Mediations on Metrics

Odysseas Vavourakis

December 2023

Setup

We are building a quality control tool for MRI data. The purpose of the tool is, ultimately, to reduce the need for manual screening. There are two ways to achieve this:

- (correctly) label a fraction of the input data as **definitely clean**
- (correctly) label a fraction of the input data as **definitely dirty**

We expect to work with mostly clean data, so we anticipate greater human-effort reductions from the former, but we can do both simultaneously.

We use a model architecture that produces one stochastic point estimate of the "artefact probability" p of each input image per inference run (using e.g. inference-time dropout). After N inference runs (a hyperparameter) on an image, there are several sensible ways to collate the results, e.g.:

(A) compute sample μ (or max) and σ of the probability estimates $\{p\}$

(B) turn the estimates into class predictions given a fixed threshold π (a hyperparameter), then compute sample μ (or max) and σ of the predicted classes

(C) do kernel density estimation from the point estimates (need to choose associated hyperparameters), then work from there

We consider the first two options in what follows.

Given a *fixed* model that has given predictions $\{p\}$ on a *fixed* validation set, and also given a *fixed* method (e.g. (A) or (B) above) to compute μ and σ from these predictions, we choose thresholds (η, θ, τ) (and potentially π for option (B)).

- η is the maximally allowed prediction uncertainty σ above which images are flagged for manual review
- θ is the decision threshold on μ below which images automatically pass QC;
- τ is the decision threshold on μ above which images automatically fail QC ;
- all other images are flagged for manual review.

In other words: images with

- $\sigma > \eta$ are flagged for manual review (FFMRed) due to model uncertainty;
- $\sigma \leq \eta$ and $\theta \leq \mu \leq \tau \leq 100\%$ are FFMRed due to class uncertainty;
- $\sigma \leq \eta$ and $\theta < \tau < \mu \leq 100\%$ are rejected as definite artefacts;
- $\sigma \leq \eta$ and $\mu < \theta < \tau < \mu \leq 100\%$ are accepted as definitely clean.

See also Fig. 1.

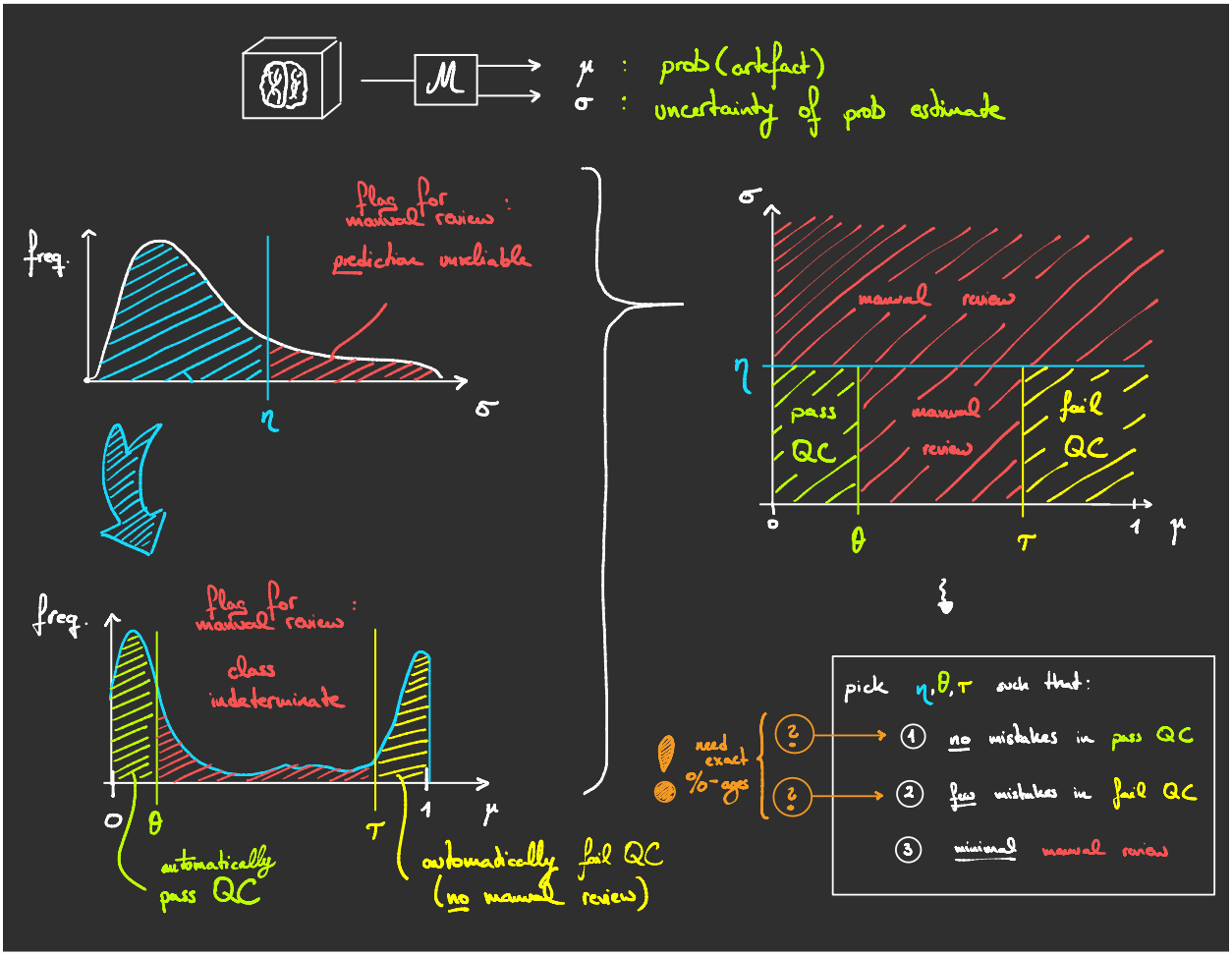

Figure 1: Schematic of Evaluation Setup

Objectives

In order of priority, we must ensure that

1. images that pass QC are definitely clean
    - i.e. Negative Predictive Value 100%, Purity 100%, UDM 100% (or something close?)
2. images that are rejected are definitely artefacts
    - i.e. Purity 0%, precision on this subset 100% (or something close?)
3. manual screening is significantly reduced
    - i.e. $DFFMR(\eta, \tau, \theta) = FFMR/all$ is small

Picking an Operating Point (η, θ, τ)

Given fixed a model, predictions on a test-set, collation method etc.,

1. calculate the above metrics for a (η, θ, τ) (and possibly π) grid
2. successively filter by Objective 1, 2 and 3
3. if no grid point satisfies all objectives, relax Objectives in reverse order or improve the model

We can define the purity of the empty set as 0. This allows us to only set $\tau \neq 100\%$ if there is truly a benefit to ternary (rather than binary) classification, i.e. if this setup allows us to exclude more images from manual review without sacrificing certainty on the rejected set.

Things to Consider

- it might be smart if η were a function of θ (and/or τ) and μ
    - i.e. an uncertain prediction μ near the decision threshold θ (or τ) is worse than a prediction of equal uncertainty further from the threshold, where variation within the uncertainty range will not change the class assignment
    - e.g. we could consider a (bi-)parabolic parametrisation of this function $\eta_{\theta,\tau}(\mu)$ with minima at θ and τ, though the exact functional form is up for debate