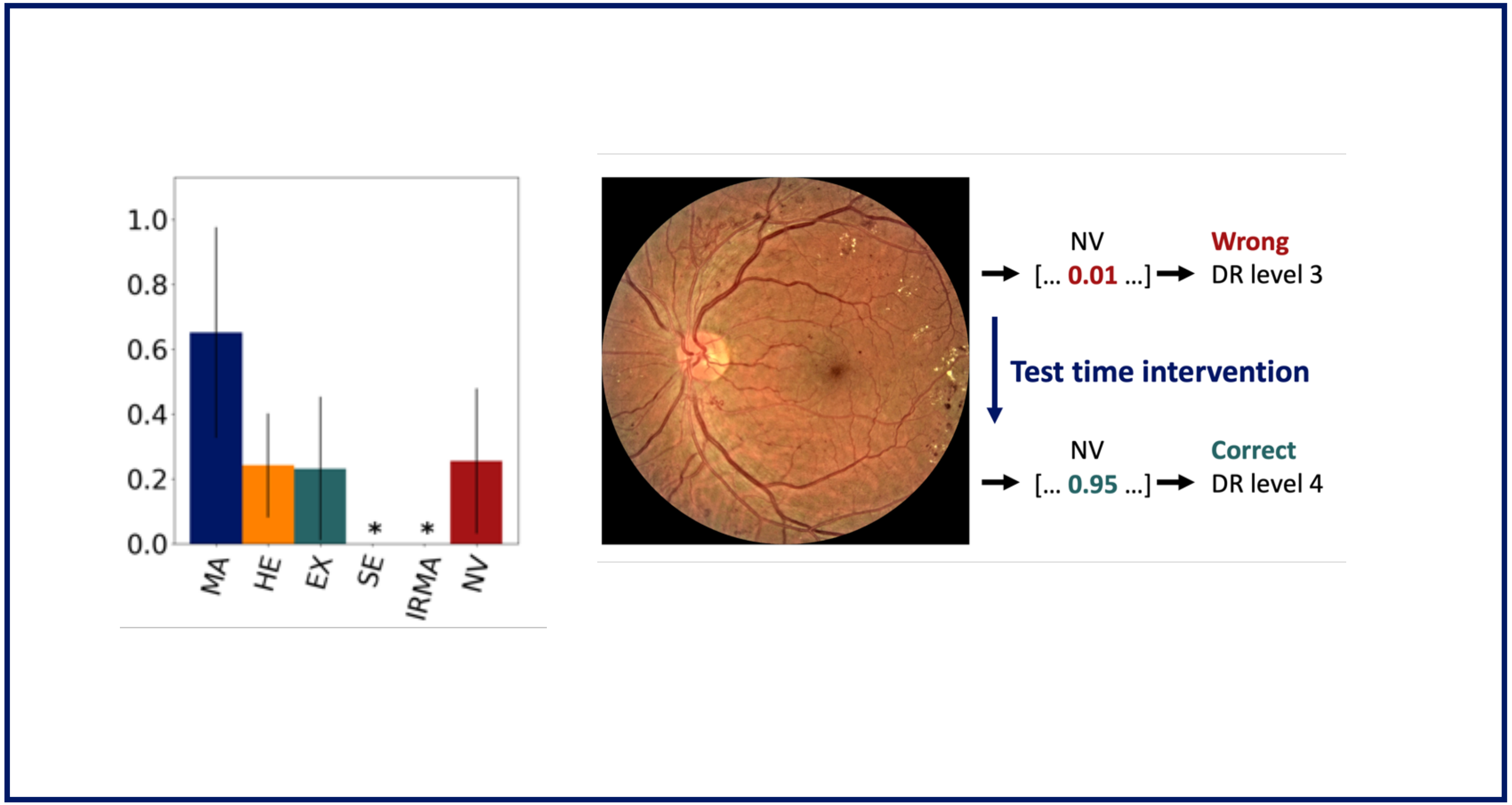
Looking into Concept Explanation Methods for Diabetic Retinopathy Classification

Andrea M. Storås and
Josefine V. Sundgaard

iMIMIC October 8, 2023



NV
→ [... 0.01 ...] → DR level 3
Wrong

Test time intervention

NV
→ [... 0.95 ...] → DR level 4
Correct

simulamet
Simula Metropolitan Center for Digital Engineering AS

novo nordisk®

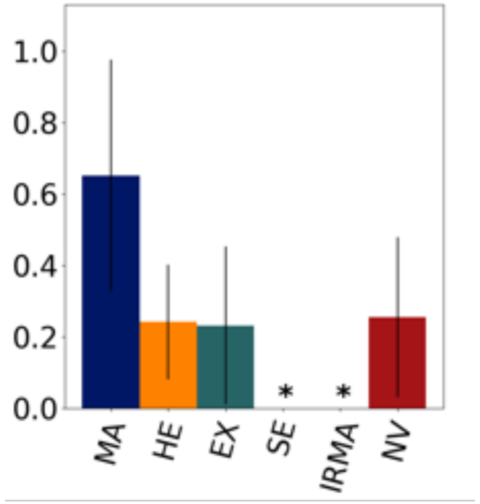
Technical University of Denmark
DTU

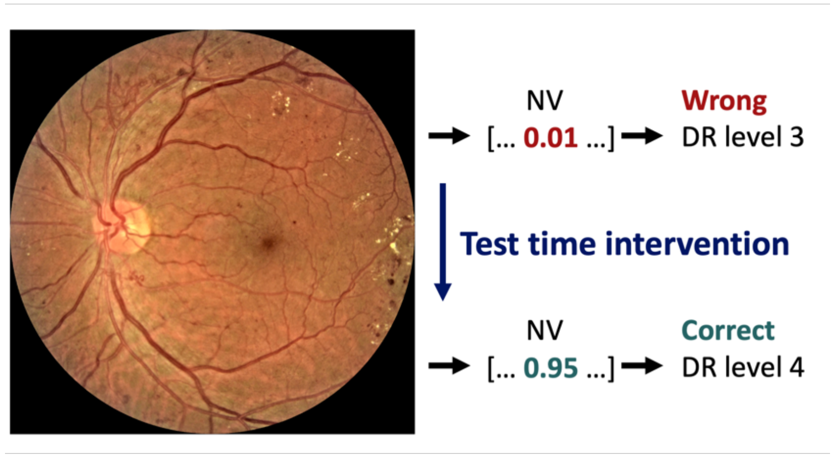
Fundus images are from Zhou, Y. et al. (2021): A Benchmark for Studying Diabetic Retinopathy: Segmentation, Grading and Transferability. https://doi.org/10.1109/TMI.2020.3037771

Looking into Concept Explanation Methods for Diabetic Retinopathy Classification

Andrea M. Storås and
Josefine V. Sundgaard

iMIMIC October 8, 2023





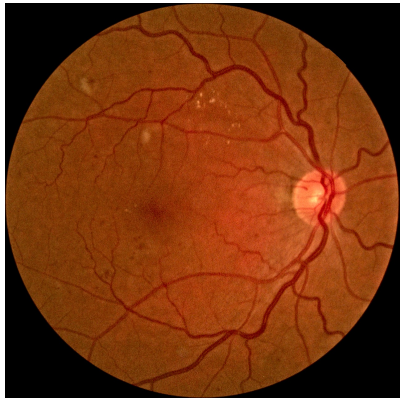
NV → [... 0.01 ...] → Wrong DR level 3

Test time intervention

NV → [... 0.95 ...] → Correct DR level 4

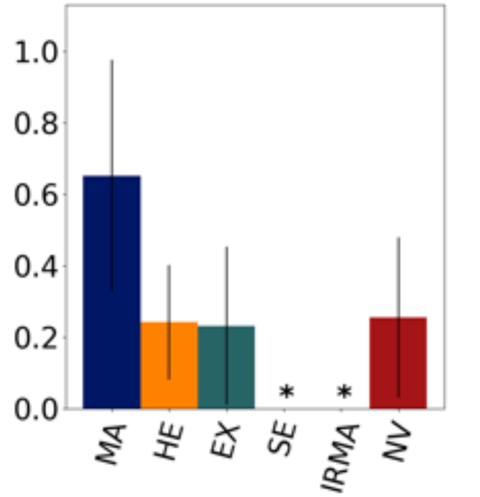
Fundus images are from Zhou, Y. et al. (2021): A Benchmark for Studying Diabetic Retinopathy: Segmentation, Grading and Transferability. https://doi.org/10.1109/TMI.2020.3037771

This talk compares two concept explanation methods for deep learning-based diabetic retinopathy (DR) grading



DR and DL

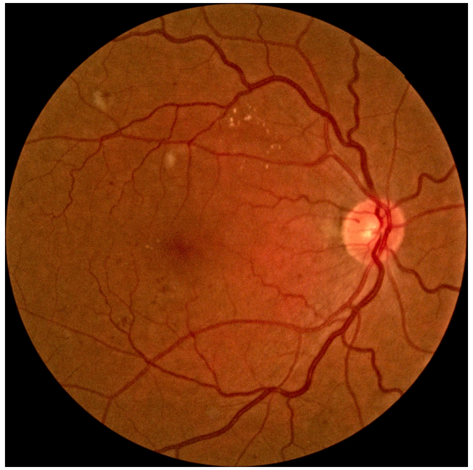
Concept explanations



Results and discussion

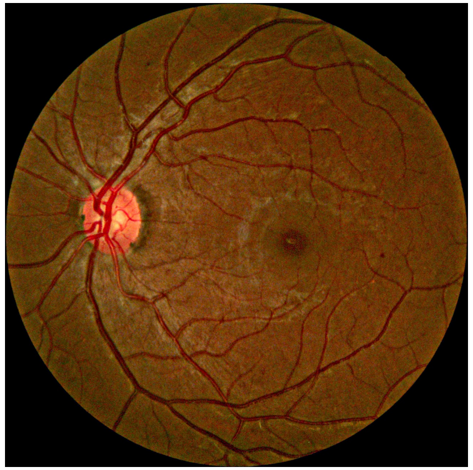
DR is graded from 0 to 4 based on findings in fundus images

Level 0



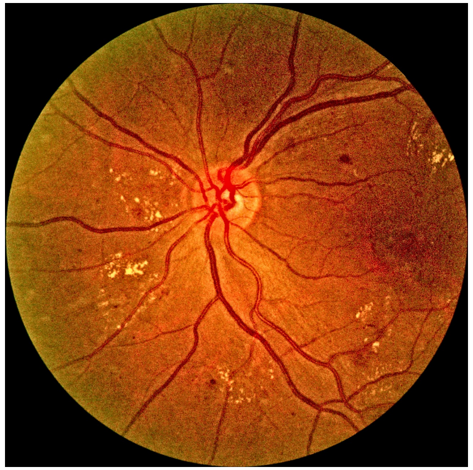
No abnormalities

Level 1



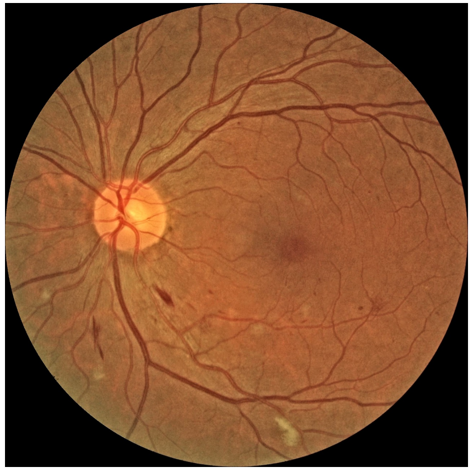
Microaneurysms (MA) only

Level 2



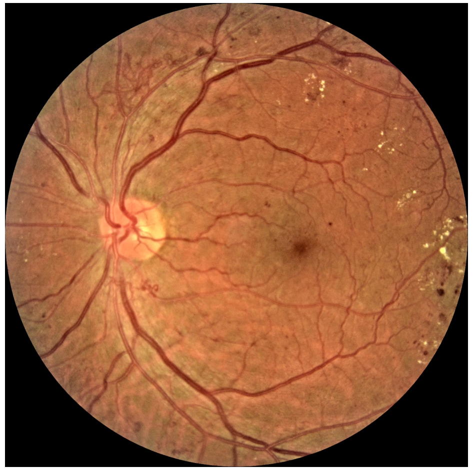
More than MA, but less severe than level 3

Level 3



No signs of proliferative DR and either >20 intraretinal hemorrhages in each quadrant, definite venous beadings in 2+ quadrants or prominent intraretinal microvascular abnormalities

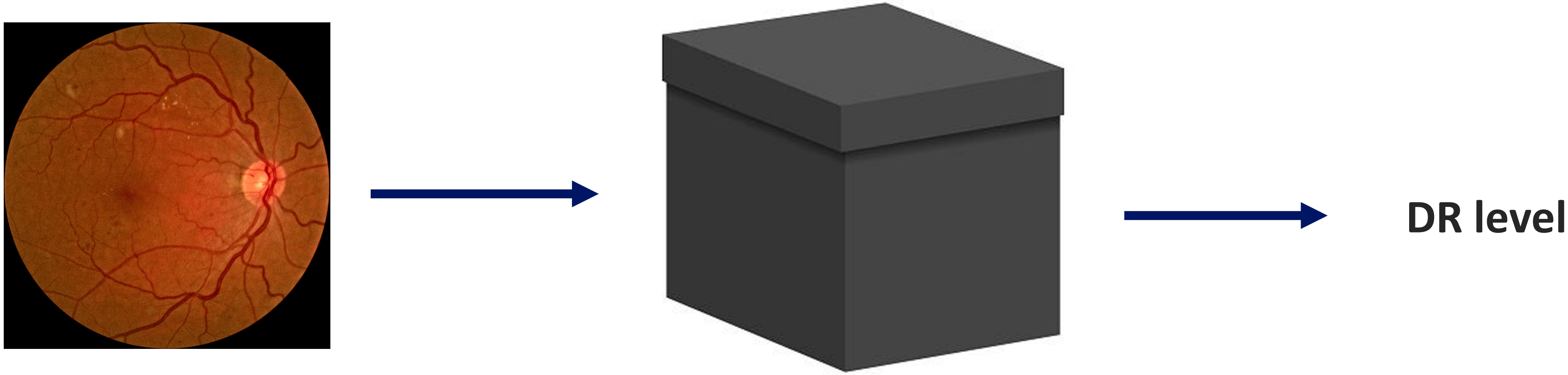
Level 4



Neovascularization and/or vitreous/preretinal hemorrhage

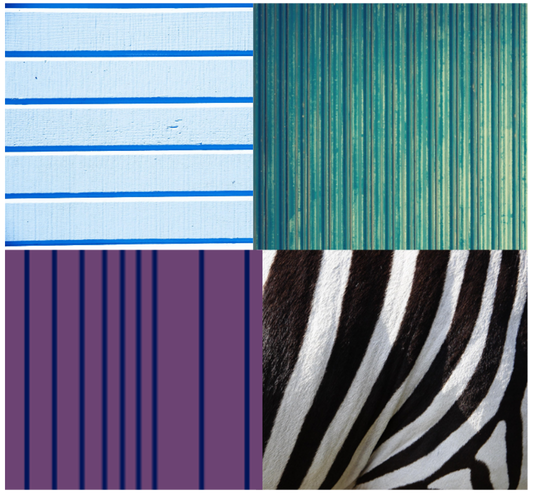
Wilkinson, C. et al. (2003). Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales.
Doi: https://doi.org/10.1016/S0161-6420(03)00475-5

Deep learning can grade fundus images, but less work has been done on explaining the models

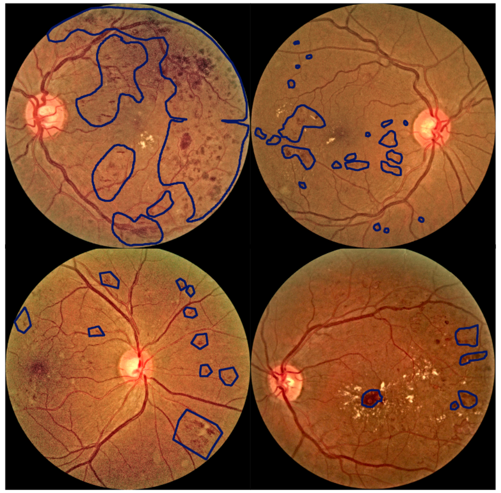


DR level

Concept-based explanations have several advantages that heatmaps lack



Stripes



Hemorrhages

Concept-based explanations have several advantages that heatmaps lack

- User-defined concepts
- Adapt to use case
- Quantify the concept importance for the model
- Explain a group of images

Concept-based explanations have several advantages that heatmaps lack

- User-defined concepts
- Adapt to use case
- Quantify the concept importance for the model
- Explain a group of images

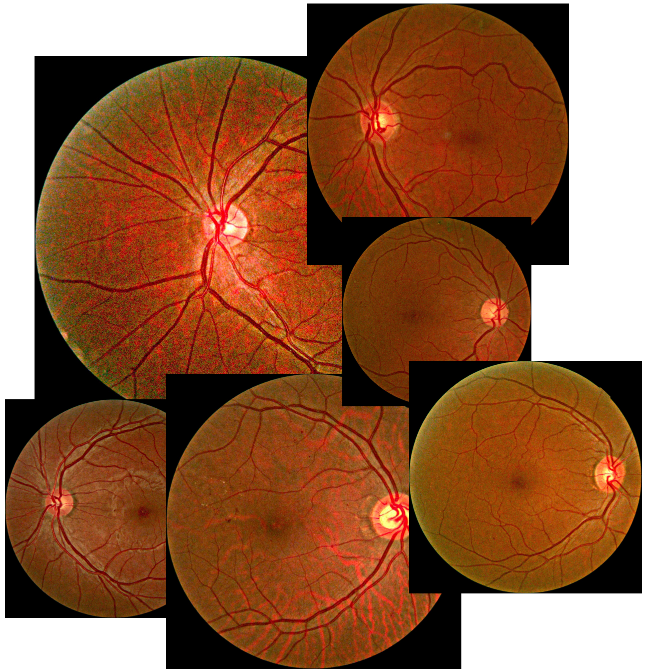
→ **We compare two concept-based methods for explaining deep neural networks grading DR**

Six concepts representing relevant medical findings for DR grading were defined

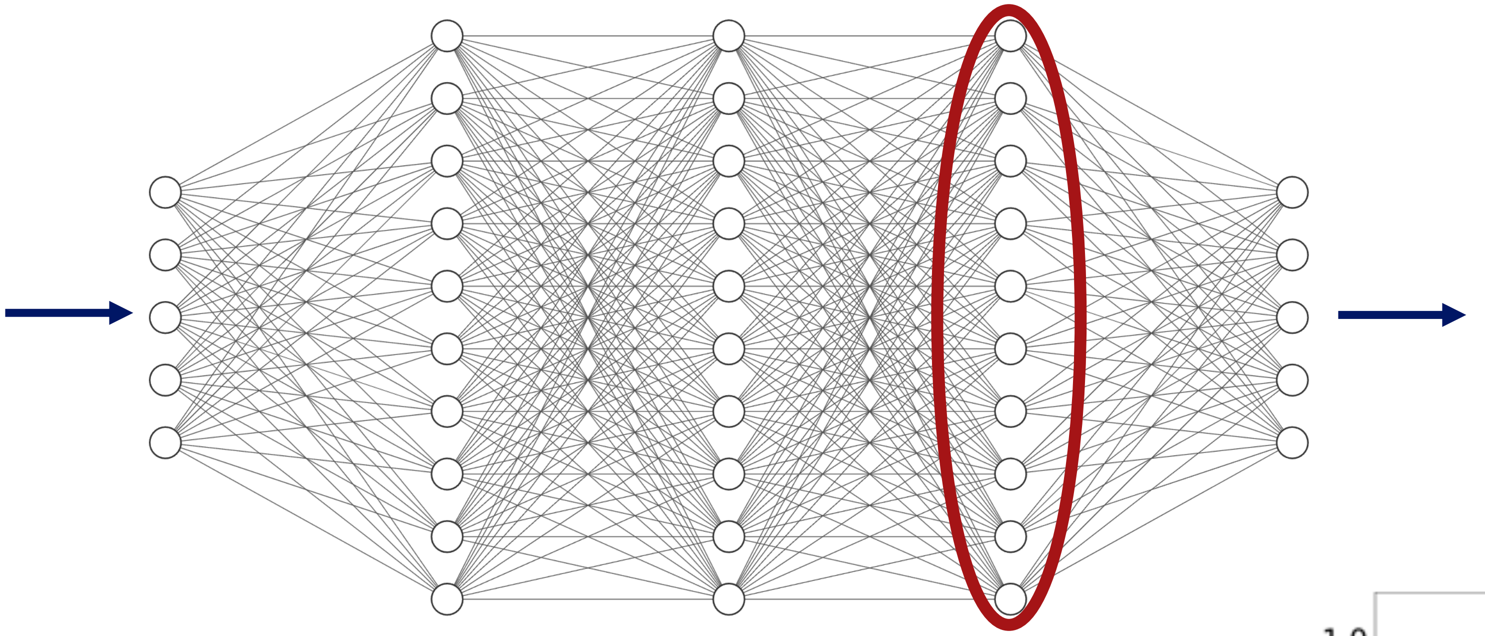
- Microaneurysms (MA)
- Hemorrhages (HE)
- Hard exudates (EX)
- Soft exudates (SE)
- Intraretinal microvascular abnormalities (IRMA)
- Neovascularization (NV)

1. Testing with Concept Activation Vectors (TCAV)

Input images

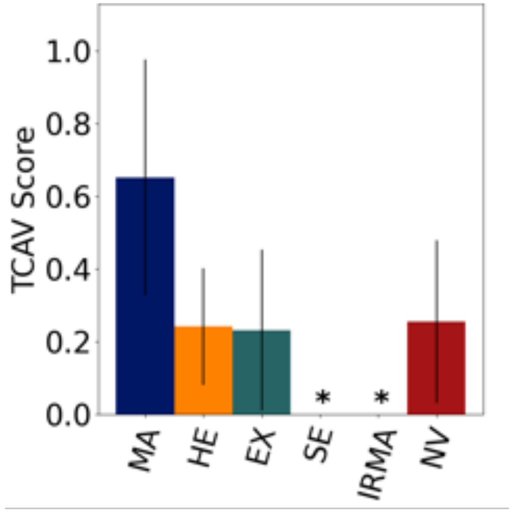


Deep neural network



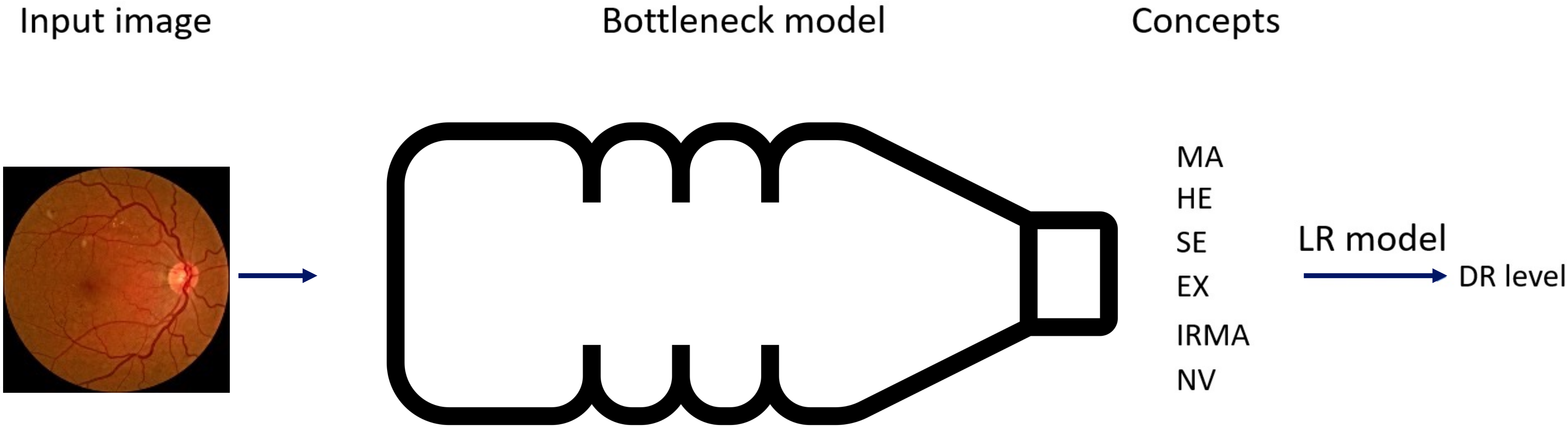
DR level

TCAV



Kim, B. et al. (2018). Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). URL: https://proceedings.mlr.press/v80/kim18d.html.

2. Concept Bottleneck Models (CBMs)



Input image

Bottleneck model

Concepts

MA
HE
SE
EX
IRMA
NV

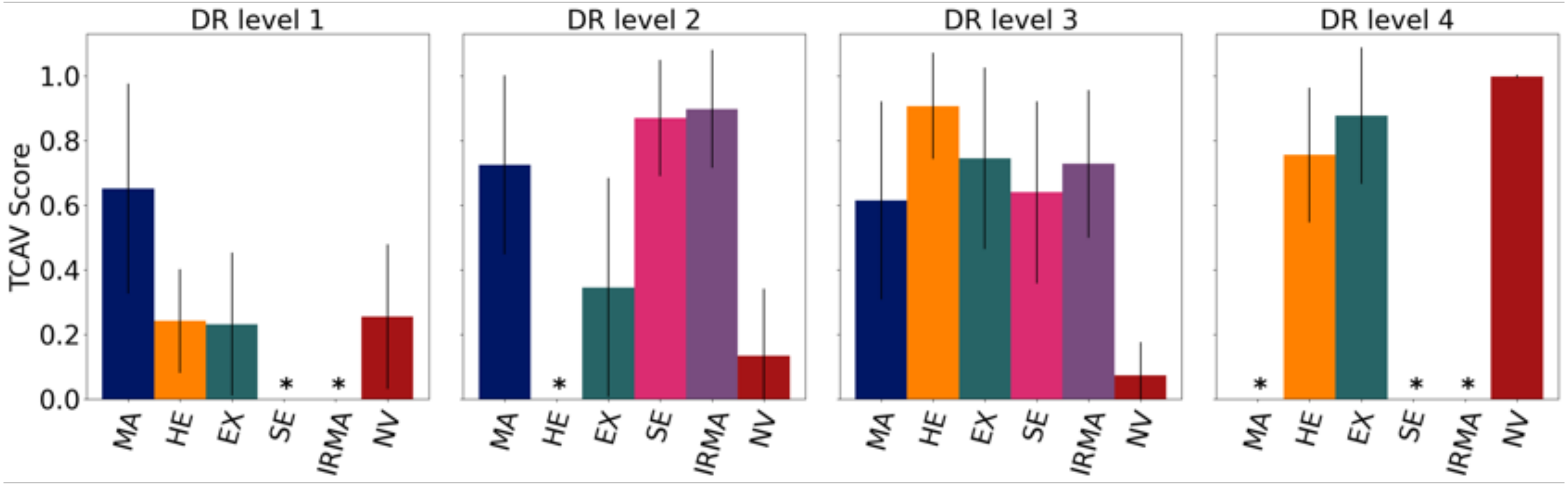
LR model

DR level

Koh, P.W. et al. (2020). Concept Bottleneck Models.

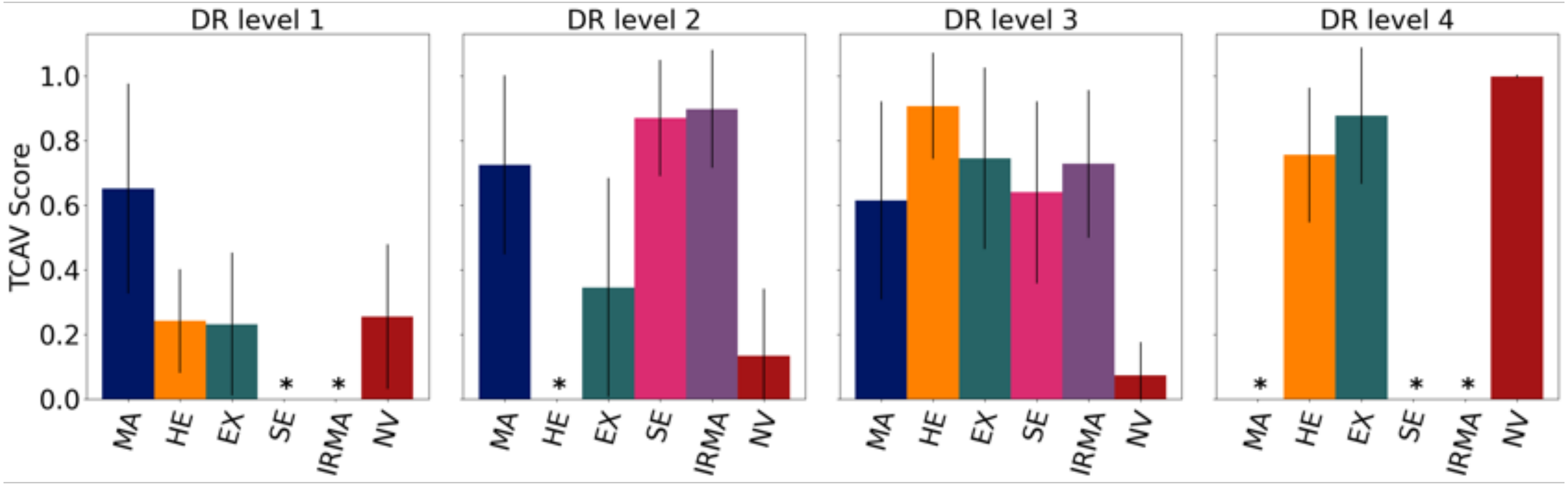
URL: https://proceedings.mlr.press/v119/koh20a.html.

Results TCAV: Concept ranking is highly in line with diagnostic criteria of DR



*: Insignificant concept

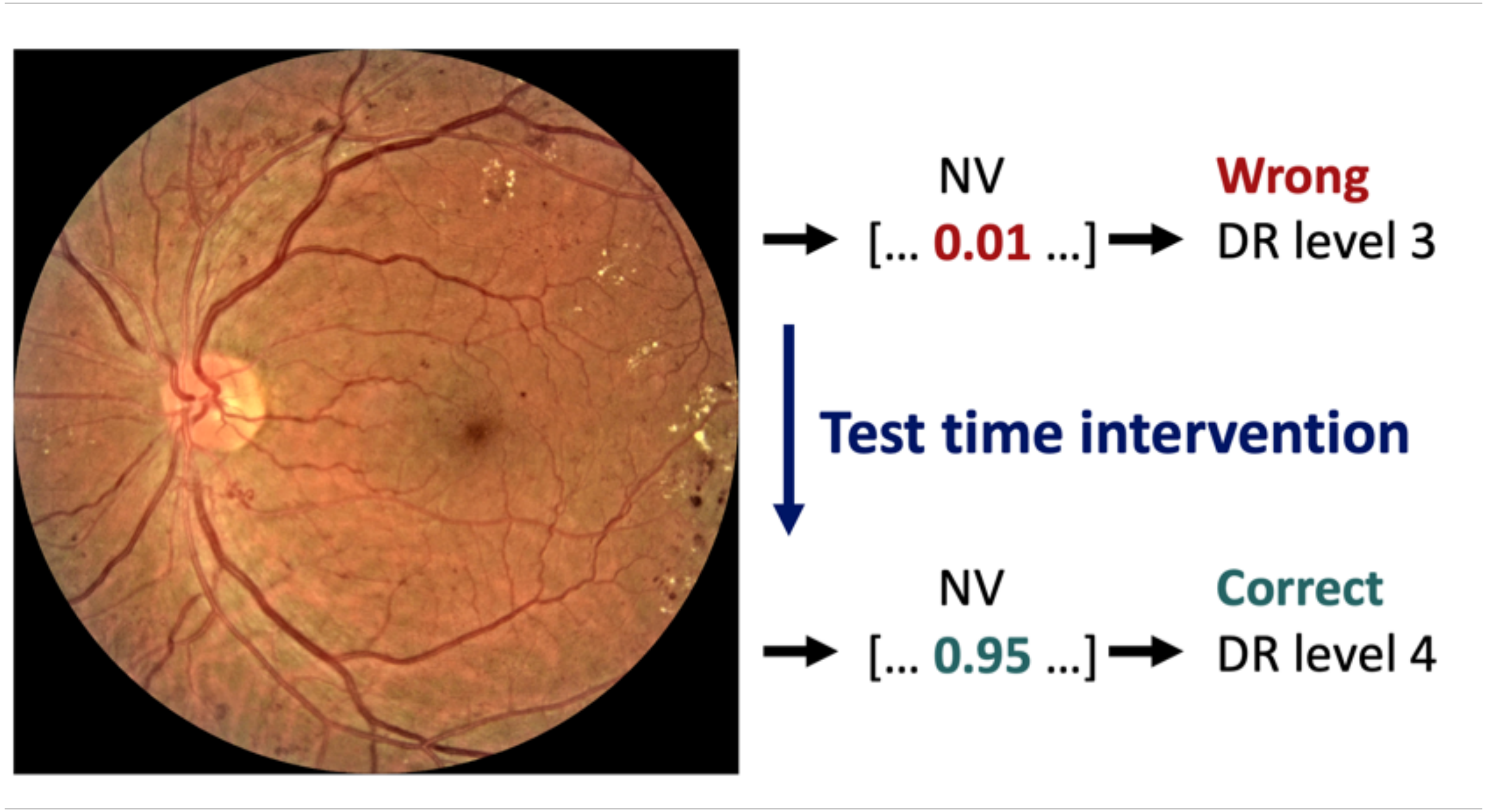
Results TCAV: Concept ranking is highly in line with diagnostic criteria of DR



Increasing severity

*: Insignificant concept

Results CBMs: Test time intervention on predicted concepts improve model accuracy



NV

→ [... 0.01 ...] → DR level 3 Wrong

Test time intervention

NV

→ [... 0.95 ...] → DR level 4 Correct

Results for DR grading: The CBMs do not generalize well to fundus images from external test datasets, probably due to limited training data

Model	No. of concepts	Accuracy	Balanced accuracy	F1 score	MCC	Precision
TCAV	-	**81.2%**	**62.3%**	**0.612**	**0.615**	**0.613**
CBM	4	71.9%	44.8%	0.429	0.416	0.454
CBM	6	24.8%	39.9%	0.257	0.095	0.318

Results for DR grading: The CBMs do not generalize well to fundus images from external test datasets, probably due to limited training data

Model	No. of concepts	Accuracy	Balanced accuracy	F1 score	MCC	Precision
TCAV	-	**81.2%**	**62.3%**	**0.612**	**0.615**	**0.613**
CBM	4	71.9%	44.8%	0.429	0.416	0.454
CBM	6	24.8%	39.9%	0.257	0.095	0.318

To conclude, concept explanations are promising for deep learning-based DR grading

CBMs allow for intervention at test time, but require datasets annotated with both concepts and target labels

↓

TCAV provides the best trade-off between model performance and explainability for DR grading

Questions?