Studying the Impact of Augmentations on Medical Confidence Calibration

Adrit Rao[1,2], Joon-Young Lee[3], Oliver Aalami[2]

[1]Palo Alto High School, [2]Stanford University, [3]Adobe Research

Introduction

- CNNs are often prone to **overconfidence**, impacting the **reliability** of uncertainty measures
- This can affect the **clinical confidence** in medical image analysis systems
- Modern augmentations show promise in both performance improvement and calibration on ***general benchmarks***
- This study aims to validate modern augmentation effectiveness in **medical confidence calibration** across various modalities (CT, CXR, MRI, and Derm).
- Unconventional image modifications, such as feature combination or removal, **may yield varied effects on medical images.**

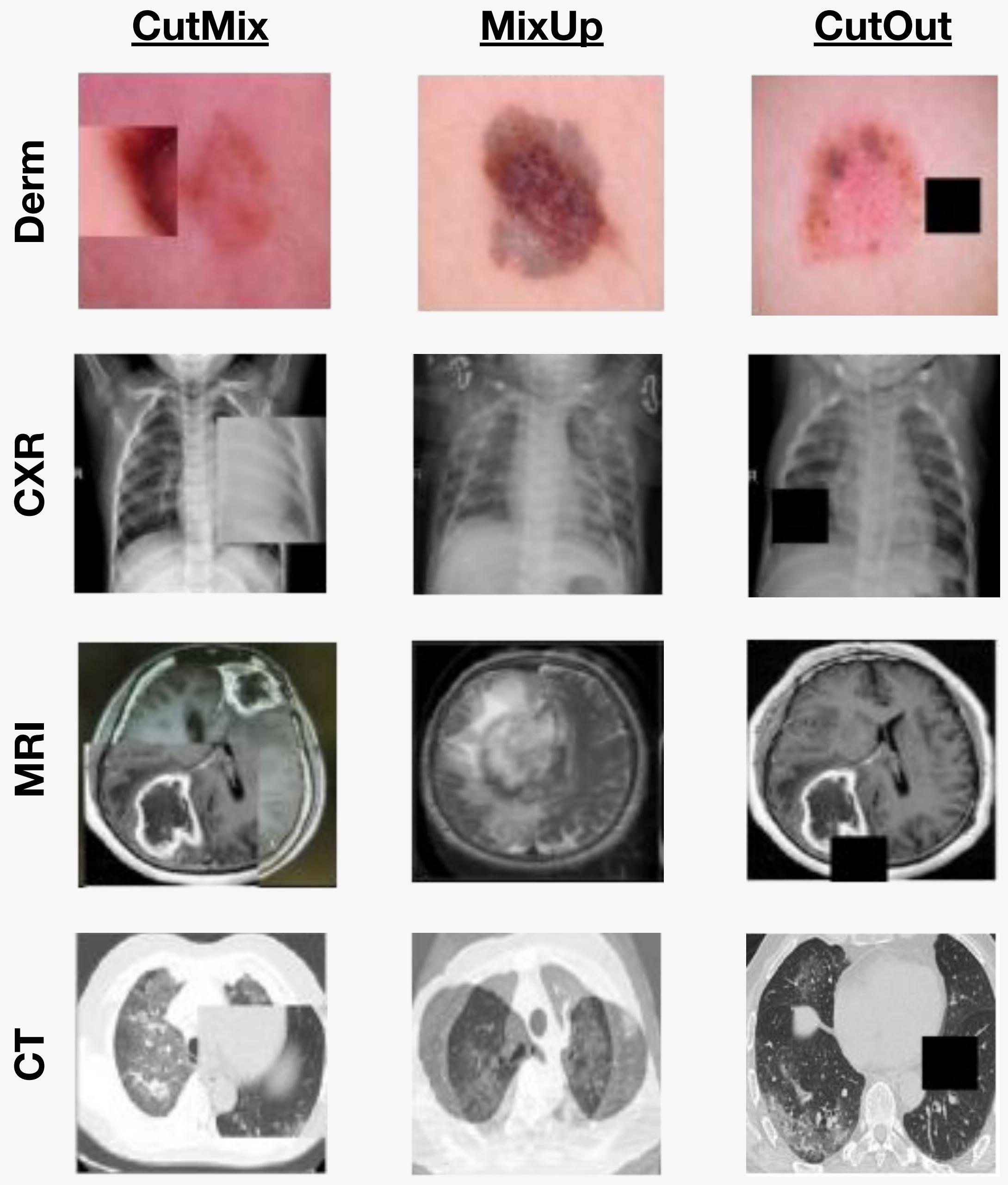

Figure 1: Samples of modern augmentations performed on different medical image modalities

Methods

- Train **four ResNet CNNs** on a medical dataset (one baseline, three augmented) and validate with **calibration and performance metrics** (ECE, reliability plotting, AUROC, accuracy)
- **Compare augmentation effects** on different **model sizes** (ResNet-50 & 101) and **modalities** using the results

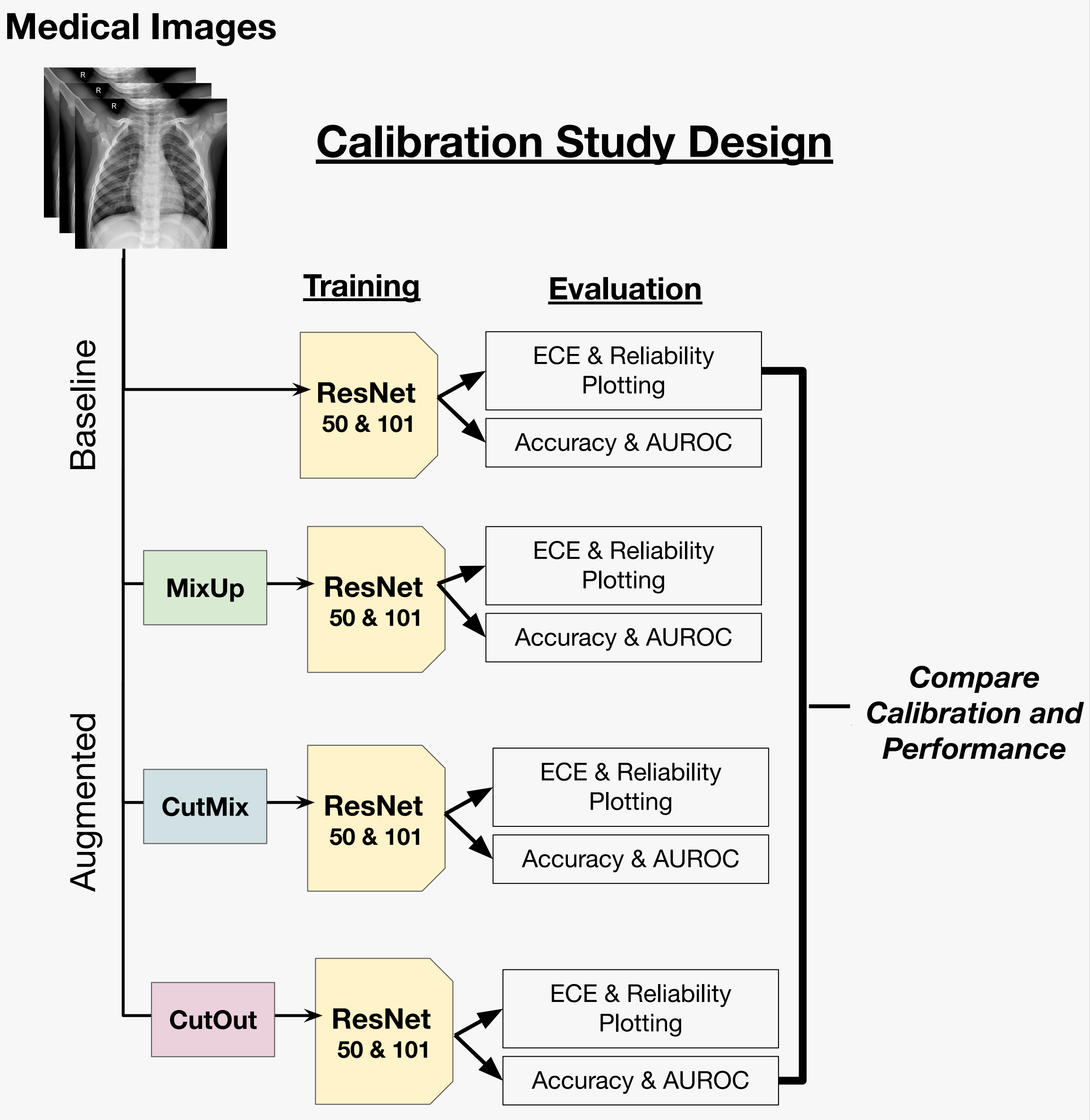

Figure 2: Design of our augmentation evaluation study

Qualitative Results

- **Reliability plots are generated** for ResNet-50 and ResNet-101 with each augmentation across the four medical image modalities
- Addition of modern augmentations **typically improves the line fit**, indicating **enhanced calibration**
  - **CutOut** in certain cases (such as CXR pneumonia) can be seen **significantly reducing line fit**
- **Standard ResNet**, in comparison, often exhibits notably **lower levels of line fit**

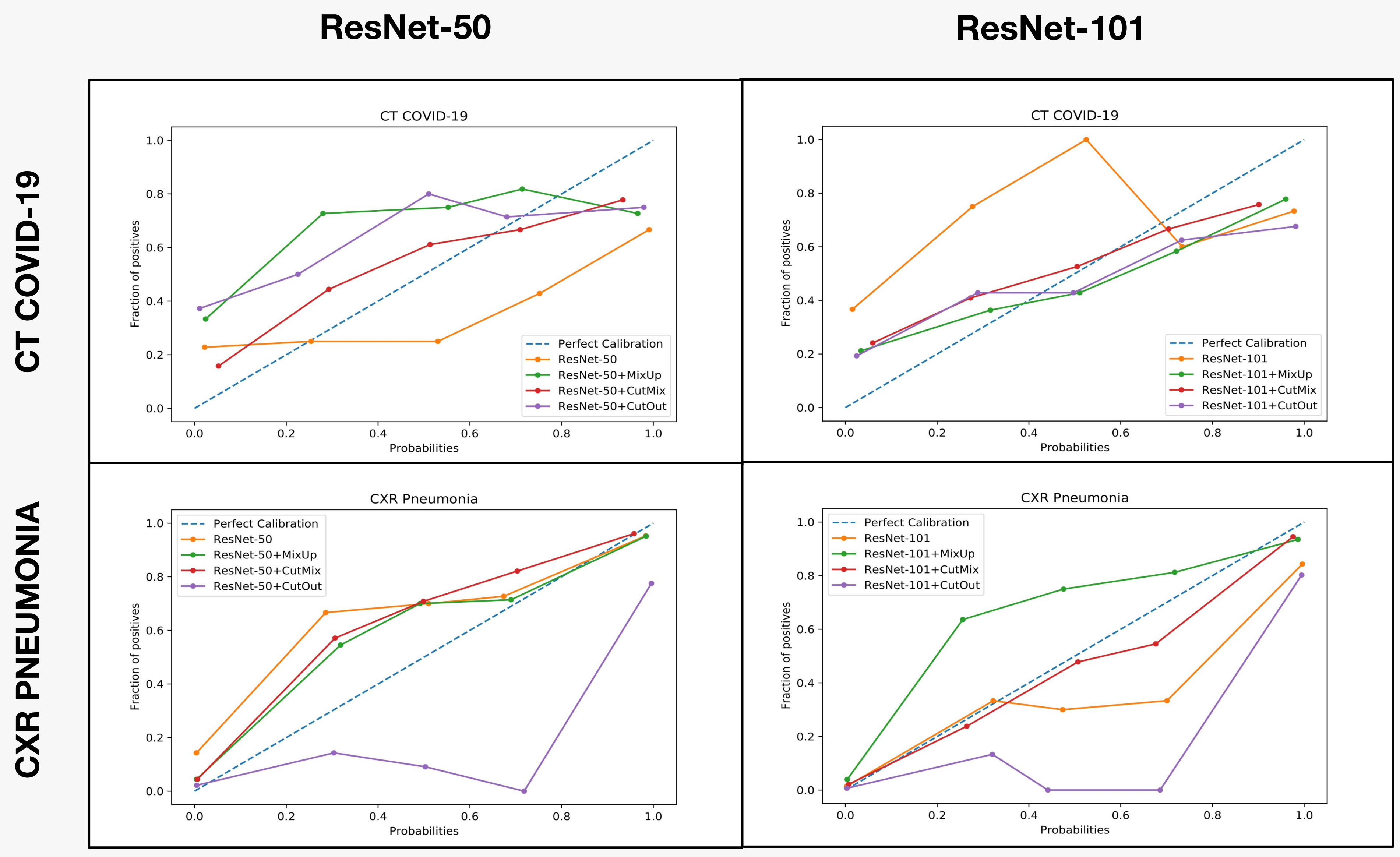

Figure 3: Confidence Calibration Reliability Plots for Modern Augmentations on CXR and CT modalities

Quantitative Results

- Quantitative results cover **performance** (accuracy and AUROC) and **calibration** (ECE) **evaluations**.
- **Augmented models generally improve accuracy and AUROC**, with exceptions in ResNet-101 MRI Tumor tests and ResNet-50 CT COVID-19.
- Regarding calibration, augmentation notably **reduces the ECE score**, especially with **MixUp** and **CutMix**
- However, **CutOut's** impact on calibration is mixed, with **both reductions and increases observed**

Model	Augmentation	Accuracy	AUROC
ResNet-50	None	0.927	0.944
ResNet-50	MixUp	**0.944**	**0.980**
ResNet-50	CutMix	0.941	0.977
ResNet-50	CutOut	0.917	0.941
ResNet-101	None	0.872	0.902
ResNet-101	MixUp	**0.939**	0.976
ResNet-101	CutMix	0.933	**0.977**
ResNet-101	CutOut	0.886	0.915

(a) CXR Pneumonia

Model	Augmentation	Accuracy	AUROC
ResNet-50	None	**0.700**	0.724
ResNet-50	MixUp	0.680	**0.757**
ResNet-50	CutMix	0.653	0.754
ResNet-50	CutOut	0.633	0.656
ResNet-101	None	0.653	0.706
ResNet-101	MixUp	**0.706**	**0.765**
ResNet-101	CutMix	0.613	0.708
ResNet-101	CutOut	0.673	0.746

(b) CT COVID-19

Dataset	Model	Baseline	MixUp	CutMix	CutOut
Derm	ResNet-50	0.1812	0.1424 (-0.0388)	**0.1286 (-0.0526)**	0.1726 (-0.0086)
Derm	ResNet-101	0.1676	0.1020 (-0.0656)	**0.0973 (-0.0703)**	0.1967 (+0.0291)
CXR	ResNet-50	0.0675	0.0409 (-0.0266)	**0.0351 (-0.0324)**	0.0750 (+0.0075)
CXR	ResNet-101	0.1150	0.0340 (-0.081)	0.0448 (-0.0702)	0.1024 (-0.0126)
MRI	ResNet-50	0.3419	0.3675 (+0.0256)	**0.1259 (-0.2416)**	0.2874 (-0.0801)
MRI	ResNet-101	**0.2665**	0.3675 (+0.101)	0.3487 (+0.0822)	0.3770 (+0.1105)
CT	ResNet-50	0.2866	0.2361 (-0.0505)	**0.1909 (-0.0957)**	0.3367 (+0.0501)
CT	ResNet-101	0.3237	0.1975 (-0.1262)	0.2382 (-0.0855)	0.2464 (-0.0773)

(e) ECE All Datasets

Table 1: Performance-based metrics (a-b) and ECE calibration metrics (e)

Conclusion

- Our study has shown the potential of modern augmentations to increase performance & calibration of medical image analysis algorithms across a variety of imaging modalities
- By increasing the reliability of uncertainty measures through augmentations, we can:
  - **Prevent clinical misinterpretations**
  - **Increase clinical confidence in medical AI**
  - **Increase the accuracy of medical AI**

Augmentation	↑Calibration	↓Calibration
MixUp	6	2
CutMix	**7**	**1**
CutOut	4	4

Table 2: Numerical summary of calibration effects