GOAL
Understand concepts learnt by a neural network classifying infectious disease (Tuberculosis (TB) and Monkeypox) image behavior

OVERVIEW
- We seek to explain behavior by using a post-hoc concept extraction (CE) technique
- The presented method (NMFx) is based on nonnegative matrix decomposition (NMF) from Collins [1] and is flexible to work in unsupervised, semi-supervised or weakly supervised fashion, and provided labels do not need to correspond to the labels that the underlying network was trained for

OPTIMIZATION DETAILS
Let $X \in \mathbb{R}^{n_1 \times n_2}$ denote the nonnegative data matrix of $n_2$ data points in $\mathbb{R}^{n_1}$. Lee and Seung [2] propose to decompose $X$ into a topic matrix $A$ and a weight matrix $S$ using the following, Frobenius-norm optimization objective:

$$\min_{A,S} \|X - AS\|_F^2$$

Here, $A \in \mathbb{R}^{n_1 \times k}$ denotes the topic matrix with $k$ topics and $S \in \mathbb{R}^{k \times n_2}_{\geq 0}$ denotes the representative weight matrix.

NMF with Image Label Supervision
When information about the data points’ labels is available, we can encode it into $Y \in \mathbb{R}^{l \times n_2}$, a binary label matrix where columns correspond to data points in $X$ and rows represent their class membership. The resulting objective becomes:

$$\min_{A,B,S} \|X - AS\|_F^2 + \lambda \|Y - BS\|_F^2$$

RESULTS
Visual explanations extracted using NMFx in TB classification (VGG-16) correspond to to areas used most in diagnosis (Topic 1).

In contrast to baselines (ECLAD[3] and the method of Oramas[4]), the NMFx method identifies larger and more consistently positioned regions in the input X-ray images.

CONCLUSION
- CE techniques are a promising visual explanation technique for understanding infectious disease classification using neural networks
- CE using NMFx is a lightweight and versatile method for analyzing NN decisions

BIBLIOGRAPHY
[1] Edo Collins, Radhakrishna Achanta, Sabine Susstrunk. Deep feature factorization. ECCV 2018.
[2] Daniel D Lee, H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. Nature 1999.
[3] Andres Felipe Posada-Moreno, Nikita Surya, Sebastian Trimpe. ECLAD: Extracting concepts with local aggregated descriptors, 2022.
[4] Jose Antonio Oramas Mogrovejo, Kaili Wang, Tinne Tuytelaars. Visual explanation by interpretation: Improving visual feedback capabilities of deep neural networks. In ICLR, 2019