Towards Trustworthy Deep Learning: Applying Mondrian Conformal Prediction to Intracranial Hemorrhage Detection

Cooper Gamble, Undergraduate Research Intern, Mayo Clinic; Shahriar Faghani, MD

Introduction
Deep learning (DL) has earned its place in medical imaging by performing classification, segmentation, generation, and detection tasks with impressive efficiency and accuracy. By default, out-of-the-box DL tools give statistical estimates about the certainty of their predictions, but these figures cannot be interpreted as probabilities or confidence scores because they are uncalibrated. Our approach replaces these estimates with statistical guarantees for predictions by calibrating DL outputs. We apply Mondrian Conformal Prediction (MCP) to Intracranial hemorrhage (ICH) detection, one of many severe conditions for which DL models have been trained as clinical support systems.

Hypothesis
Mondrian Conformal Prediction supports trustworthy deep learning by providing statistical guarantees for model predictions.

Methods
We selected the CQ500 dataset, which is composed of 193,317 scans from 491 patients. It contains challenging cases (n=52539) in which one or more of the readers disagree about a patient’s diagnosis, and we designated the remaining patients’ slices as definite cases. We trained a YOLOv8 model on both positive-only and balanced subsets of our dataset to localize and identify the ICH types. Finally, we performed MCP by sorting the confidence scores of our calibration set. We reported MCP’s accuracy in flagging challenging cases.

Results
The mAP of 0.547 during validation and 0.411 during testing on definite samples. Our model was best at localizing and identifying IVH, IPH, and SDH instances with respective mAPs of 0.995, 0.694, 0.443 during validation. At a p-value threshold of 0.2, MCP identified 100% of challenging cases by generating a prediction set which contained both the presence and absence of at least one hemorrhage type.

Conclusion
We built an ICH detection model whose performance rivals state-of-the-art models, but which can also output prediction sets with statistical guarantees. Furthermore, applying MCP to improve trustworthiness demonstrated perfect accuracy in flagging challenging cases. Continued external validation is necessary for clinical adoption, but this study is a promising first step in that direction.

Statement of Impact
Trust is a critical component of practical deep learning tools. We offer a deployable, statistically rigorous, and task-and model-agnostic approach to increase trustworthiness in DL models by calibrating and filtering their predictions.
Figure 1: Prediction and ground truth for a definite case. Checked boxes indicate labels that were included in the model's prediction set.
Figure 1: Prediction and ground truth for a challenging case. Checked boxes indicate labels that were included in the model's prediction set.
Figure 3: Interface for Mondrian Conformal Prediction with YOLOv8 model (available at https://huggingface.co/spaces/cgamble/gradio_mcp).

Demo link: https://www.abstractscorecard.com/uploads/Tasks/upload/20687/HAPYGCNX-1583345-3-ANY(3).gif

**Keywords**
Conformal Prediction; Brain Hemorrhage; Uncertainty Quantification; Object Detection