AJNR

Radiomics Approach Fails to Outperform Null Classifier on Test Data

J.B. Colby

AJNR Am J Neuroradiol 2017, 38 (11) E92-E93
doi: https://doi.org/10.3174/ajnr.A5326
http://www.ajnr.org/content/38/11/E92

Radiomics Approach Fails to Outperform Null Classifier on Test Data

It is with great pleasure that I read the recent article, the accompanying commentary, wide popular press, and lively ongoing discussion in the community regarding "Computer-Extracted Texture Features to Distinguish Cerebral Radionecrosis from Recurrent Brain Tumors on Multiparametric MRI: A Feasibility Study" by Tiwari et al.[1]

With increasingly ubiquitous, cheap, computing infrastructure and the commoditization of high-quality machine-learning algorithms, multivoxel and multimodal pattern classification techniques, so-called "radiomics," are increasingly being used to incorporate subtle but useful imaging features into our routine clinical decision-making as radiologists. To this end, I commend the authors on their well-thought-out design and implementation of a machine-learning classifier for differentiating tumor recurrence versus radiation necrosis in treated primary and metastatic human brain tumors.

The authors used a state-of-the-art but straightforward method incorporating the following: 1) image texture-based feature extraction, 2) feature selection/reduction via minimum redundancy maximum relevance, 3) generalization performance estimation of a support vector machine classifier, and most important, 4) an external layer of cross-validation to ensure that the feature selection and performance estimation steps were unbiased (ie, not overfit).

The authors widely acknowledged that this is, indeed, an early feasibility study, using limited retrospective data at hand, and this point has already been further explored by other commenters.[2] Additionally, however, there was no discussion of base rate effects or inclusion of a null classifier. A discussion here will hopefully be useful in both understanding the authors' specific results and generalizing to the future because we aim to target the specific clinical scenarios where these advanced techniques may have their greatest clinical utility.

Diagnostic testing can be framed in the Bayesian sense of a pretest (prior) probability, which is updated by some new evidence, to yield a posttest (posterior) probability.[3] The pretest probability is often informed by some general knowledge about the background prevalence (ie, base rate) in the community. The new evidence typically takes the form of a test result for the individual. Consider the fringe cases: On the one hand, we can imagine a clinical scenario where the base rate is 50%. Under this regimen, similar to the authors' training data, incorporation of individual data—even of marginal reliability—will nudge us in favor of 1 group and improve clinical diagnostic accuracy. On the other hand, as the base rate approaches 0% or 100%, even the best diagnostic tests will be useless in practice. For example, consider the challenge of identifying an uncommon disease in a hypothetic healthy population of 1000 individuals. If we examined the whole population, given a supposed baseline prevalence of 0.001 (1 in a 1000), even a terrific "rule out" screening diagnostic test with 100% sensitivity and 95% specificity will result in 1 true-positive test result and 49.95 false-positive test results, for an overall positive predictive value of only approximately 2%.

The crux of the issue then lies in the middle gray zone: As the base rate slides more in favor of 1 group, the bar rises for any additional candidate predictors/features to be "worth it" in terms of the marginal discriminating information they provide with respect to their inherent variability and accompanying measurement error. In the present feasibility study, tumor recurrence was only slightly more prevalent than radiation necrosis among the primary brain tumor training data (12 of 22 cases, or 55%). Therefore, a null classifier incorporating only this information would perform with 55% accuracy on average (the null information rate) and would be beaten handily by the authors' imaging-based classifier, which achieved 75% estimated generalization accuracy via cross-validation-based resampling on the training data. This 75% number is the benchmark we would like to compare against human performance; however, such an analysis was not performed in the present study.

In the holdout test sample, however, the recurrence group was much more enriched. Therefore, while it may seem impressive that the imaging-based classifier attained 91% accuracy (10/11 cases) and this was the main headline widely publicized in the popular press, we would, in fact, have attained the exact same diagnostic accuracy by ignoring all the machine-learning algorithms, relying solely on our general knowledge of the base rate that tumor recurrence is more common and assigning every holdout test case to the "recurrence" class label without looking at a single image.

http://dx.doi.org/10.3174/ajnr.A5326

This leads to several important discussion points: Because there are so many subtle ways for classification experiments to be methodologically invalidated, there is a strong intuitive desire to see the methods tested on a truly independent test set, held out from the get-go, as was done here. This approach does have the desired effect of making us feel more comfortable with the methods; however, it also has negative effects. Not only does it make less data available for training, thereby decreasing the quality of the classifier, but, as we see here, the test set itself may be biased due to small sample size or other effects. This then provides yet another argument in favor of data-sharing and "reproducible research" in neuroimaging,[4] whereby the community could easily validate the authors' methods, confirm their cross-validated results, and obviate a separate holdout test.

It is also worth revisiting the uncomfortable fact that humans are useful but flawed statistical machines (and hence clinical decision-makers), subject to a variety of cognitive biases that have been explored in the literature on the psychology of decision-making during the past half-century.[5] We underestimate the value of base rate information compared with individual information, overestimate the generalizability of our talents, overestimate the confidence/precision of our estimates, and are able to be systematically nudged by a variety of factors, including the arbitrary sequence in which cases are presented. In particular, as the validity of a task decreases (ie, the signal gets smaller, subtler, or more complex) or accompanying uncertainty increases, the consistency of our approach to intuitive reasoning suffers, and the net effects of these underlying biases may dominate.[6] Although it was not investigated here, we may speculate that some (or all) of these effects may help explain why the "experts" performed even more poorly than would be expected on the test data. On the bright side, ample data do suggest that the performance of expert intuitive reasoning under this regimen can be successfully augmented by the introduction of even simple algorithms,[7] as evidenced in our field by the success of the Breast Imaging Reporting and Data System, the Liver Imaging Reporting and Data System, and so forth.

In summary, while widely publicized, the presented radiomics approach fails to outperform a null classifier on the given test set. Conversely, we are unable to compare the classifier cross-validated performance estimates on the training set with human performance because this analysis was not performed. If one looks forward, this interesting article describes a state-of-the-art radiomics classifier, though it highlights the importance of base rate effects and other cognitive bias when evaluating the usefulness of such a classifier and again argues in favor of both enhanced data-sharing in neuroimaging and enhanced incorporation of our expert intuitive reasoning into more structured frameworks for clinical decision-making.

REFERENCES

1. Tiwari P, Prasanna P, Wolansky L, et al. **Computer-extracted texture features to distinguish cerebral radionecrosis from recurrent brain tumors on multiparametric MRI: a feasibility study.** *AJNR Am J Neuroradiol* 2016;37:2231–36 CrossRef Medline
2. Schweitzer AD, Chiang GC, Ivanidze J, et al. **Regarding "Computer-Extracted Texture Features to Distinguish Cerebral Radionecrosis from Recurrent Brain Tumors on Multiparametric MRI: A Feasibility Study."** *AJNR Am J Neuroradiol* 2017;38:E18–19 CrossRef Medline
3. Elstein AS, Schwartz A, Schwarz A. **Clinical problem solving and diagnostic decision making: selective review of the cognitive literature.** *BMJ* 2002:324:729–32 Medline
4. Poldrack RA, Baker CI, Durnez J, et al. **Scanning the horizon: towards transparent and reproducible neuroimaging research.** *Nat Rev Neurosci* 2017;18:115–26 CrossRef Medline
5. Tversky A, Kahneman D. **Judgment under uncertainty: heuristics and biases.** *Science* 1974:185:1124–31 CrossRef Medline
6. Kahneman D, Klein G. **Conditions for intuitive expertise: a failure to disagree.** *Am Psychol* 2009;64:515–26 CrossRef Medline
7. Grove WM, Zald DH, Lebow BS, et al. **Clinical versus mechanical prediction: a meta-analysis.** *Psychol Assess* 2000;12:19–30 Medline

J.B. Colby
Department of Radiology and Biomedical Imaging
University of California, San Francisco
San Francisco, California