Sparse Activations for Interpretable Disease Grading

Presented by DJOUMESSI Kerol

Kerol Djoumessi[1], *Indu Ilanchezian*[1], *Laura Kühlewein*[1], *Hanna Faber*[1], *Christian Baumgartner*[1], *Bubacarr Bah*[2], *Philipp Berens*[1], *Lisa M. Koch*[1]

[1] *University of Tübingen, Germany*
[2] *African Institute for Mathematical Sciences, South Africa*

July 12, 2023

Motivation

What is interpretability?

Interpretability is the degree to which a human can understand the cause of a decision (Miller 2019).

Motivation

What is interpretability?

Interpretability is the degree to which a human can understand the cause of a decision (Miller 2019).

Why do we need interpretable models?

- Trust AI
- Bias detection

Motivation

What is interpretability?

Interpretability is the degree to which a human can understand the cause of a decision (Miller 2019).

Why do we need interpretable models?

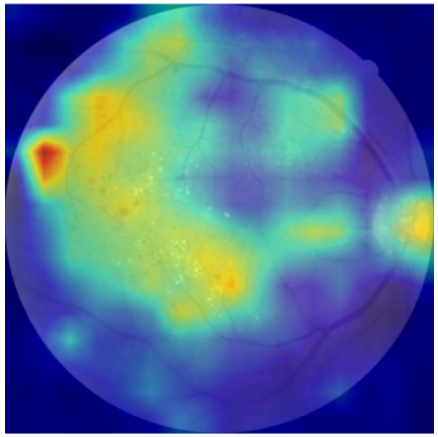
- Trust AI
- Bias detection
- Learning tools
- Policy/legal considerations

Interpretability methods

Why did a model make a specific prediction?

- Attribution maps: where a model looks

Post-hoc explanation

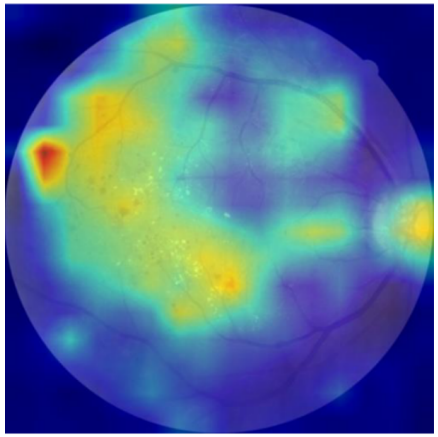


Interpretability methods

Why did a model make a specific prediction?

- Attribution maps: where a model looks

Post-hoc explanation



Interpretable-by-design models

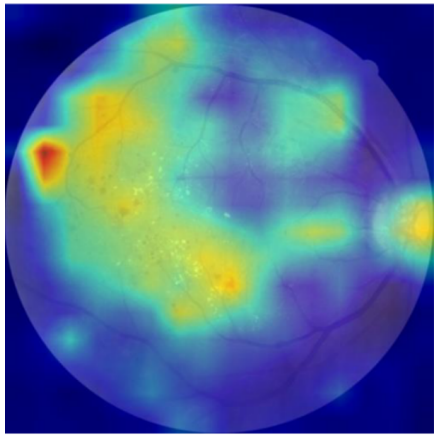
- **Linear models**
- Prototypes-based models

Interpretability methods

Why did a model make a specific prediction?

- Attribution maps: where a model looks

Post-hoc explanation



Interpretable-by-design models

- **Linear models**
- Prototypes-based models

The explanation may look different depending on reason for the explainability.

Attribution maps for medical images: drawbacks

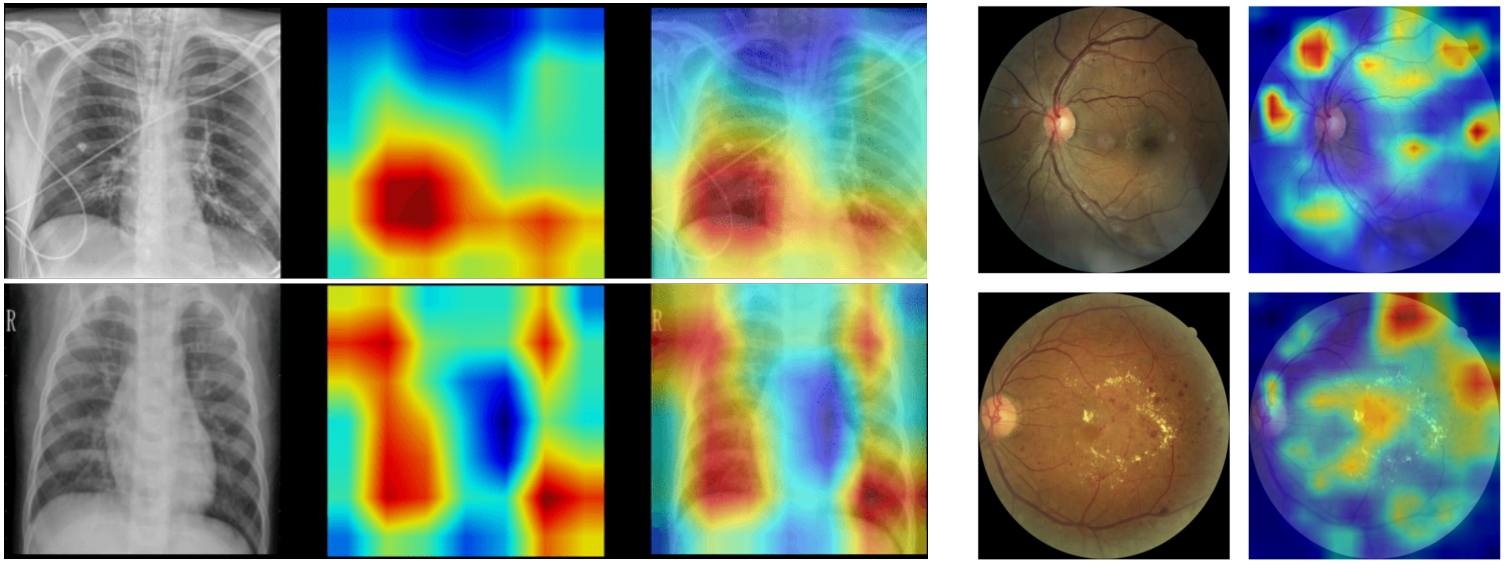


Figure reproduced from Bhowal et al. (2021)

Attribution maps for medical images: drawbacks

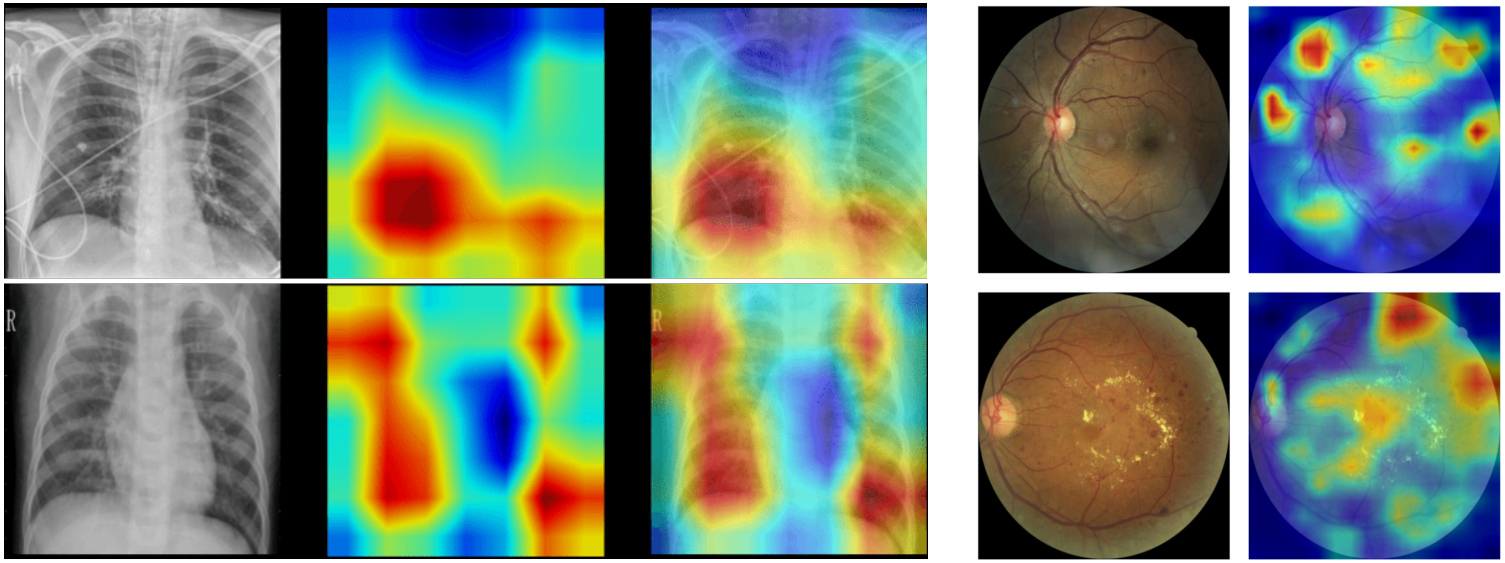


Figure reproduced from Bhowal et al. (2021)

- Highly variable
- Approximate CNNs
- No actionable insights
- Often coarse-grained evidence

Attribution maps for medical images: drawbacks

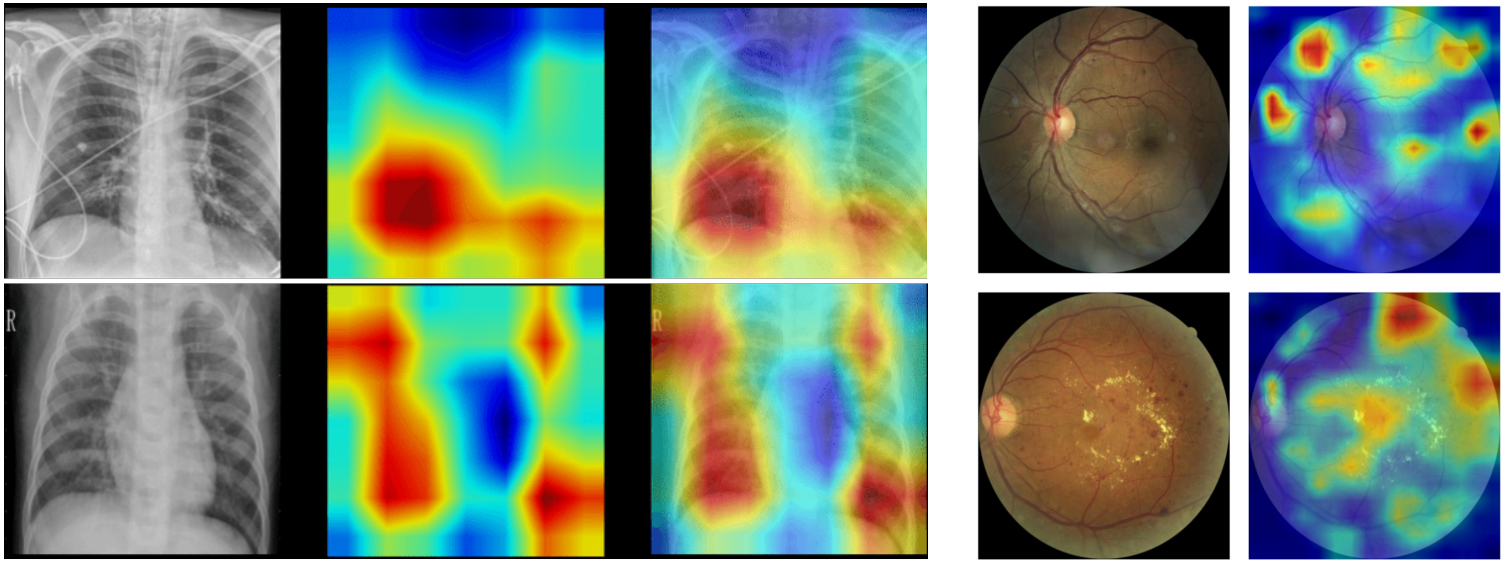


Figure reproduced from Bhowal et al. (2021)

- Highly variable
- Approximate CNNs
- No actionable insights
- Often coarse-grained evidence

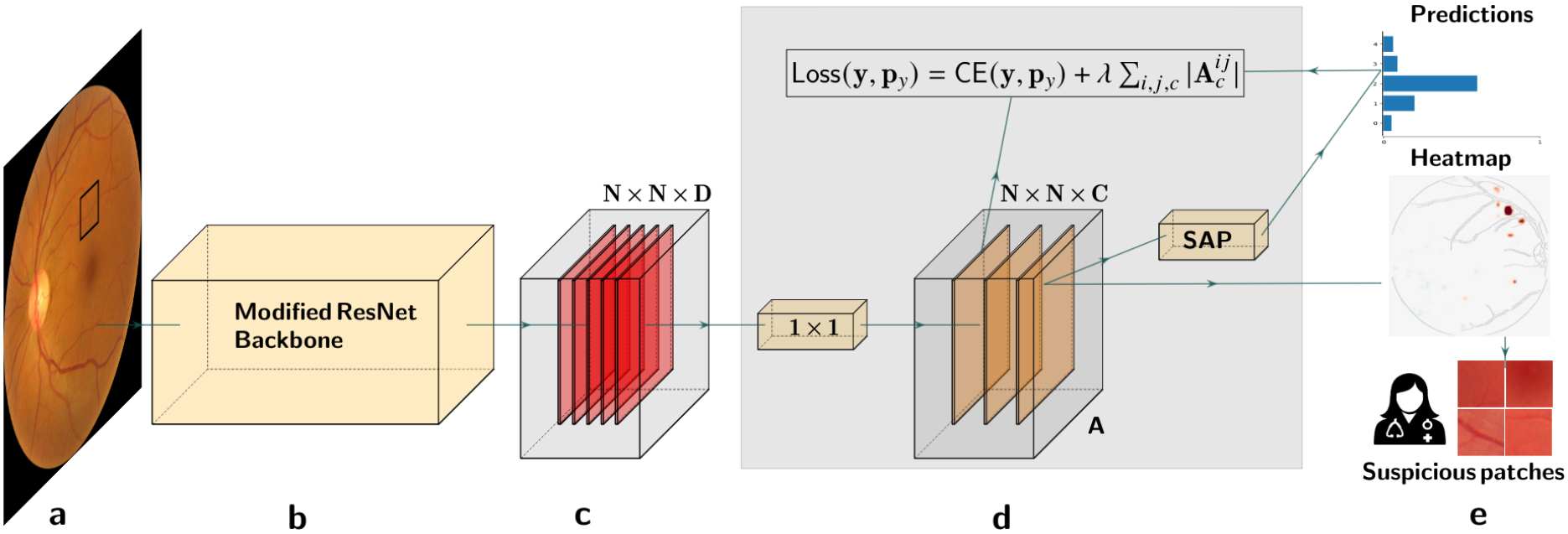
- Not inherently interpretable
- Poor performance in Med. data
  - Reproducibility
  - Lesion localizations

Contribution: sparse BagNet

We propose an inherently interpretable model that combines the feature extraction capabilities of DNNs with the advantages of sparse linear models in interpretability.

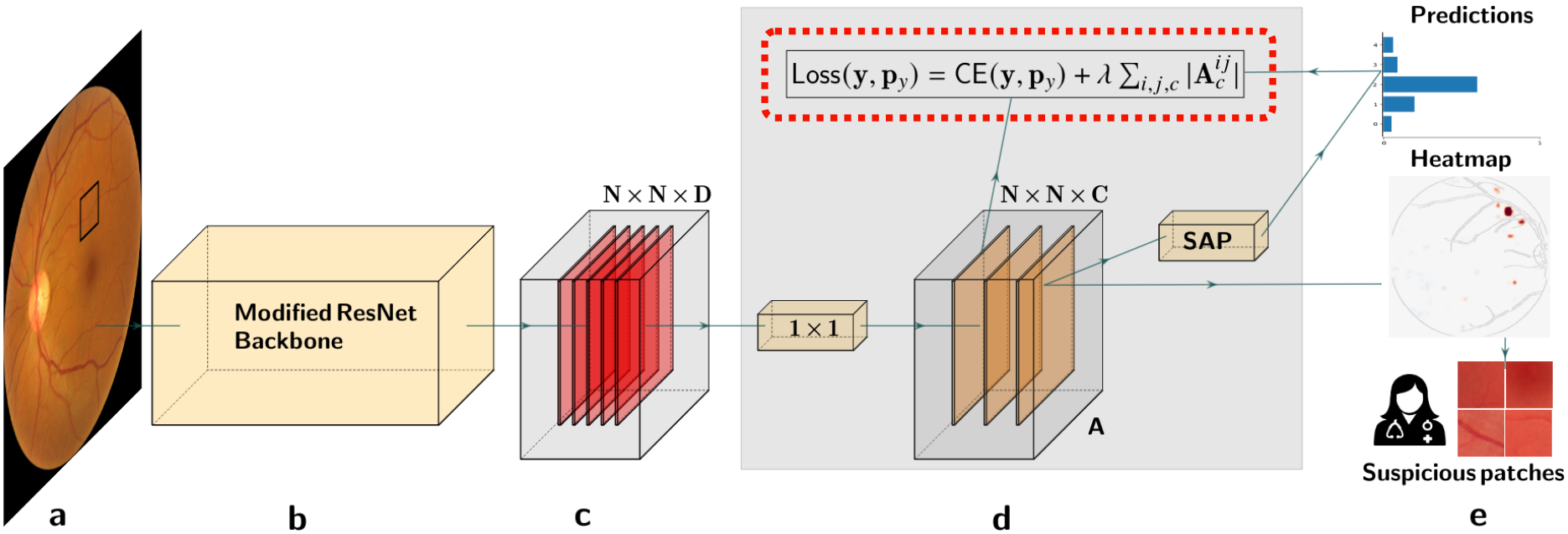
Contribution: sparse BagNet

We propose an inherently interpretable model that combines the feature extraction capabilities of DNNs with the advantages of sparse linear models in interpretability.



Contribution: sparse BagNet

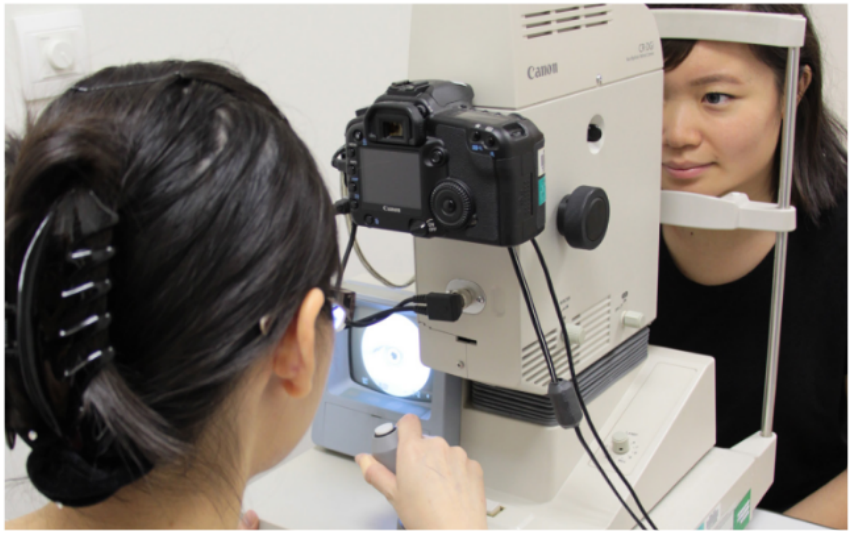
We propose an inherently interpretable model that combines the feature extraction capabilities of DNNs with the advantages of sparse linear models in interpretability.



Application: Diabetic Retinopathy (DR) detection

About DR

- Microvascular complications
- Leading cause of blindness
- Affects 1 in 3 diabetes patients



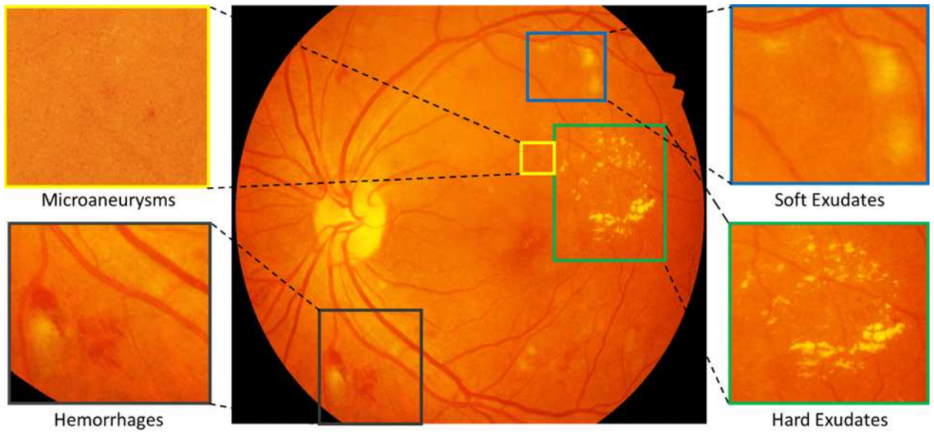
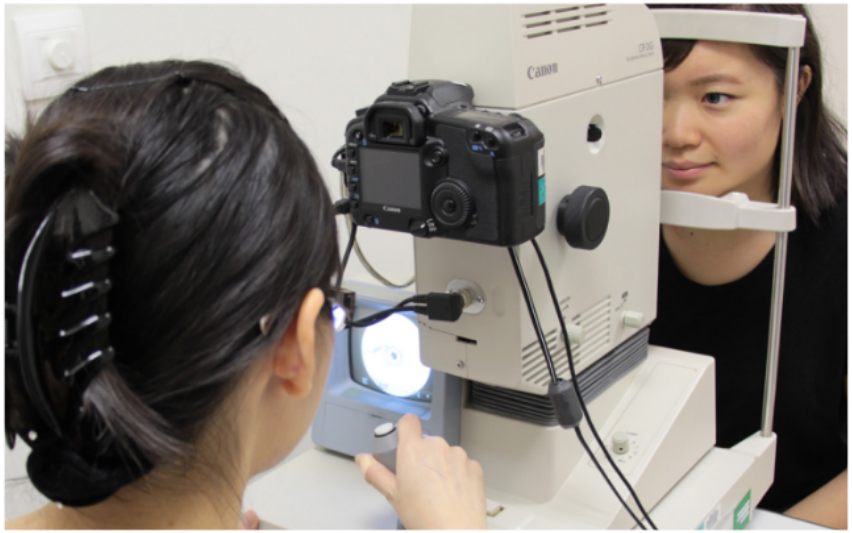


Figure reproduced from Porwal et al. (2018)

Application: Diabetic Retinopathy (DR) detection

About DR

- Microvascular complications
- Leading cause of blindness
- Affects 1 in 3 diabetes patients



Managing DR

- Diagnosis and grading based on retinal fundus images
- Early diagnosis and treatment
- Regular monitoring + screening

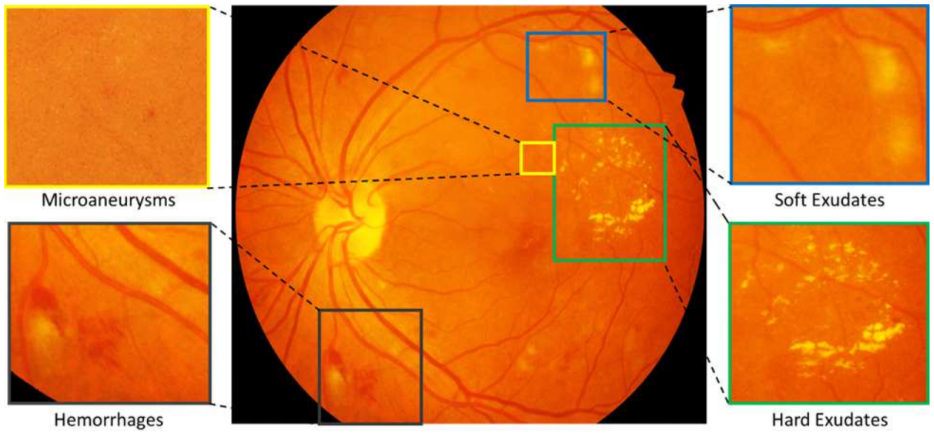


Figure reproduced from Porwal et al. (2018)

Referable DR detection: classification performances

Dataset description:

- Kaggle fundus data set for DR detection
- Referable DR classification task: $\{0,1\}$ vs $\{2,3,4\}$

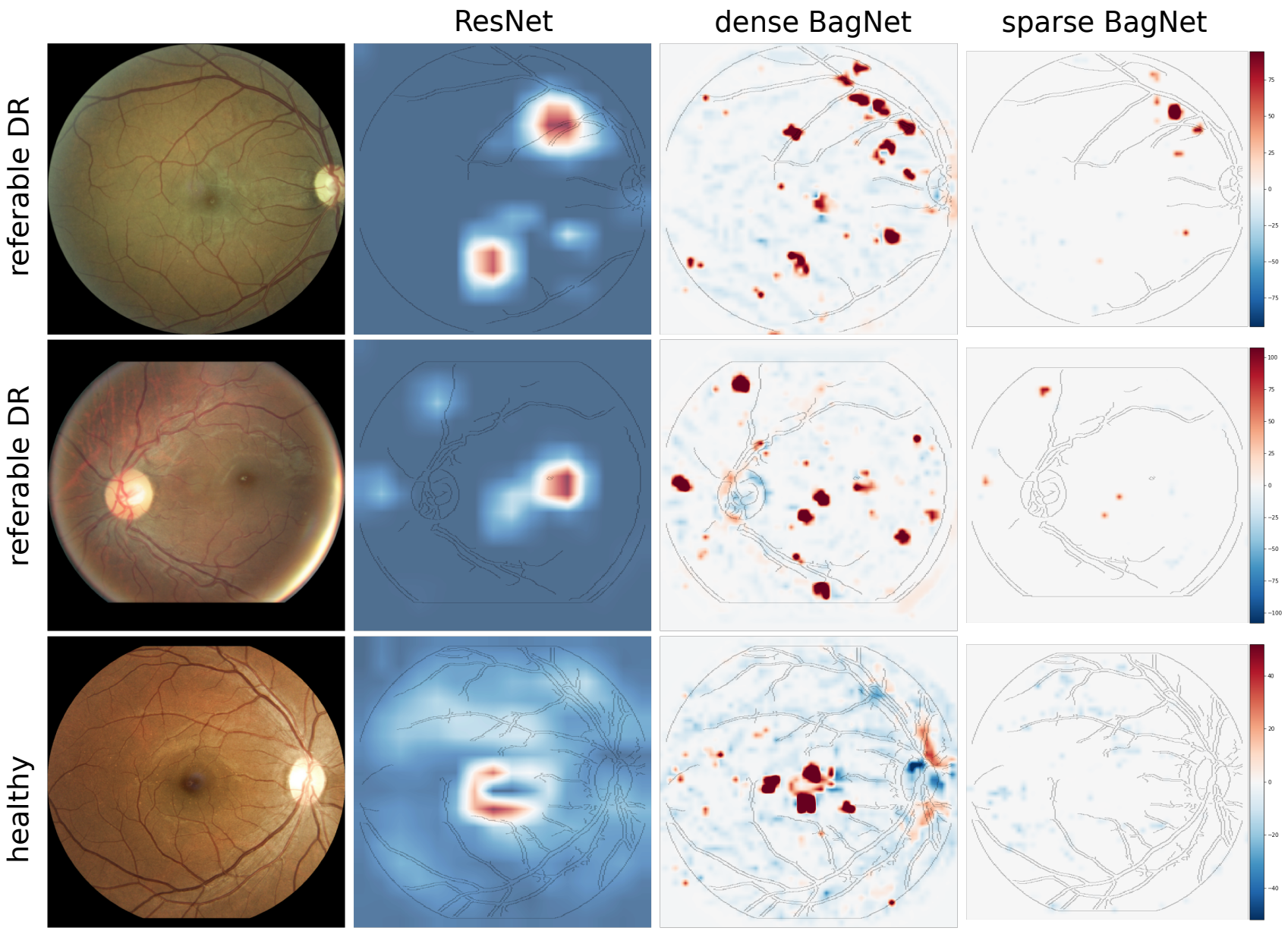
Referable DR detection: classification performances

Dataset description:

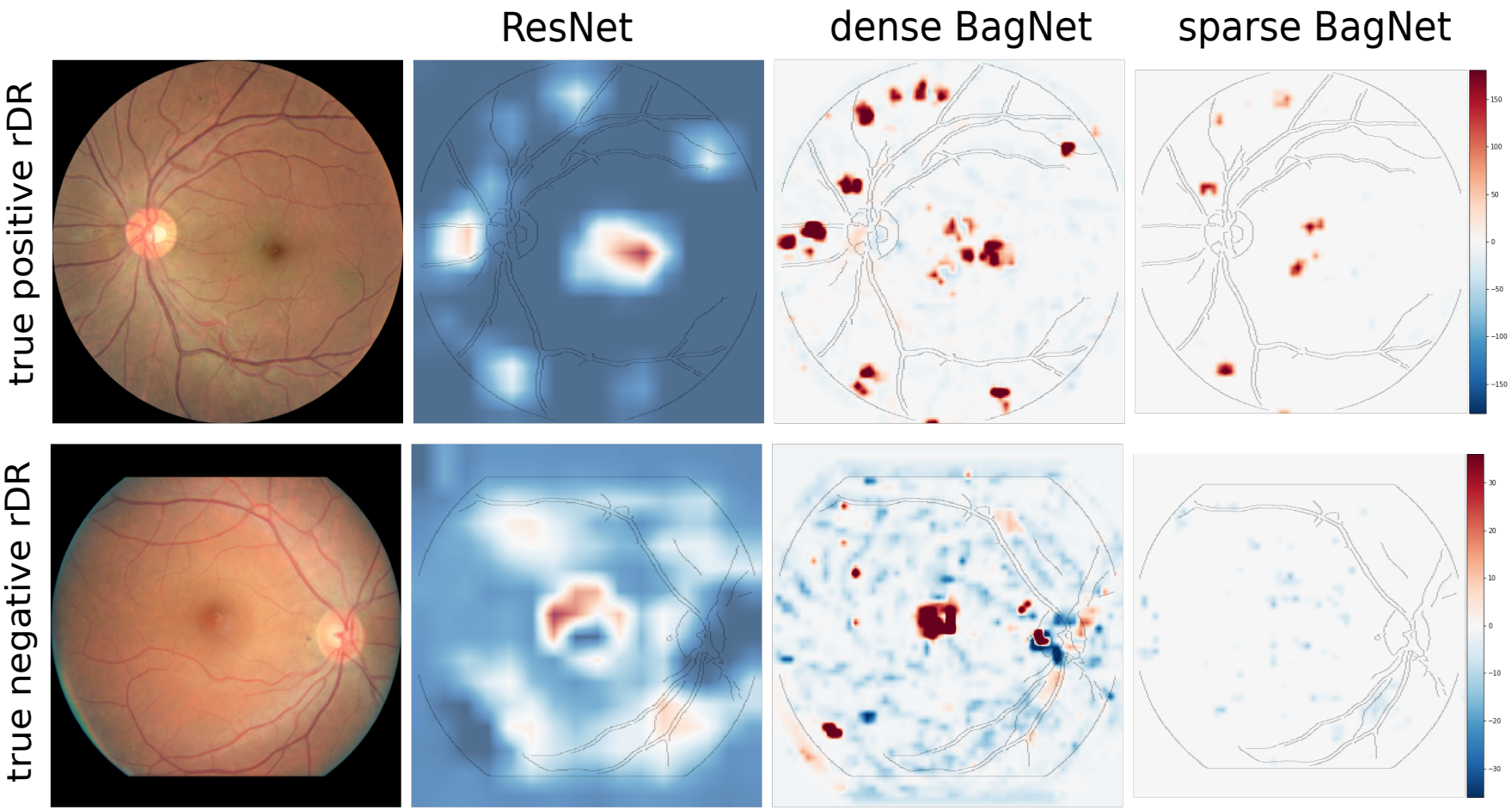
- Kaggle fundus data set for DR detection
- Referable DR classification task: {0,1} vs {2,3,4}

	Accuracy	**AUC**
ResNet-50	0.942	0.960
Dense BagNet	0.936	0.957
Sparse BagNet	0.928	0.937

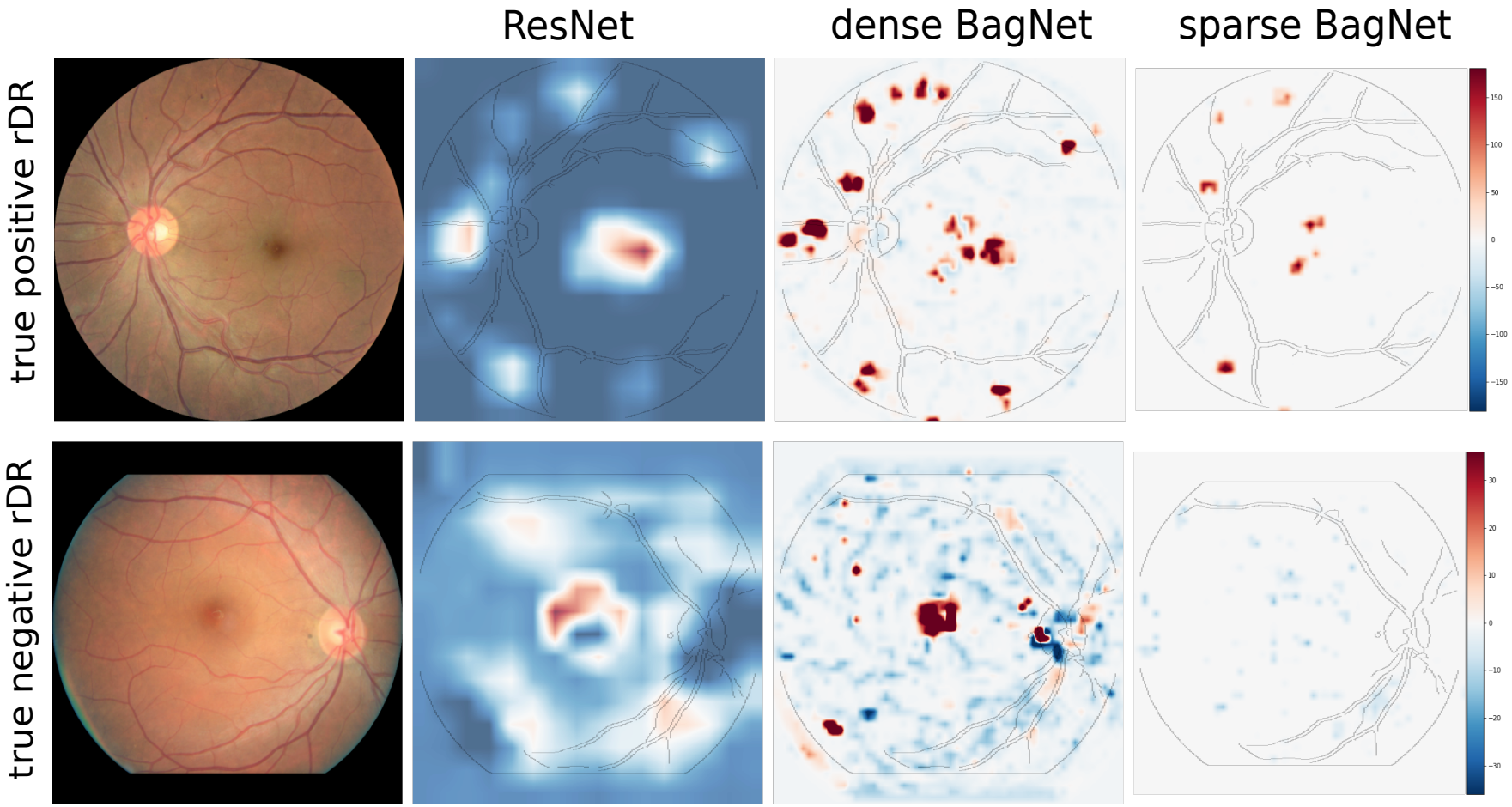
Referable DR detection: qualitative heatmap evaluation



Referable DR detection: quantitative heatmap evaluation



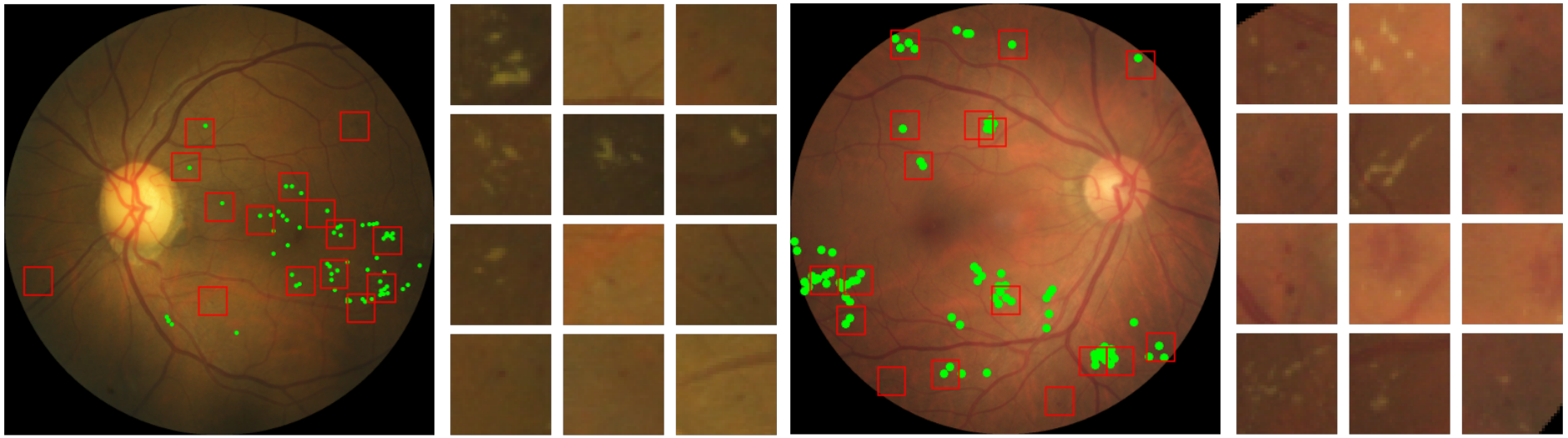
Referable DR detection: quantitative heatmap evaluation



	$\uparrow \mathbf{r}^{-}_{\mathbf{LG}}$	$\uparrow \mathbf{r}^{+}_{\mathbf{LG}}$
Dense BagNet	0.922 ± 0.03	0.145 ± 0.06
Sparse BagNet	0.991 ± 0.04	0.374 ± 0.33

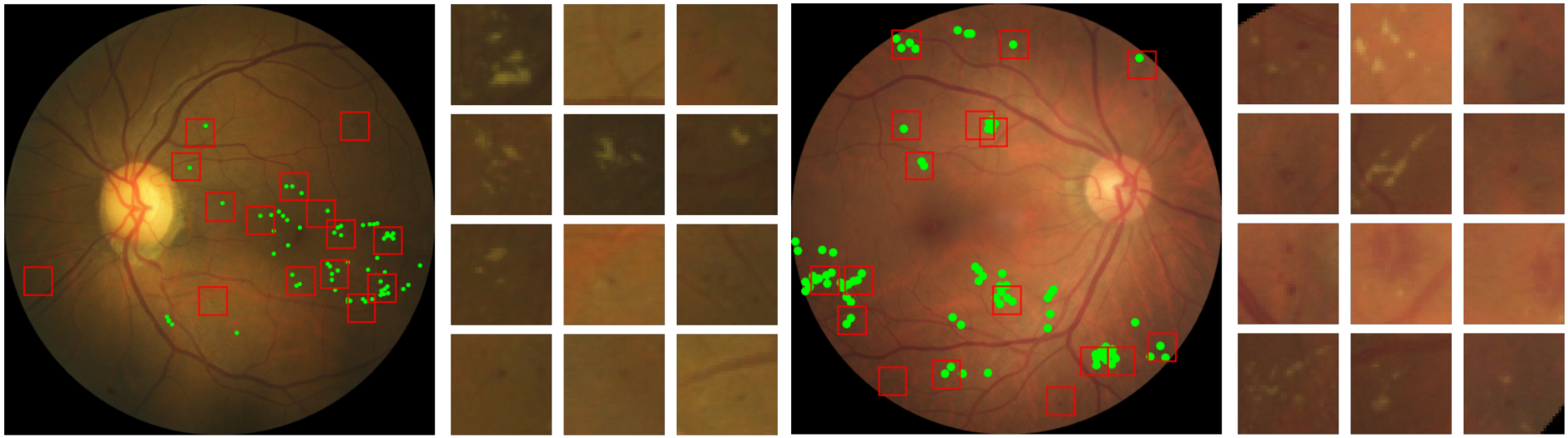
Referable DR detection: quantitative heatmap evaluation

- Lesion annotations (15 images)



Referable DR detection: quantitative heatmap evaluation

- Lesion annotations (15 images)



	Precision
Dense BagNet	0.219 ± 0.1
Sparse BagNet	0.791 ± 0.1

Multiclass DR detection: classification performances

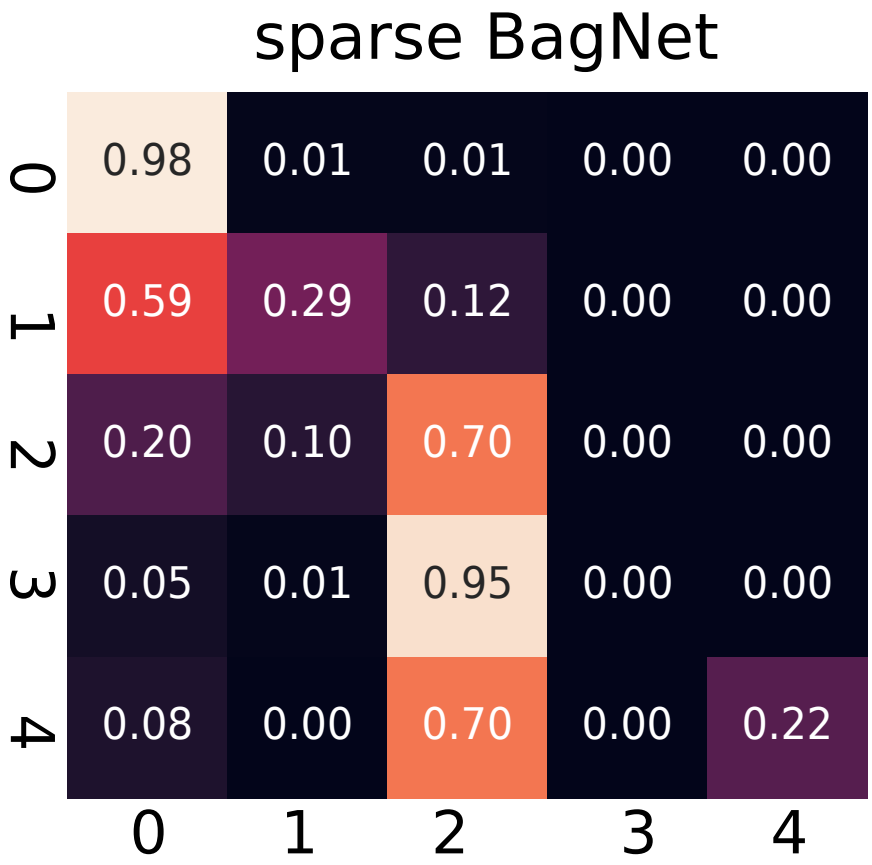
- Multiclass task
- $\{0\}$ vs $\{1\}$ vs $\{2\}$ vs $\{3\}$ vs $\{4\}$

	Acc.
ResNet-50	0.862
Dense BagNet	0.864
Sparse BagNet	0.850

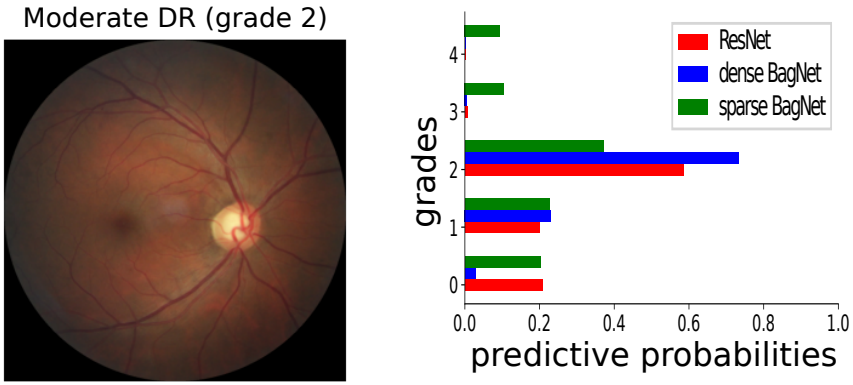
Multiclass DR detection: classification performances

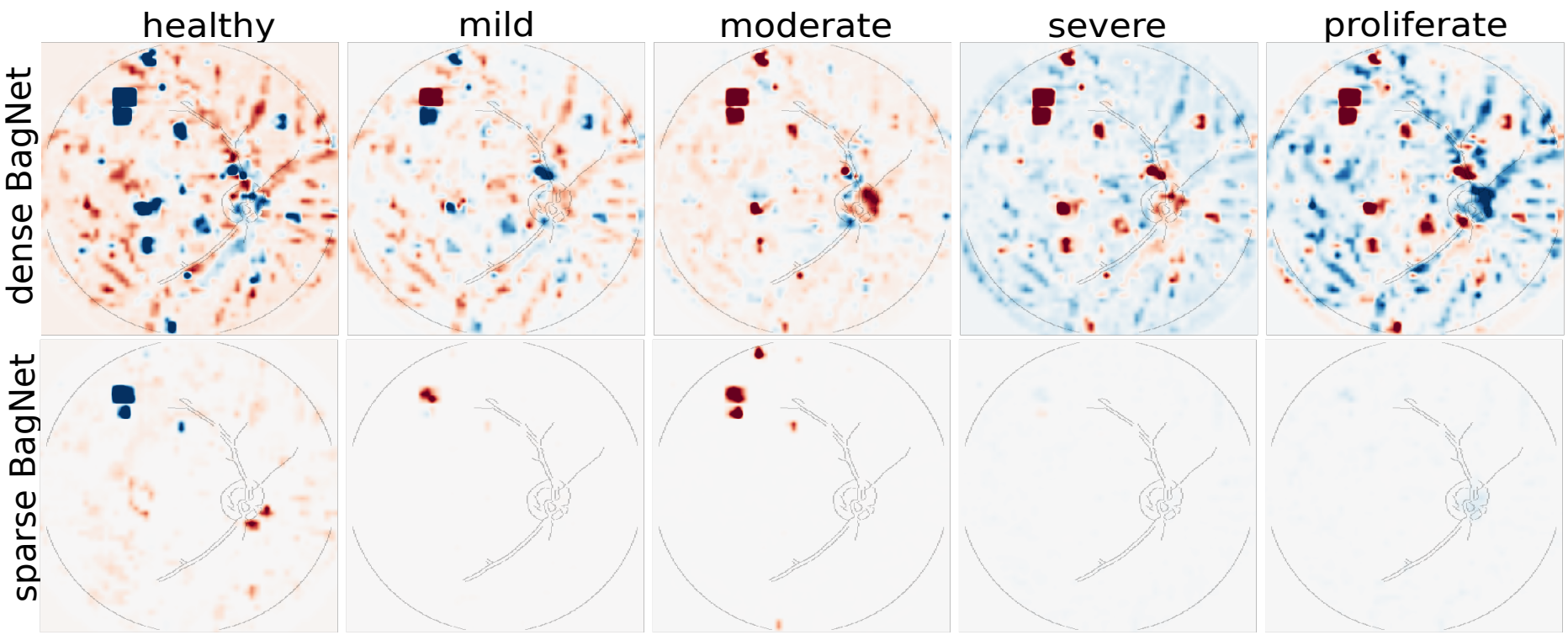
- Multiclass task
- {0} vs {1} vs {2} vs {3} vs {4}

	Acc.
ResNet-50	0.862
Dense BagNet	0.864
Sparse BagNet	0.850



Multiclass DR detection: qualitative heatmap evaluation





Conclusion

Key contributions

1. BagNet's modifications
2. Inherently interpretable attribution map
3. Quantitative metrics to access the interpretability

Conclusion

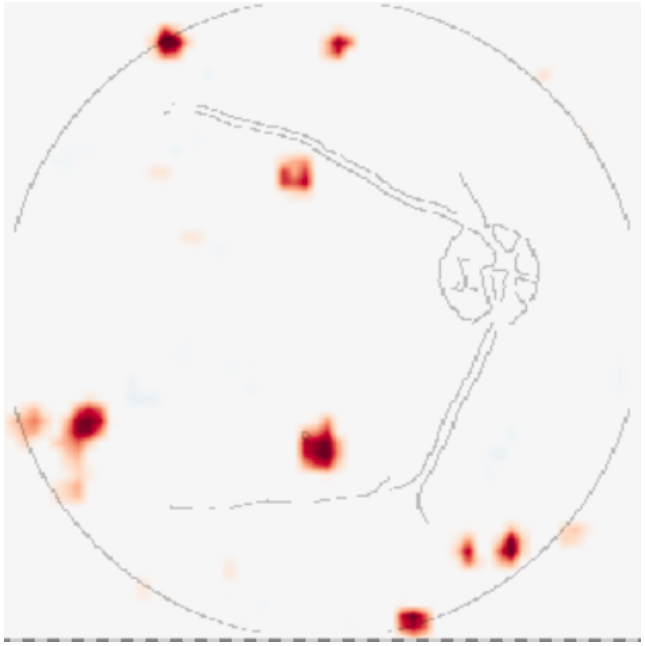
Key contributions

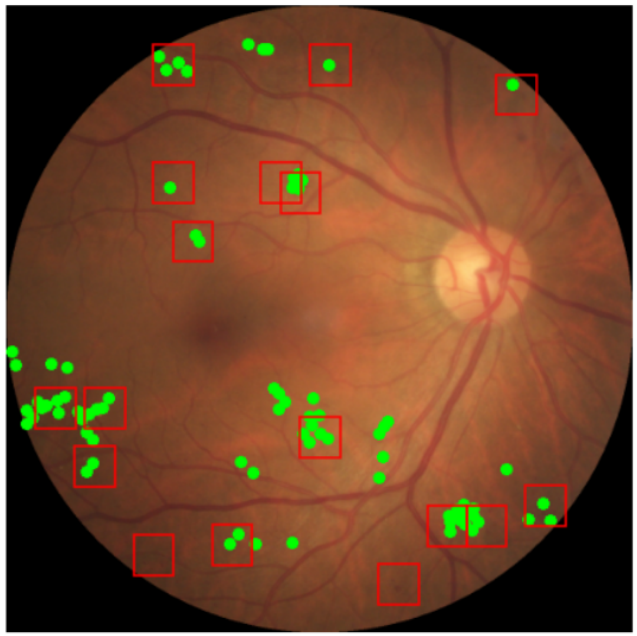
1. BagNet's modifications
2. Inherently interpretable attribution map
3. Quantitative metrics to access the interpretability

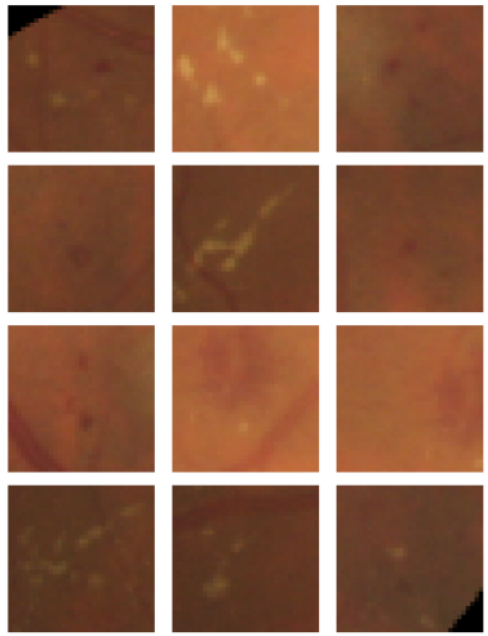
Summary

- Enforcing sparse heatmap improves the interpretability
- Suitable for small lesions distributed over the image
- Comparable SOTA performance on DR detection

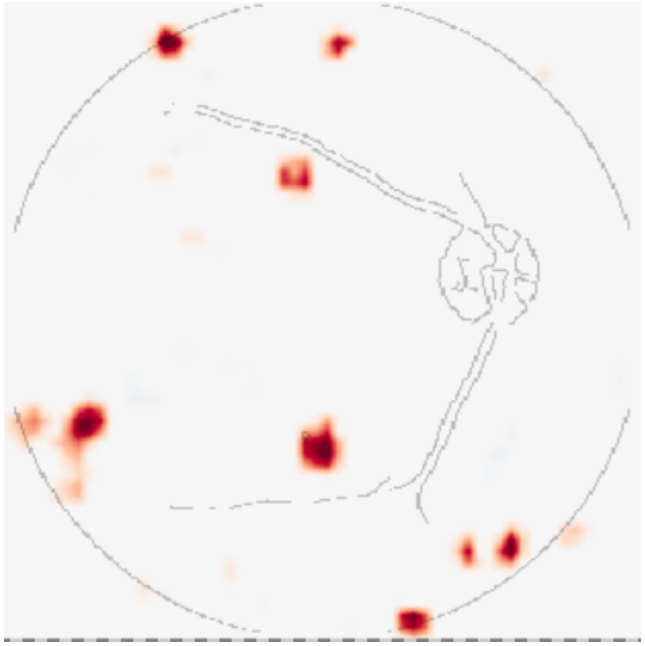
Next step

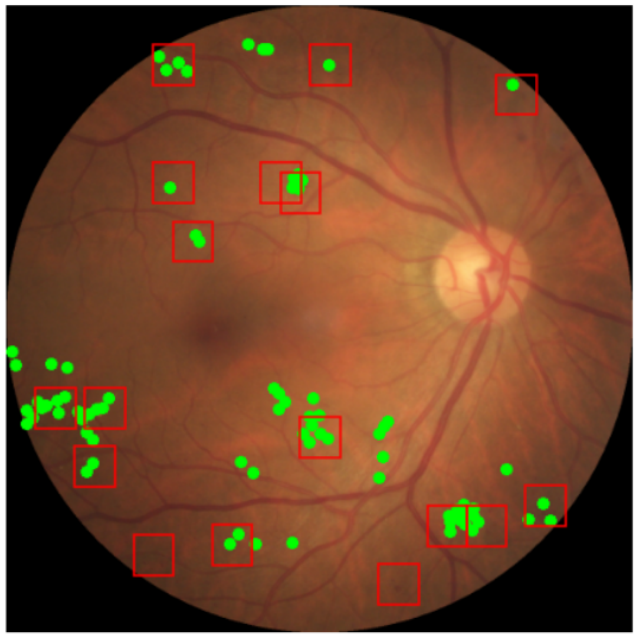


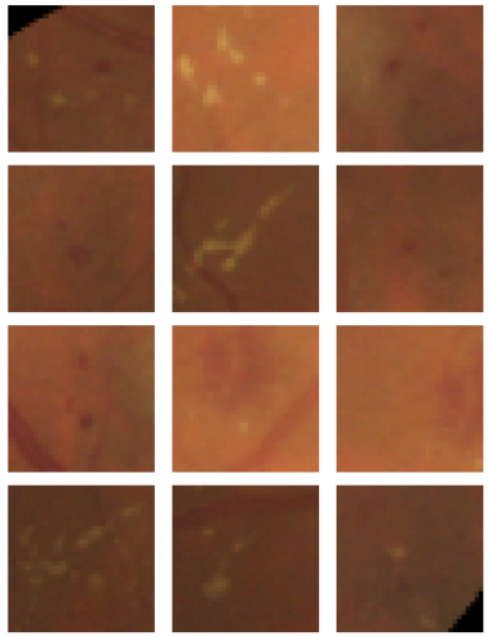




Next step



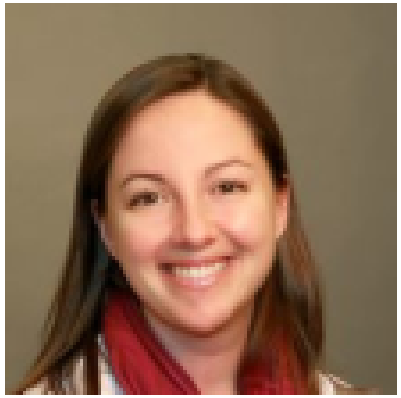
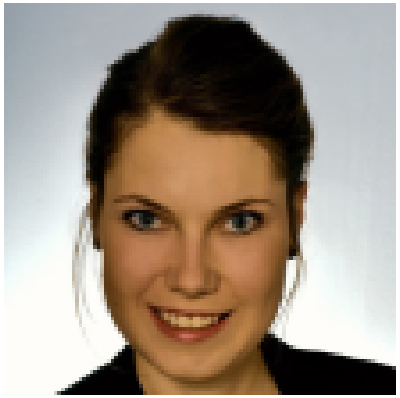




Clinical study

Does highlighting suspicious areas of the image improve diagnostic performance?

Acknowledgements



GitHub

Paper

Thank you for your attention!

Poster location: **W08**

kerol.djoumessi@uni-tuebingen.de

Dataset

Dataset Description:

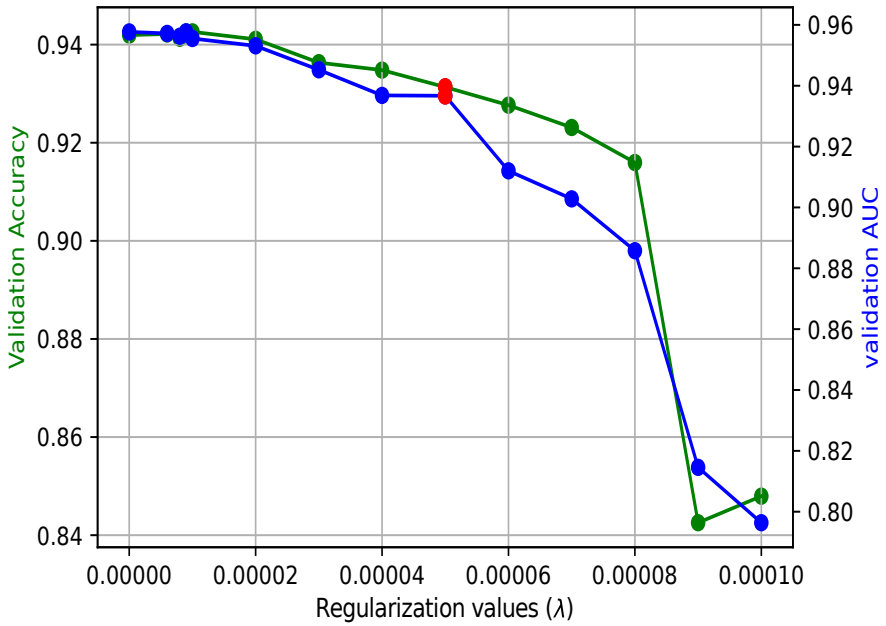
- Kaggle dataset: $88,702$
- Quality filtering: $45,923$
- Referable task: $\{0,1\}$ vs $\{2,3,4\}$
- Multiclass task: $\{0\}$ vs $\{1\}$ vs $\{2\}$ vs $\{3\}$ vs $\{4\}$

Method: Model's architecture

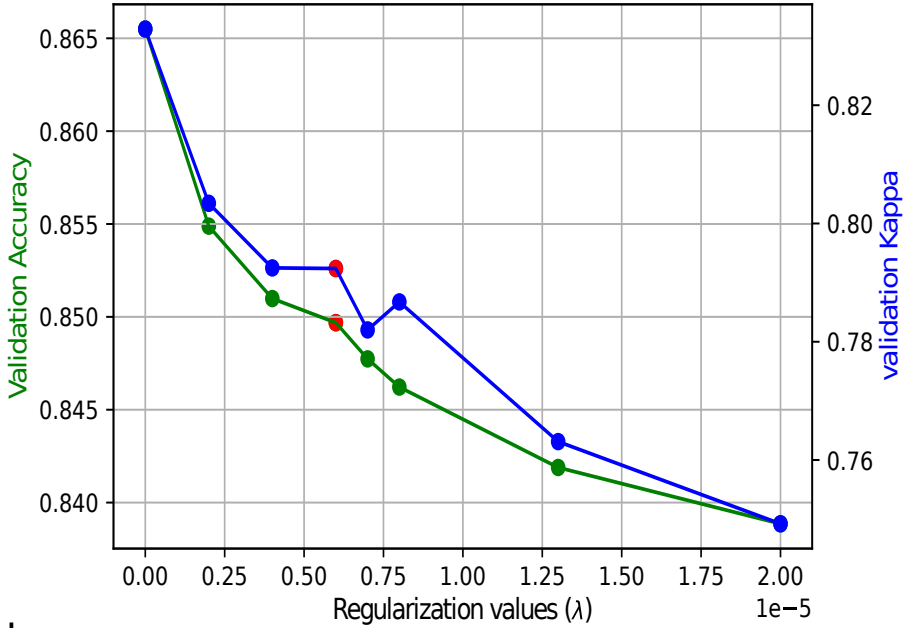
- BagNet encodes the input:
  - $f_{bag}(X) = A' \in R^{N \times N \times D}$
  - $A' \xrightarrow{C_{1\times1}} A \in R^{N \times N \times C}$
  - $\mathbf{A} \xrightarrow{\mathbf{SAP}} \mathbf{l_c} \in \mathbf{R^{1 \times C}}$
- Patch-based model

- Small receptive field
- Heatmap = evidence map
- λ controls the sparsity
- Suspicious patch visualization

Sparsity hyperparameter λ



a



b

Sparse BagNet limitations