Reproducibility Project Proposal. CS598 DL4H in Spring 2023

Quan Nguyen
quanhn2@illinois.edu

Group ID: 11
Paper ID: 27

1 Original paper goal

The authors discuss the challenges of implementing effective medical Artificial Intelligence (AI) tools in clinical practice. Despite massive investment into medical AI development with the hope of lowering healthcare costs and improving efficiency, translation into clinical deployment has proved to be extremely challenging. The difficulties range from regulatory considerations to the lack of transparency in machine learning models.

To address these challenges, the authors proposes the use of conformal prediction methods for medical imaging applications. Conformal predictions can better correspond to clinical decision-making intuition and provide meaningful statistical guarantees. The authors argue that conformal prediction sets naturally convey a measure of uncertainty by the number of items contained in their set, which is similar to how doctors routinely express uncertainty in the form of comparative sets to arrive at a diagnosis, known as a ”differential diagnosis.” The use of conformal predictions methods can also promote uncertainty quantification to facilitate greater trust in medical black-box models and to detect bias in protected patient demographics.

2 Approach taken by the original authors

The paper showcases how conformal predictors offer a more clinically intuitive representation of model uncertainty by empirically evaluating them on a dermatology dataset for skin lesion classification using Fitzpatrick skin type as a group attribute. The study characterizes two conformal use-cases, namely ”rule-in” and ”rule-out,” and shows how conformal predictors can be adapted to yield equalized coverage at a subgroup level.

Besides, the authors also compare conformal uncertainty against epistemic uncertainty to demonstrate how group-calibrated conformal predictors better represent relevant subgroup differences such as disease prevalence of malignant skin conditions.The following methods are compared: a non-conformal baseline (Naive), adaptive prediction sets (APS) (Romano, Sesia, and Candes 2020), regularized adaptive prediction sets (RAPS) (Angelopoulos et al. 2021), group adaptive prediction sets (GAPS), and group regularized adaptive prediction sets (GRAPS).

3 Hypothesis to verify

3.1 Performance of the methods used are similar

The study concludes that all methods of prediction-sets have the same accuracy at $\alpha = 0.1$.

3.2 Group conformal methods slightly better performance

GAPS and GRAPS perform better at identifying malignant cases in general and worse than other methods for the lightest skin tones.

3.3 Conformal methods versus group variants

The authors observe that GAPS and GRAPS have higher set size disparity compared to APS and RAPS.

3.4 Relationship between set size and epistemic uncertainty

The study concludes that there is a strong correlation between set size and epistemic uncertainty.

3.5 Study conclusion

The paper finds that group conformal methods are more effective in describing the uncertainty of subgroups from diverse data distributions than regular conformal methods.

4 Study ablation plan

The study uses a pre-trained ResNet-18 model as the deep learning classifier. The first ablation is to use an un-trained ResNet-18 model. The goal is to observe how performance of the model changes when its parameters are initialized from scratch.

The reproduced study can also use slightly different hyper-parameters when training. Some examples such as using learning rate 0.001 or 0.00001 instead of 0.0001, or using a larger batch size 32 instead of 16. These parameters might have significant or zero impact on the results.

Other ablation option is tweaking some configuration of the ResNet-18 architecture when creating the model without using pretrained-weights.

5 Obtaining the dataset

The dataset in the original paper is Fitzpatrick17k, which is comprised of 16,577 images from two dermatology sources. It has 114 skin condition labels arranged in a hierarchy, which are grouped into three dermatological categories: non-cancerous lesions (12,080), harmless lesions (2,234), and cancerous lesions (2,253). Also, each image has a Fitzpatrick skin type label, which rates the amount of melanin pigment in the skin using an ordinal 6-point scale.

The dataset is publicly available and can be accessed using metadata files and hyperlinks from this Github repository: https://github.com/mattgroh/fitzpatrick17k.

The plan is to write a Python script to automate the process of retrieving and download all the images from the public hosts.

6 Computational feasibility

In the study, all models were trained on a Nvidia A100 GPU machine. In the reproduction study, Google Colab Pro with T100 or A100 Nvidia GPU is an viable option. Other possible options are AWS EC2 Accelerated Computing instances such as p2.xlarge (4 cpus, 61Gb memory, 1 gpu) or similar. Either of the mentioned above options should be sufficient for this reproduction study.

7 Source code reusability

The reproduction study will attempt to not use the original study's source code and build it from scratch. However, implementation approaches will be heavily referenced by the source code. The link to the source code is: https://github.com/clu5/AAAI-22

References

Charles Lu, Andréanne Lemay, Ken Chang, Katharina Höbel, and Jayashree Kalpathy-Cramer. 2022. Fair conformal predictors for applications in medical imaging. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11):12008–12016.

(Lu et al., 2022)