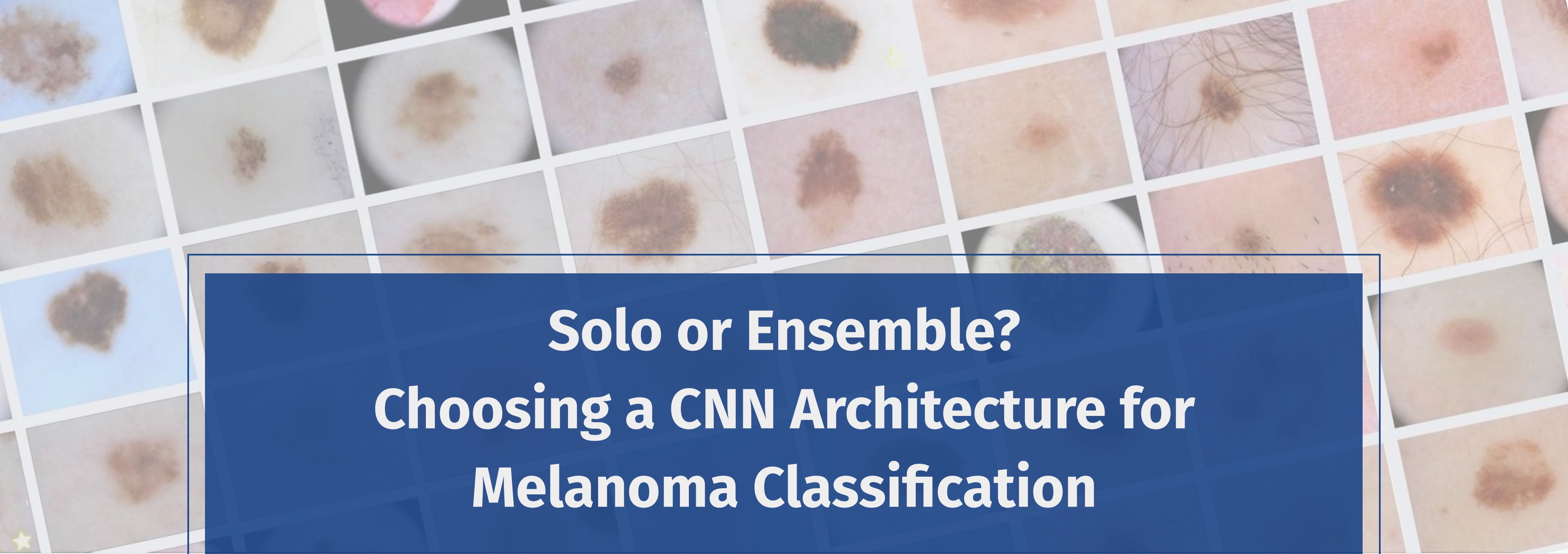


Solo or Ensemble? Choosing a CNN Architecture for Melanoma Classification

Fábio Perez[1], Sandra Avila[2], Eduardo Valle[1]
[1]Recod Lab., DCA, FEEC, University of Campinas (UNICAMP)
[2]Recod Lab., IC, University of Campinas (UNICAMP)

ISIC Workshop @ CVPR 2019

Convolutional Neural Networks

SotA for most computer vision problems, including **skin lesion analysis**

Used by all **winner submissions** in **ISIC Challenges** 2016, 2017, 2018

CNN Architectures

AlexNet

CNN Architectures

AlexNet

GoogLeNet

VGG

ZFNet

CNN Architectures

NASNet
ResNet
Inception
GoogLeNet
AlexNet
DenseNet
DualPathNet
Xception
MobileNet
ZFNet
SE-Net
VGG
ResNeXt
PNASNet
Inception-ResNet
SqueezeNet

CNN Architectures

ISIC Challenges

2016 ResNet

2017 ResNet, Inception

2018 ResNet, Inception, DenseNet, ResNeXt
PNASNet, DPN, SENet...

Transfer Learning

The **most critical factor** for model performance

SotA for most computer vision problems, including **skin lesion analysis**

Also used by all **ISIC Challenges winners**

Valle et al. (2017). Data, Depth, and Design: Learning Reliable Models for Melanoma Screening https://arxiv.org/abs/1711.00441
Menegola et al. (2017). Knowledge Transfer for Melanoma Screening with Deep Learning https://arxiv.org/abs/1703.07479

Do better ImageNet models transfer better?

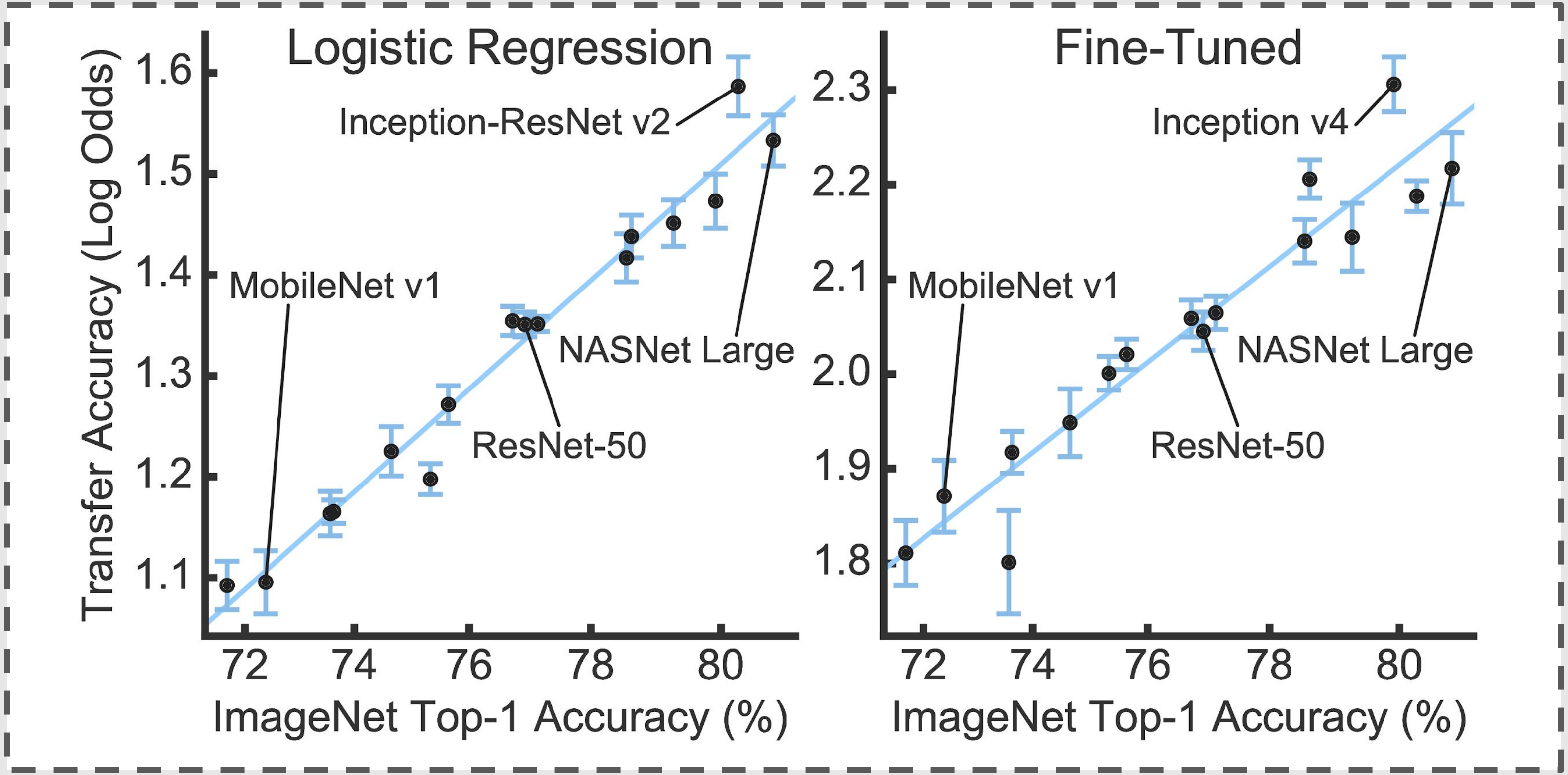
Short answer: **Yes**

For multiple natural datasets

Fine-tuning, fixed features, and random initialization

Kornblith et al. (2018) https://arxiv.org/abs/1805.08974

Do better ImageNet models transfer better?



Kornblith et al. (2018) arxiv.org/abs/1805.08974

How to predict model performance?

Experimental Design

9 architectures
× 5 splits
× 3 replicates
= **135 experiments**

Experimental Design

9 architectures

× 5 splits

× 3 replicates

= **135 experiments**

DenseNet
Dual Path Nets
Inception-v4
Inception-ResNet-v2
MobileNetV2
PNASNet
ResNet
SENet
Xception

Experimental Design

9 architectures
× 5 splits
× 3 replicates
= **135 experiments**

ISIC 2017
1750 train
500 validation
500 test

Explored factors

Architectural

Acc@1 on ImageNet
# of Parameters
Date of Publication

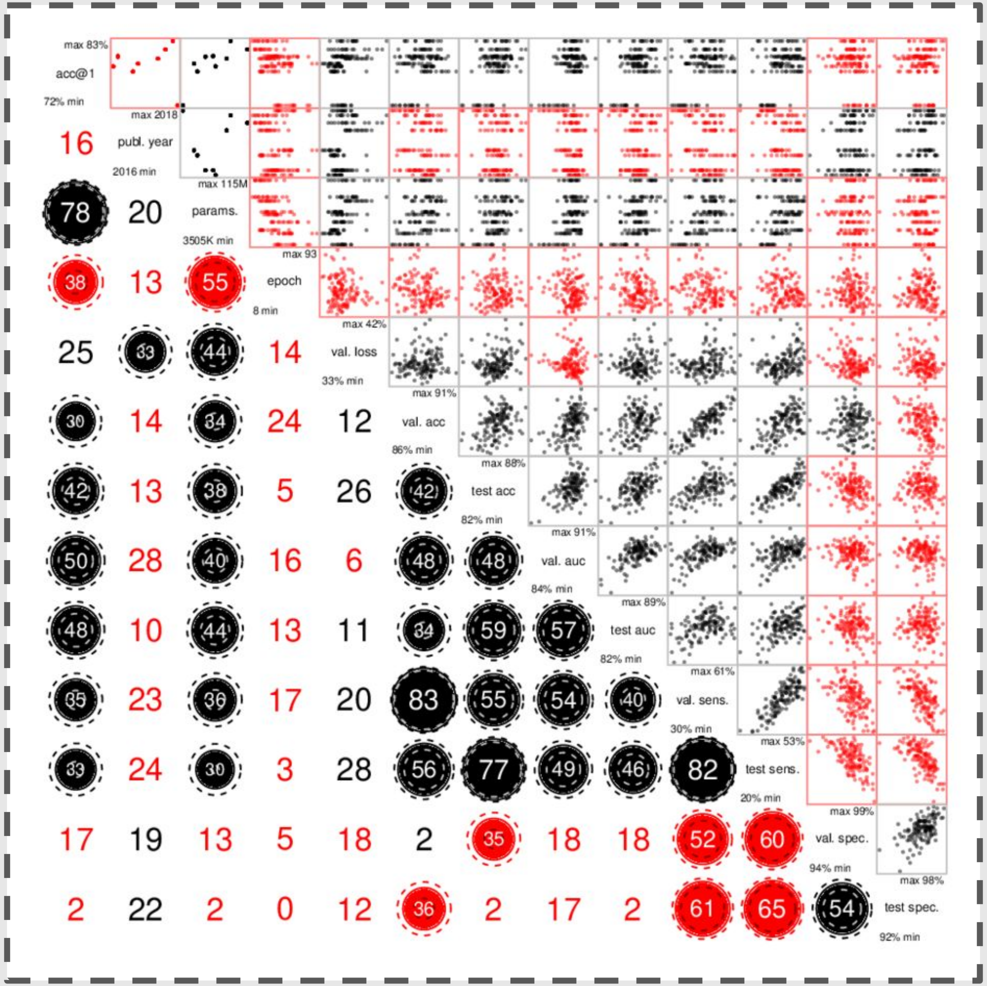
Training

AUC
Accuracy
Sensitivity
Specificity
➡ Validation
Test

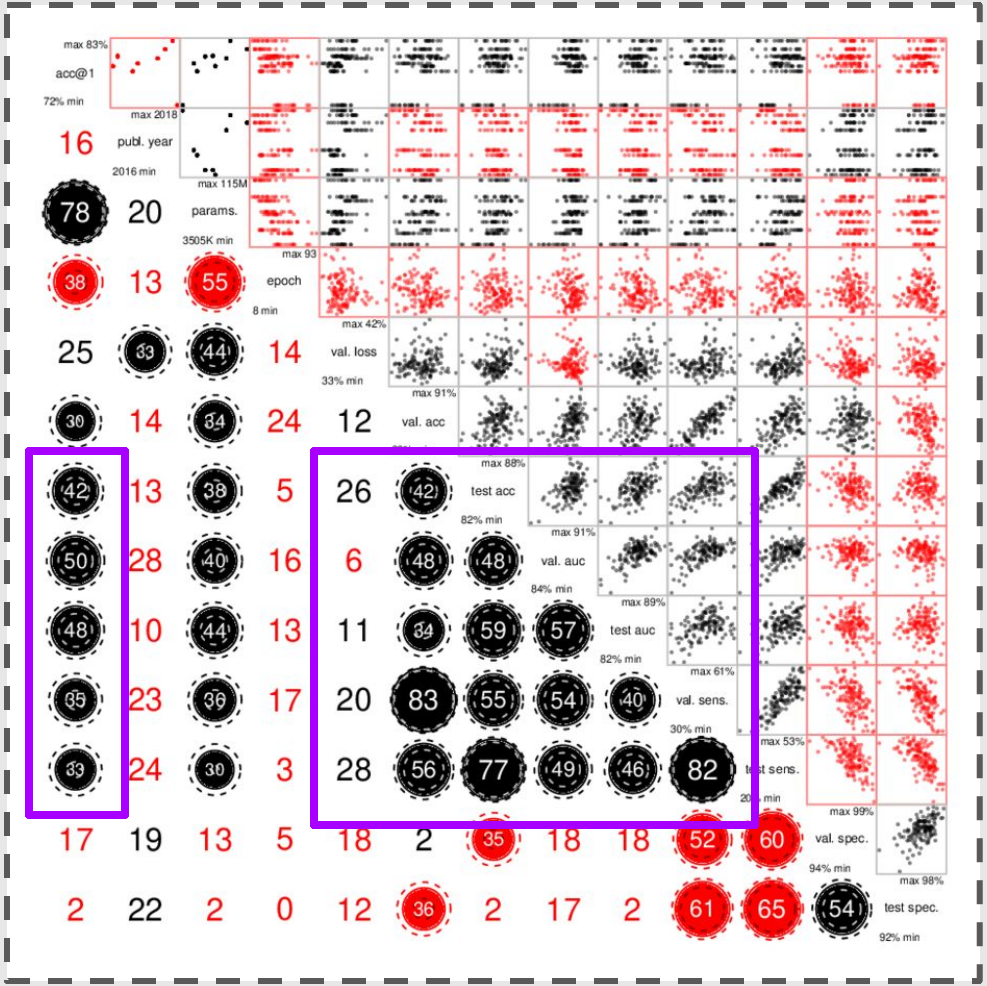
Loss ➡ Validation

# of Epochs

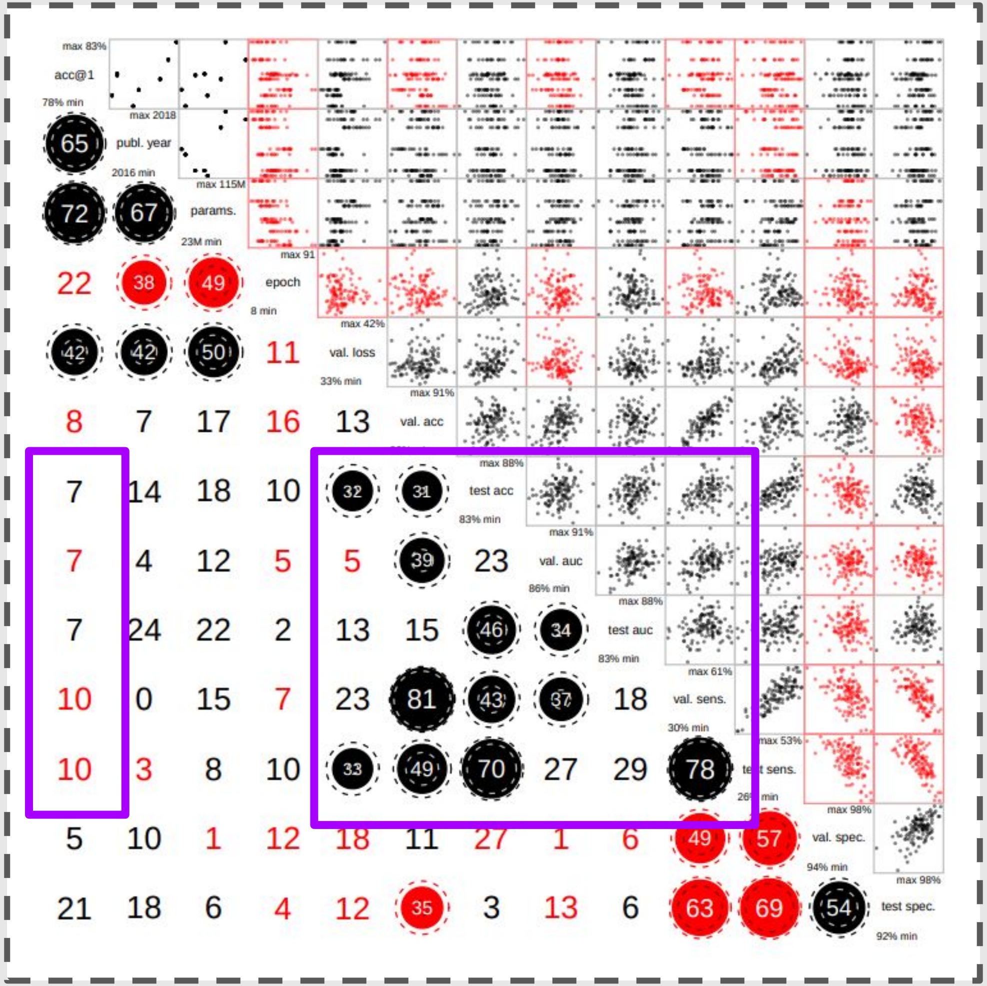
Results


max 83%
acc@1
72% min
16
max 2018
publ. year
2016 min
78 20
max 115M
params.
3505K min
38 13 55
max 93
epoch
8 min
25 33 44 14
max 42%
val. loss
33% min
30 14 34 24 12
max 91%
val. acc
86% min
42 13 38 5 26 42
max 88%
test acc
82% min
50 28 40 16 6 48 48
max 91%
val. auc
84% min
48 10 44 13 11 34 59 57
max 89%
test auc
82% min
35 23 30 17 20 83 55 54 40
max 61%
val. sens.
30% min
33 24 30 3 28 56 77 49 46 82
max 53%
test sens.
20% min
17 19 13 5 18 2 35 18 18 52 60
max 99%
val. spec.
94% min
2 22 2 0 12 36 2 17 2 61 65 54
max 98%
test spec.
92% min

Results


acc@1
max 83%
72% min
publ. year
max 2018
2016 min
params.
max 115M
3505K min
epoch
max 93
8 min
val. loss
max 42%
33% min
val. acc
max 91%
test acc
max 88%
82% min
val. auc
max 91%
84% min
test auc
max 89%
82% min
val. sens.
max 61%
30% min
max 53%
val. spec.
max 99%
94% min
test spec.
max 98%
92% min
16
78 20
38 13 55
25 33 44 14
30 14 34 24 12
42 13 38 5 26 42
50 28 40 16 6 48 48
48 10 44 13 11 34 59 57
35 23 36 17 20 83 55 54 40
33 24 30 3 28 56 77 49 46 82
17 19 13 5 18 2 35 18 18 52 60
2 22 2 0 12 36 2 17 2 61 65 54

Results
(without MobileNetV2)



Datasets

Kornblith et al. (2018)

- ➔ Multiple large datasets
- ➔ One factor: Acc@1
- ➔ Hyperparameter tuning

Ours

- ➔ ISIC 2017 (2750 images)
- ➔ Multiple factors
- ➔ “Best-practice” hyperparameters

Datasets

Kornblith et al. (2018)

- ➔ Multiple large datasets
- ➔ One factor: Acc@1
- ➔ Hyperparameter tuning
- ➔ **One split per dataset**
- ➔ **No replicates**

Ours

- ➔ ISIC 2017 (2750 images)
- ➔ Multiple factors
- ➔ “Best-practice” hyperparameters
- ➔ **Five splits**
- ➔ **Three replicates**

 19

Ensembles

Creating the Ensembles

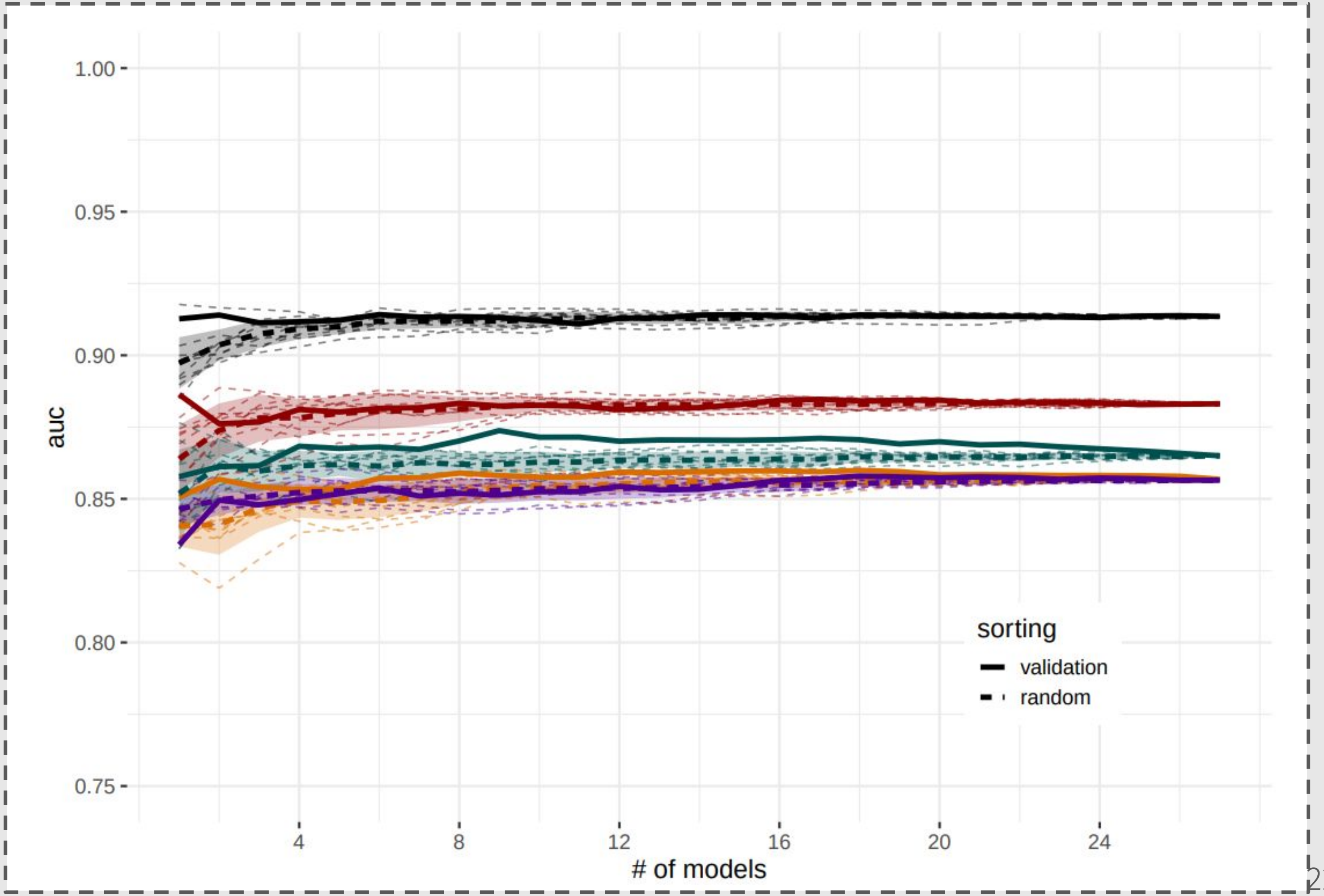
9 architectures × 3 replicates = 27 models per split

For **each split**, ensemble **1, 2, …, 27 models**

Two **strategies** for adding models:

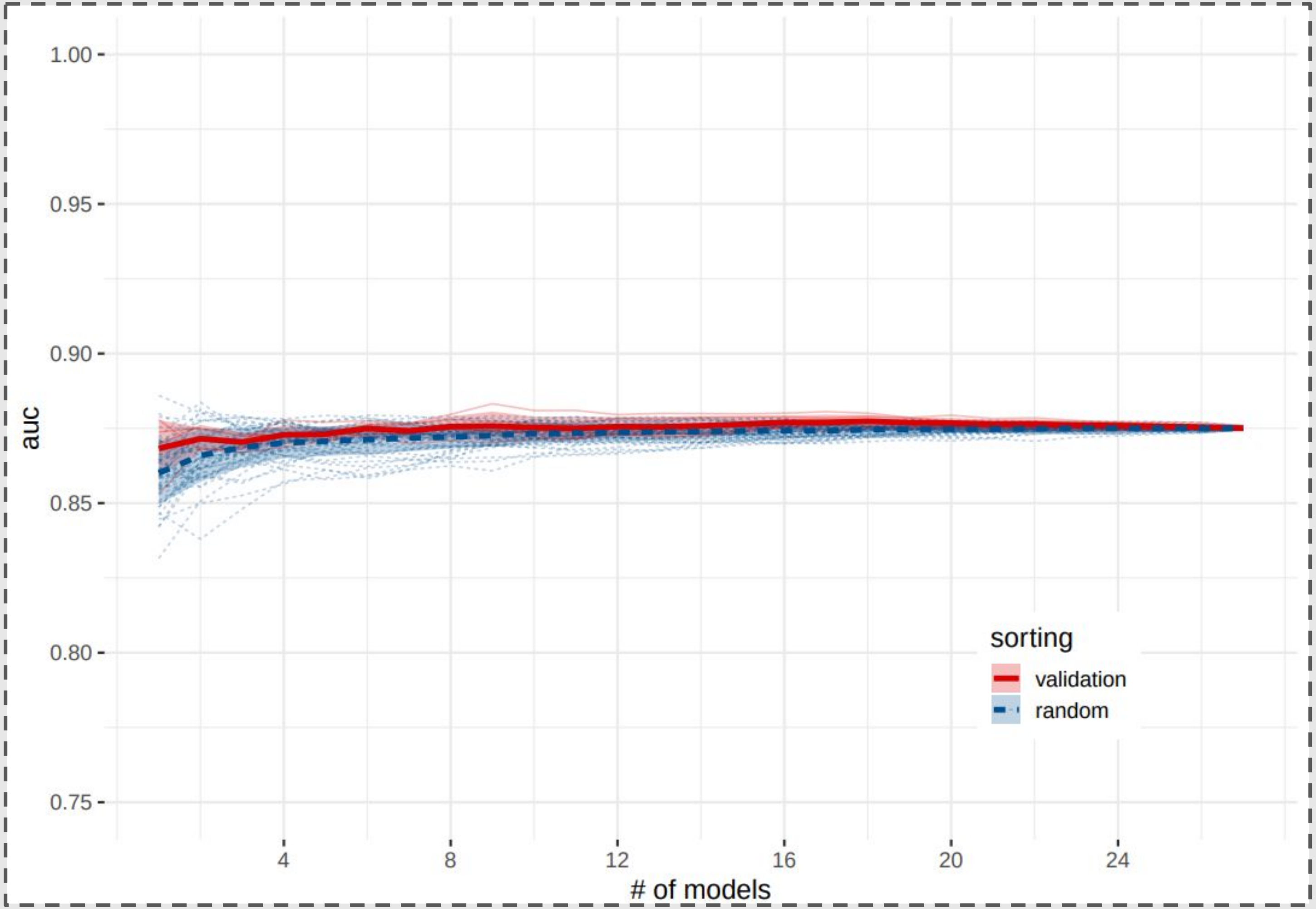
- in **random** order
- models with best **validation AUC** first

Results



1.00
0.95
0.90
0.85
0.80
0.75
auc
4
8
12
16
20
24
# of models
sorting
validation
random

Results
(normalized)



Conclusions

For the SotA models, performance on ImageNet **does not necessarily translate** to performance on melanoma detection

Validation metrics correlate with test metrics much better much better than validation loss

Ensembles are needed for stable SotA performance; large ensembles work okay from simply picking at random from a pool of SotA individual models

Acknowledgments

UNICAMP

RECOD
reasoning for complex data

CNPq

FAPESP

eScience
UNICAMP

NVIDIA

Azure

Thanks!