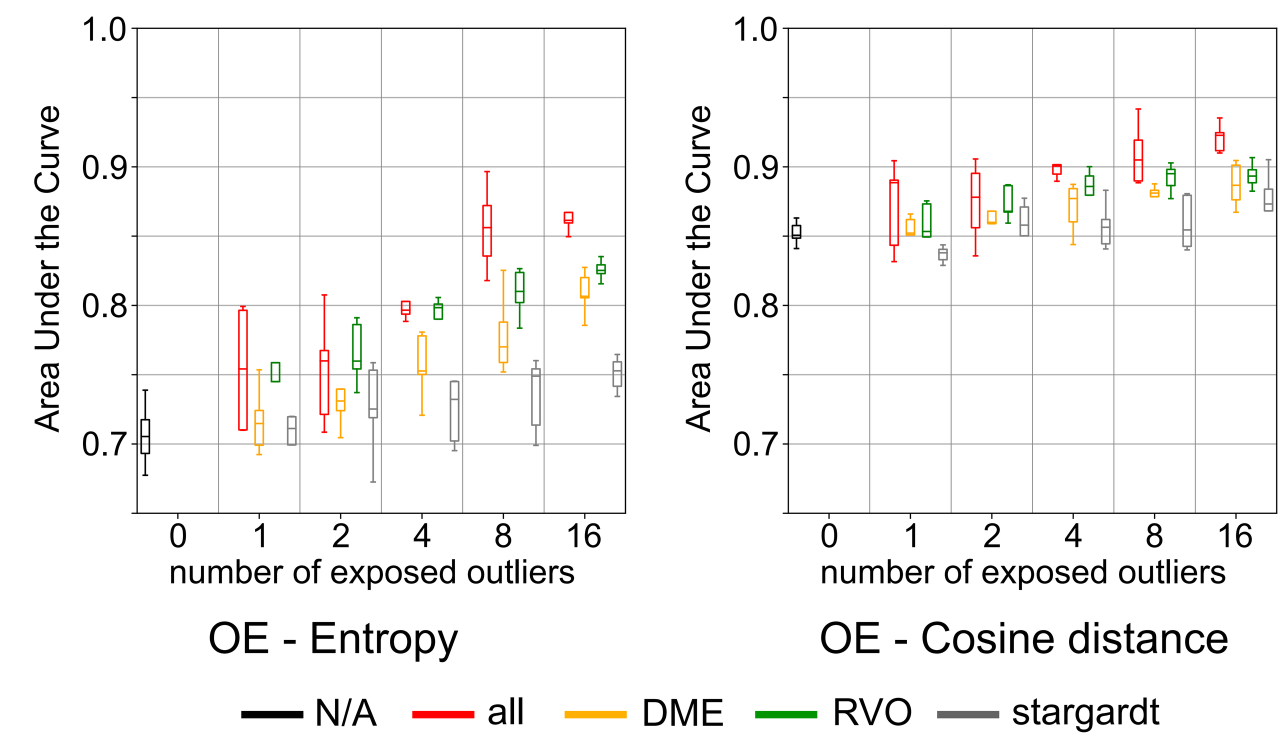

Fig. 3: OE method and score metrics performance for outlier detection, in relation with the number of exposed outliers. Colors: class of the exposed outliers; "all": all classes; "N/A": no exposure.

Conclusion

Combining **OE with the Cosine distance improved the outlier detection performance by 30%** compared to the baseline with MP scoring.

Exposing the network to a few non-AMD examples improves the detection of unrelated pathologies in the context of automated AMD screening, making the **DL systems more reliable and trustworthy.**

References

[1] Hendrycks, D. & Gimpel, K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. 5th Int. Conf. on Learn. Represent. ICLR 2017 1–12 (2017).

[2] Hendrycks, D., Mazeika, M. & Dietterich, T. Deep anomaly detection with outlier exposure. 7th Int. Conf. on Learn.Represent. ICLR 2019 1–18 (2019).

[3] Guha Roy, A. et al. Does your dermatology classifier know what it doesn't know? Detecting the long-tail of unseen conditions. Med. Image Analysis 75, 102274, DOI: 10.1016/j.media.2021.102274 (2022)