UNIVERSITÄT ZU LÜBECK

dfki ai Artificial Intelligence in Medical Imaging

Privacy-Friendliness of Feature Extractors: Empirical Insights, Metrics and Correlations

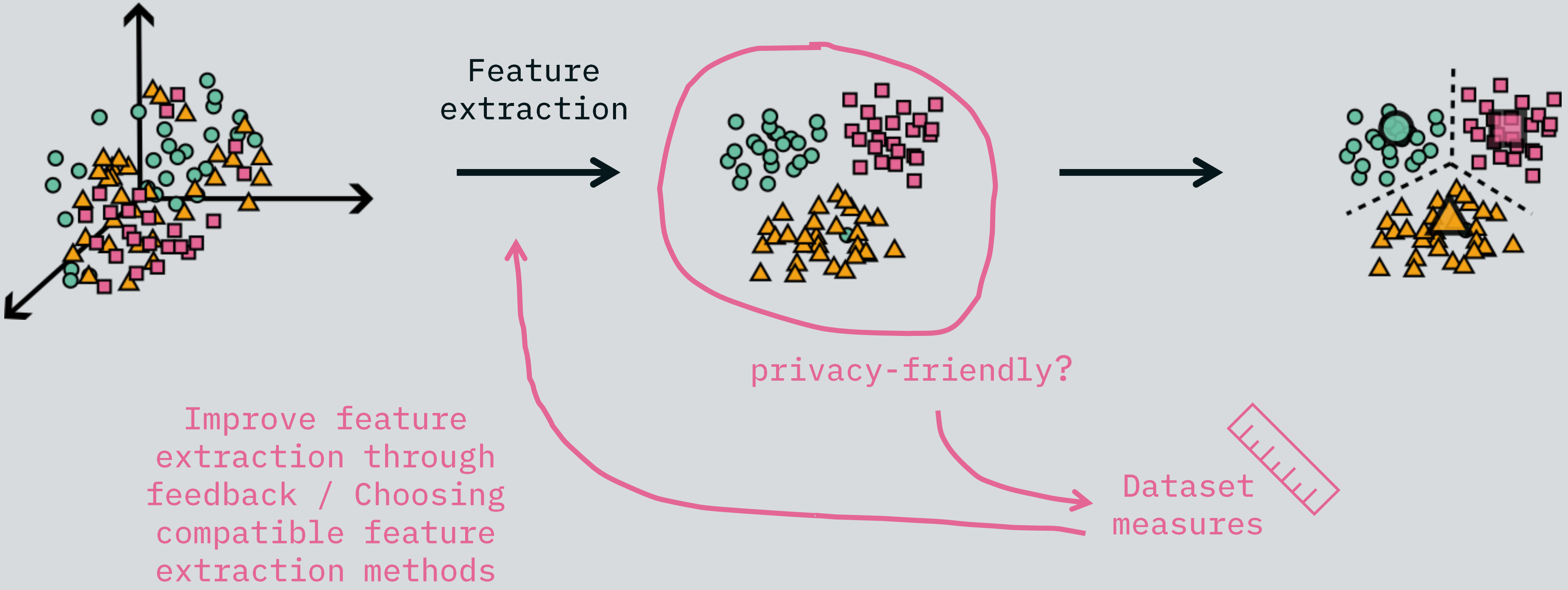
Nele Brügge

29.01.2026

Bundesministerium für Forschung, Technologie und Raumfahrt

Finanziert von der Europäischen Union
NextGenerationEU

Feature extraction for differentially-private machine learning



Feature extraction for differentially-private machine learning

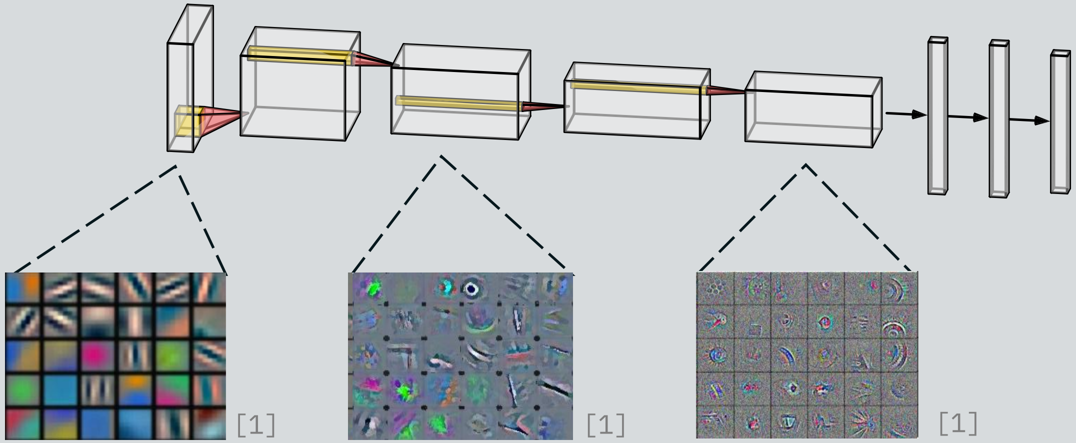
Recap Differential Privacy (DP):

$$Pr[M(D) \in S] \leq e^{\varepsilon} \cdot Pr[M(D') \in S] + \delta$$

⟶ an adversary cannot confidently infer information about a specific individual from the output of a randomized algorithm

- Protects against all known and unknown attacks
- Privacy loss ε can be quantified
- Multiple mechanisms can be composed

Using a public dataset

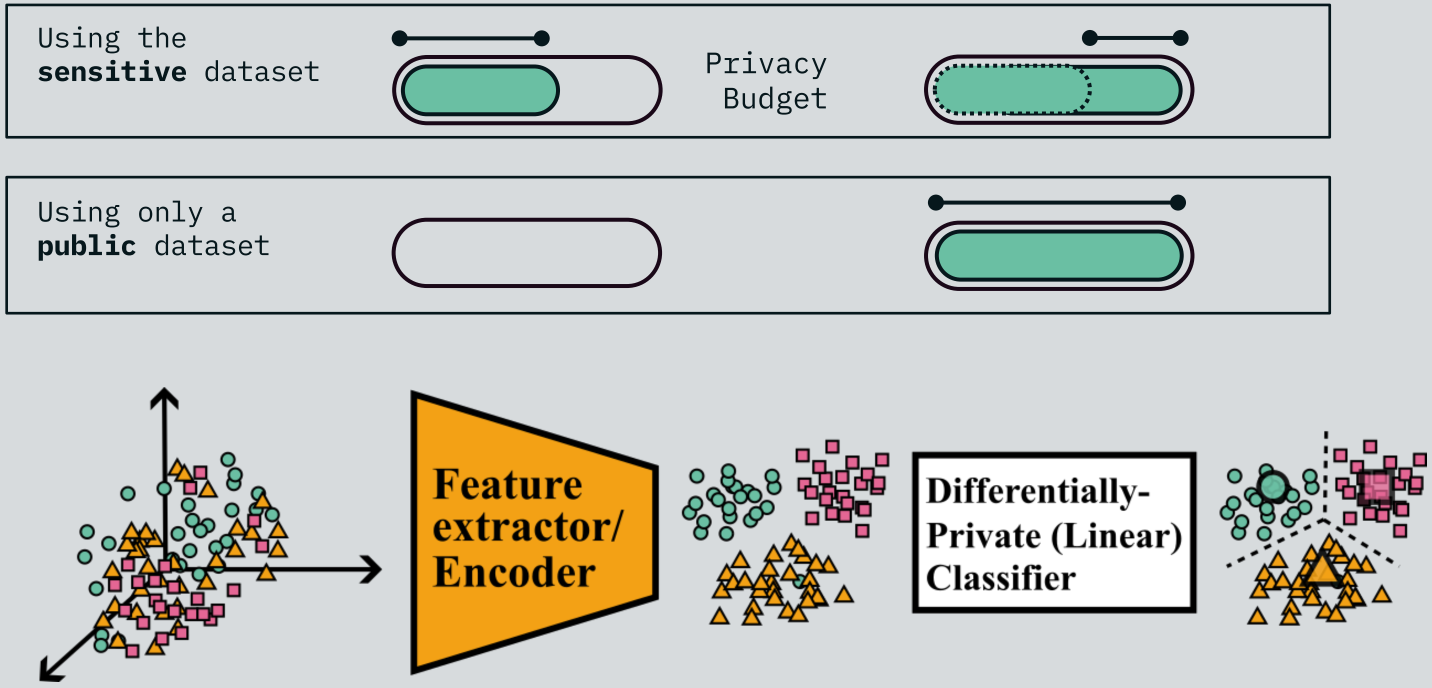


[1] [1] [1]

[1] Zeiler, M.D., Fergus, R. (2014). Visualizing and Understanding Convolutional Networks. In: European Conference on Computer Vision– ECCV 2014, pp 818–833.

Feature extraction for differentially-private machine learning

- Strong feature extraction improves downstream learning tasks
- Reduction of dimensionality, pretraining the network without using sensitive data
- Reduces downstream tasks to convex/linear learning problems, for which robust privacy-friendly algorithms exist



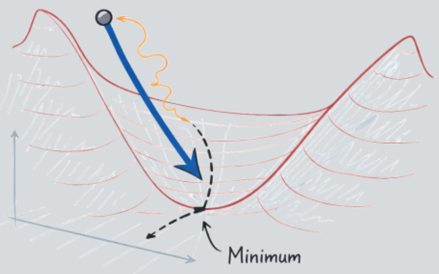
Research Questions

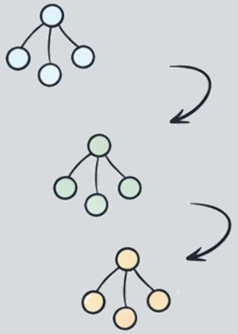
How should we design feature spaces for different classifiers e.g. DP-SGD, Gradient-Boosted Decision Trees?

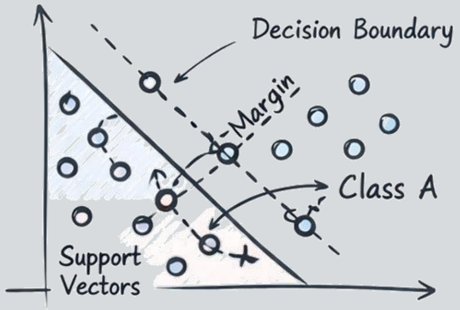
Which measures are relevant in high-dimensional spaces?

Which off-the-shelf feature extraction models lead to good utility-privacy trade-offs?

Can we use certain dataset property measures to choose suitable feature extractors for downstream DP classification?







Experiments

Artificial Intelligence in Medical Imaging

- Used toy datasets with diverse pre-trained feature extractors
- Evaluated separability-, entropy-, and clustering-based measures (supervised, unsupervised, and DBSCAN-based)
- Linked these measures to DP performance and the DP vs. non-DP gap as well as the normalized DP balanced accuracy

Tested example datasets

- CIFAR10 / CIFAR100
- Oxford flowers
- Oxford pets
- MedMNIST (blood, path, organ, breast)
- food101

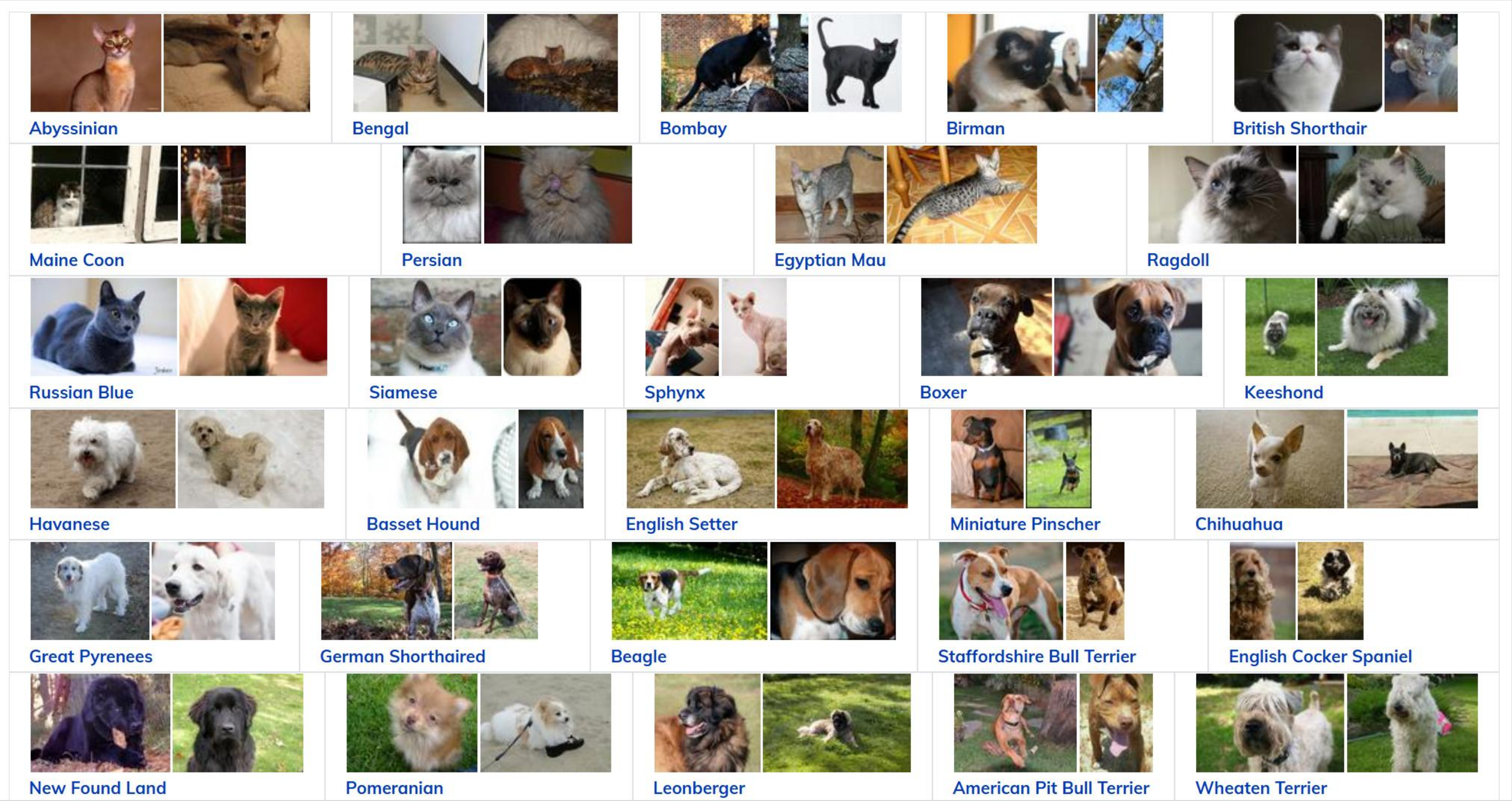
Pretrained models

- DINOv1, DINOv2, DINOv3 (ViT, ConvNext, tiny, small, base)
- SimCLR (ResNet-50)
- MoCov2 (ResNet-50)
- ... (many more)

Experiments

Tested example datasets

- CIFAR10 / CIFAR100
- Oxford flowers
- Oxford pets
- MedMNIST (blood, path, organ, breast)
- food101



O. M. Parkhi, A. Vedaldi, A. Zisserman, C. V. Jawahar
Cats and Dogs, IEEE Conference on Computer Vision and Pattern Recognition, 2012

Experiments

DINOv2 vit-small

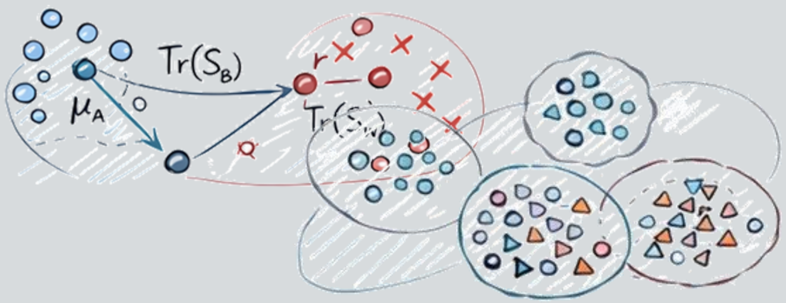
ResNet-50,
pretrained on ImageNet

Metrics

Current metrics mainly focus on separability, cluster structure and subpopulations

Examples:

- Silhouette Score
- **Calinski-Harabasz Index (normalized Trace Ratio)**
- Davies-Bouldin Index
- Class Granularity Index
- Geometrical Separability Index
- Fisher Discriminant Ratio
- Prototype Separability



$$CH = \frac{\mathrm{tr}(S_B)/(K-1)}{\mathrm{tr}(S_W)/(N-K)}$$

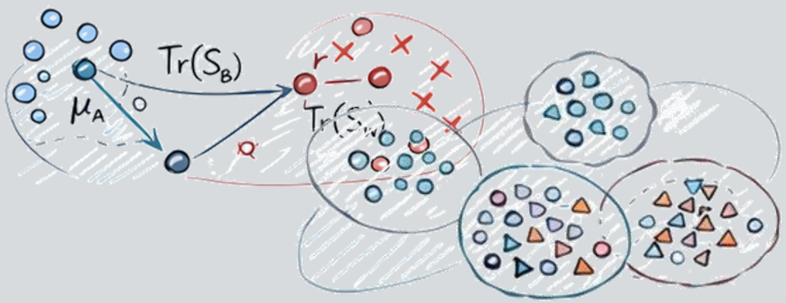
with $\mathrm{tr}(S_B) = \sum_{k=1}^{K} n_k \left\|C_k - C\right\|^2$

and $\mathrm{tr}(S_W) = \sum_{k=1}^{K} \sum_{i=1}^{n_k} n_k \left\|X_{i,k} - C_k\right\|^2$

n_k	Number of observations in cluster k
C_k	Centroid of cluster k
$X_{i,k}$	The i-th observation of cluster k
K	Number of clusters

Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics*, *3*(1), 1–27.

Metrics



Current metrics mainly focus on separability, cluster structure and subpopulations

Examples:

- Silhouette Score
- Calinski-Harabasz Index (normalized Trace Ratio)
- Davies-Bouldin Index
- Class Granularity Index
- **Geometrical Separability Index**
- Fisher Discriminant Ratio
- Prototype Separability

$$GSI = \frac{\sum_{i=1}^{n}\left(f(x_i) + f(x_i')\right)}{n}$$

n	Dataset size
f	Target function
x	Dataset
x_i'	Nearest neighbour of x_i

Thornton, C. Separability is a Learner's Best Friend. In: 4th Neural Computation and Psychology Workshop (1998)

Heading

Artificial Intelligence in Medical Imaging

Class Granularity Index + Subpopulation Isolation Score (DBSCAN-based):

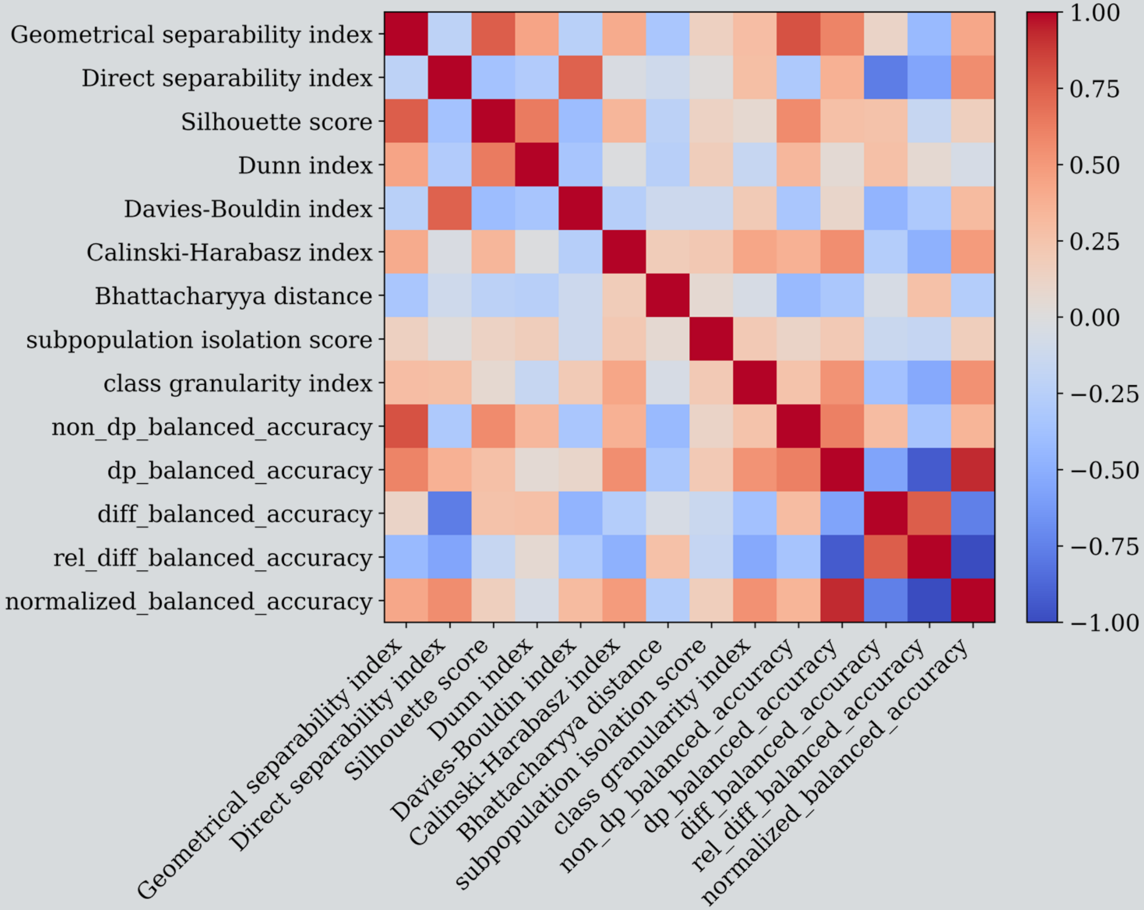
To what extent do intra-class sub-populations impact the performance and stability of differentially private classification models?

Experiments (DP-SGD)

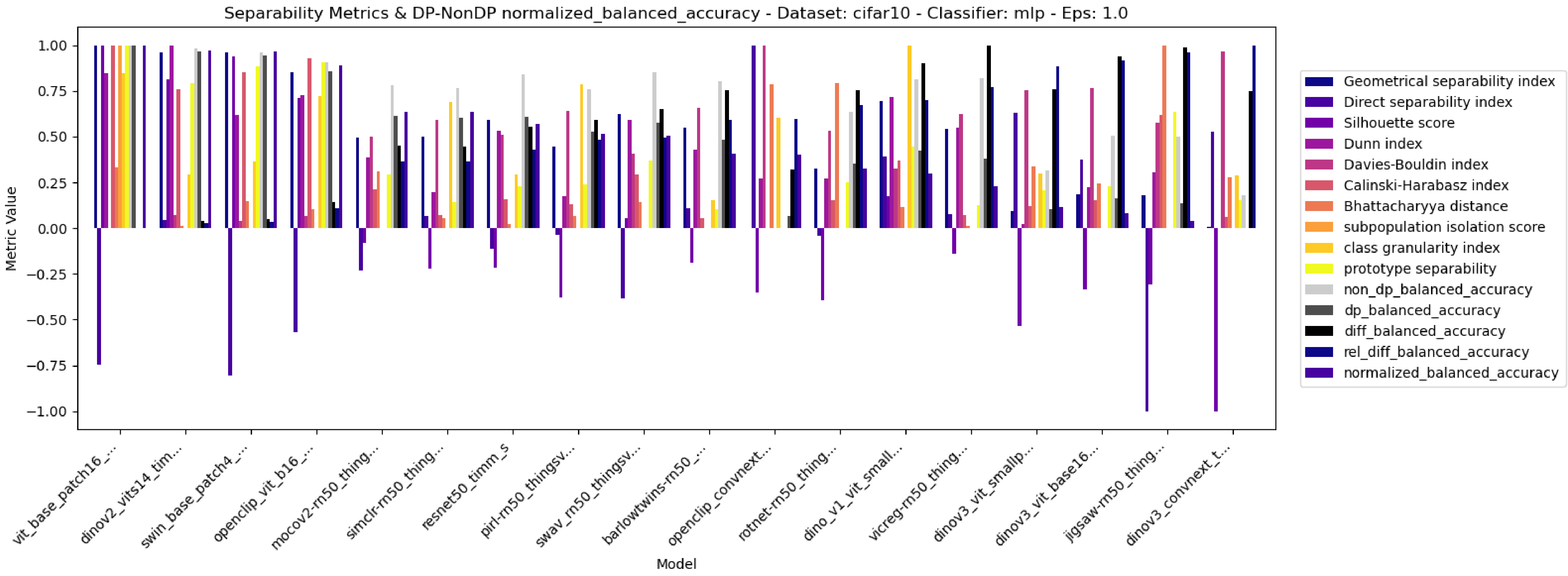
Correlation of measures and metrics with the DP classification accuracy / the gap between DP and non-DP classification accuracy, averaged over all datasets and all feature extraction models ($\varepsilon = 0.5$)

Highest correlations:

- Calinski-Harabasz index
- Direct separability index
- Class granularity index
- Geometrical separability index



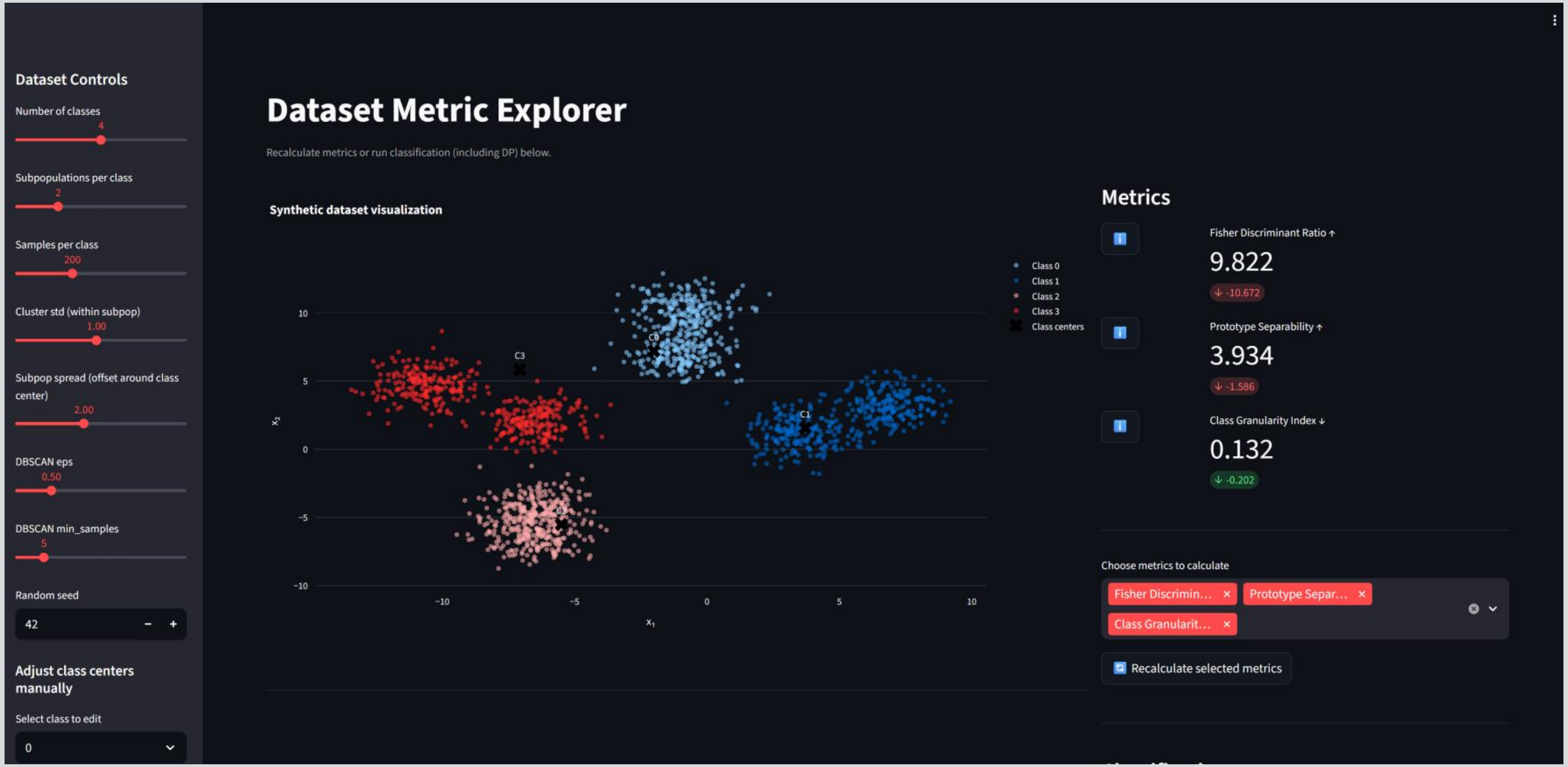
Experiments (DP-SGD)



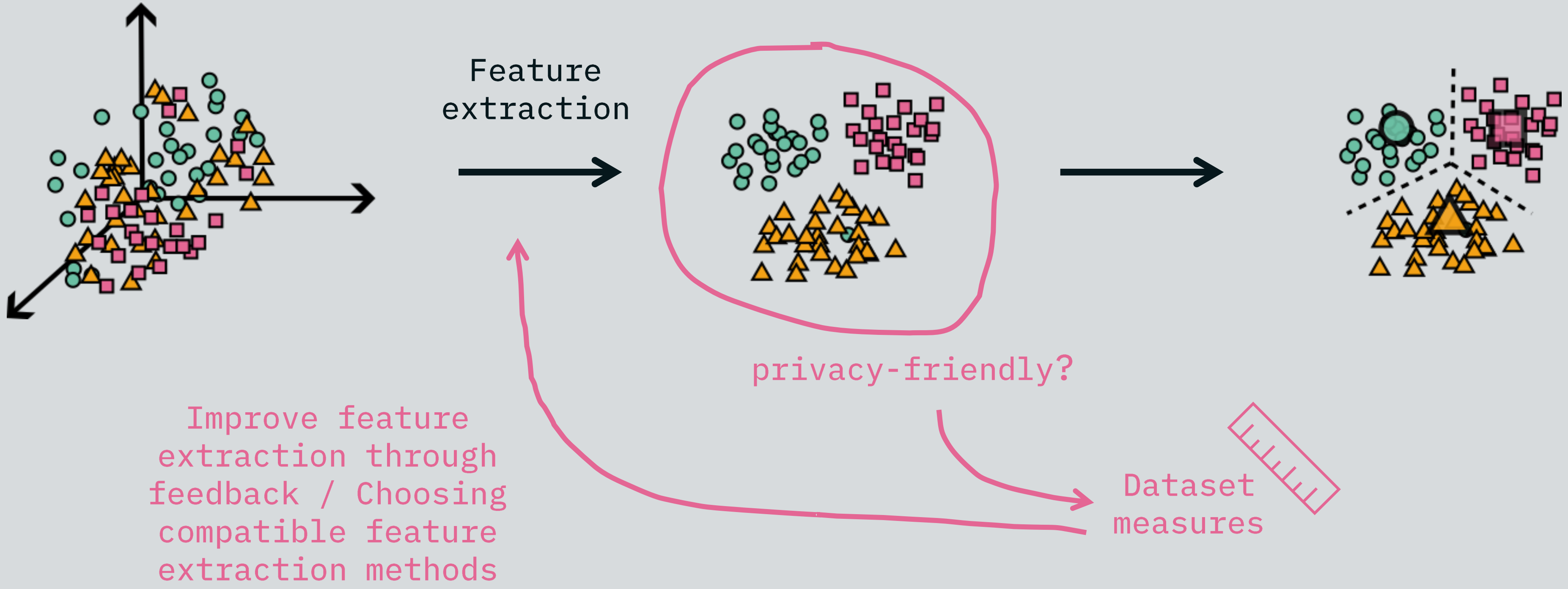
Demo

dfki ai Artificial Intelligence in Medical Imaging

- Change dataset characteristics
- Observe changes in measures
- DP/Non-DP classification



Feature extraction for differentially-private machine learning



Artificial Intelligence in Medical Imaging

Example: Facial Expression Classification

On CK+ and CelebA Datasets

©Jeffrey Cohn

Example Dataset
CK+ emotion recognition dataset

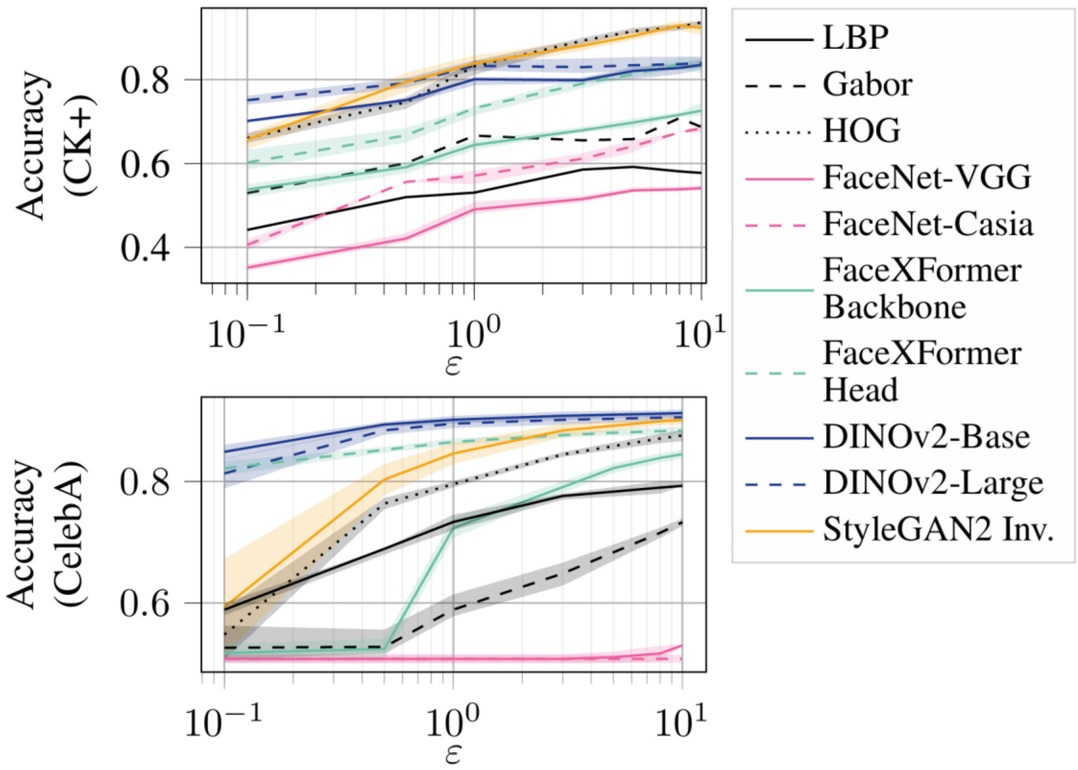
©Jeffrey Cohn

- **Multi-class classification** of 7 posed facial expressions (+neutral expression):
    - Happiness
    - Fear
    - Disgust
    - Anger
    - Sadness
    - Contempt
    - Surprise
- 123 subjects, 593 short video sequences
- Each sequence: onset (neutral) to peak formation of the facial expression

P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar and I. Matthews, "The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression," 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, San Francisco, CA, USA, 2010

Privacy Friendliness:
Facial Expression Classification

- Evaluation of feature extractors under different DP budgets
- Depending on the privacy budget used, different feature extractors can be recommended
- Calinski-Harabasz index and Geometrical Separability Index correlate with the DP performance

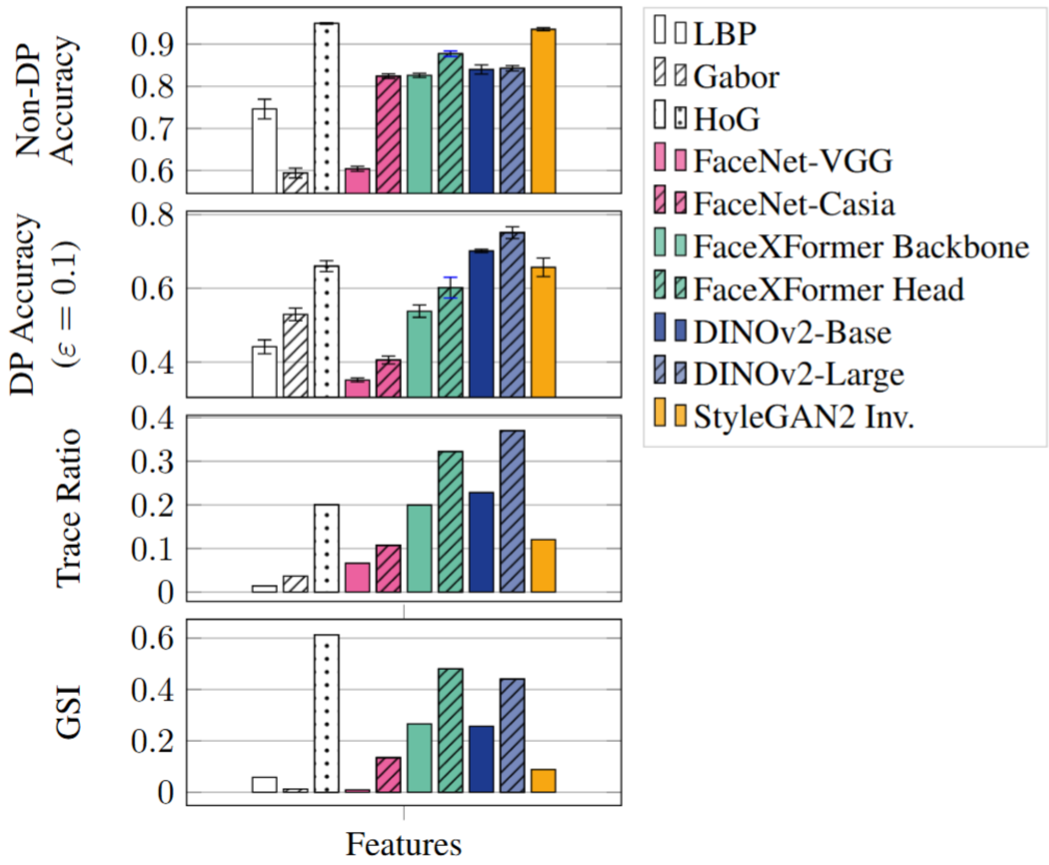


Privacy Friendliness: Facial Expression Classification

Artificial Intelligence in Medical Imaging

Evaluation of feature extractors under different DP budgets:

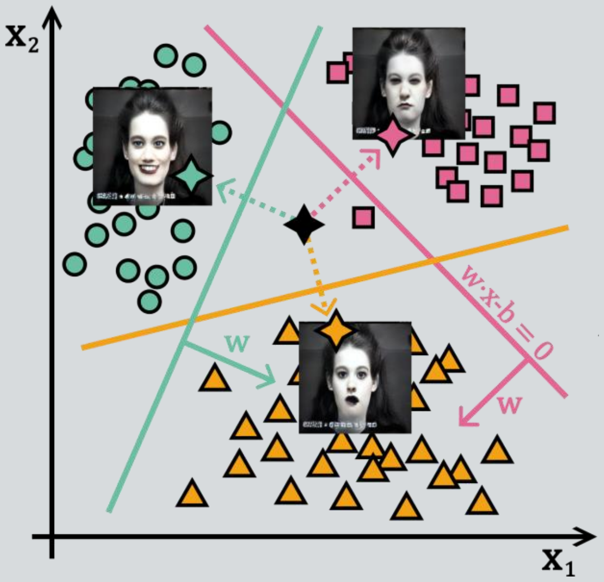
- Depending on the privacy budget used, different feature extractors can be recommended
- Calinski-Harabasz index (Trace Ratio) and Geometrical Separability Index correlate with the DP performance



Artificial Intelligence in Medical Imaging

Results Explainability

CK+ Dataset; StyleGAN features



GEFÖRDERT VOM

Bundesministerium für Bildung und Forschung

Finanziert von der Europäischen Union
NextGenerationEU

UNIVERSITÄT ZU LÜBECK

dfki ai Artificial Intelligence in Medical Imaging

Questions / Contact

Nele Sophie Brügge
nele.bruegge@dfki.de

Researcher @ DFKI Lübeck
German Research Center for Artificial Intelligence (DFKI)
Building 64, 2nd floor, room 12
Ratzeburger Allee 160
23562 Lübeck, Germany