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.

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.

Medical Images

![Calibration Study Design](image)

Figure 1: Design of our augmentation evaluation study

<table>
<thead>
<tr>
<th>Medical Images</th>
<th>Calibration Study Design</th>
</tr>
</thead>
<tbody>
<tr>
<td>Derm</td>
<td>MixUp</td>
</tr>
<tr>
<td>CXR</td>
<td>MixUp</td>
</tr>
<tr>
<td>MRI</td>
<td>MixUp</td>
</tr>
<tr>
<td>CT</td>
<td>MixUp</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Augmentation</th>
<th>Accuracy</th>
<th>AUROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>0.927</td>
<td>0.944</td>
</tr>
<tr>
<td>MixUp</td>
<td>0.644</td>
<td>0.980</td>
</tr>
<tr>
<td>CutMix</td>
<td>0.941</td>
<td>0.977</td>
</tr>
<tr>
<td>CutOut</td>
<td>0.917</td>
<td>0.941</td>
</tr>
</tbody>
</table>

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

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.

Table 2: Numerical summary of calibration effects

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.