Human Versus Machine: How Do We Know Who Is Winning? ROC Analysis for Comparing Human and Machine Performance under Varying Cost-Prevalence Assumptions

Michael Merry[1] Patricia Jean Riddle[1] Jim Warren[1]

[1] School of Computer Science, University of Auckland, Auckland, New Zealand

Methods Inf Med 2022;61:e45–e49.

Address for correspondence Michael Merry, BA BMus BSc(Hons), School of Computer Science, University of Auckland, Private Bag 92019, Auckland 1142, New Zealand (e-mail: m.merry@auckland.ac.nz).

Abstract

Background Receiver operating characteristic (ROC) analysis is commonly used for comparing models and humans; however, the exact analytical techniques vary and some are flawed.

Objectives The aim of the study is to identify common flaws in ROC analysis for human versus model performance, and address them.

Methods We review current use and identify common errors. We also review the ROC analysis literature for more appropriate techniques.

Results We identify concerns in three techniques: (1) using mean human sensitivity and specificity; (2) assuming humans can be approximated by ROCs; and (3) matching sensitivity and specificity. We identify a technique from Provost et al using dominance tables and cost-prevalence gradients that can be adapted to address these concerns.

Conclusion Dominance tables and cost-prevalence gradients provide far greater detail when comparing performances of models and humans, and address common failings in other approaches. This should be the standard method for such analyses moving forward.

Keywords

- ROC analysis
- machine learning
- performance metrics

Introduction

Many researchers use clinician performance as a benchmark for machine learning with a range of methods to compare human and machine performances.[1–7] The most common is receiver operating characteristic (ROC) analysis with the area under the curve (AUC) being the primary performance metric. As this is used commonly for comparing different models, this has come into use for comparing human and machine performance. ROC and AUC analyses provide an effective generalized summary of performance that is cost and prevalence invariant. When applied to practical applications, however, cost and prevalence assumptions are necessary to decide how to optimise the sensitivity–specificity tradeoff. In reviewing existing literature, we have identified several ways that the ROC analysis is used in the human versus machine context that can lead to erroneous results.

received
August 4, 2021
accepted after revision
October 19, 2021
published online
December 31, 2021

DOI https://doi.org/10.1055/s-0041-1740565.
ISSN 0026-1270.

Georg Thieme Verlag KG, Rüdigerstraße 14, 70469 Stuttgart, Germany

Objectives

The aim of the study is to assess whether current practices for comparing human and machine performance are appropriate, and if not, identify a more rigorous approach.

Methods

We searched the literature for papers that compare human and machine performance using keyword searches on combinations of "human," "clinician," "AI," "artificial intelligence," "machine," "model," "ROC," and "performance" as well as manually reviewing citations from notable articles especially on the topic.[8–10] We reviewed the analytical approach of the relevant papers, and identified several methods of concern related to ROC analysis. Evaluation of the methods was done from the frame of reference of being able to use the results to guide real-world decisions. We then reviewed the original papers describing the theory behind ROC analysis and identified a variant that can appropriately address the concerns raised.

Results

Concerns with Existing Approaches

Several highly cited papers that use ROC as a basis for machine–human comparison stand out[5–7] with further exemplary papers being identified after further review.[1,4] Other works looked at factors other than performance, such as speed or ease.[2,3] We identified concerns with three approaches comparing performance via ROC analysis:

1. Using mean human sensitivity and specificity.
2. Assuming humans can be approximated by ROCs.
3. Sensitivity and specificity matching.

As the specifics of the individual papers differ, we will present the core elements of the three methods, and discuss the concerns with each.

Mean Human Sensitivity and Specificity

Using the mean human sensitivity and specificity (i.e., the [sensitivity, specificity] pair of $(mean(sens_h), mean(spec_h))$ *for* $h \in$ *Human performances*) in comparison with the ROC of a model will always underestimate the human performance and put it at a disadvantage. This risks over-calling claims of model superiority. The geometric mean of a series of points will always be inside the convex hull (the smallest convex set containing all the points[11]). For example, we could have a model that for every operating point, we could find a human that was fractionally better. If we examine the ROC plot, then we would find that the ROC of the model is always dominated by human performance. However, if we take the mean sensitivity and specificity of the model, this would be within the ROC of the model (►Fig. 1) and we would claim the model is superior to human performance, which conflicts with the result of humans dominating at all operating points. This is a comparison between different metrics (the ROC of the model and the geometric mean of performances) and results in over-optimistic claims of model performance.

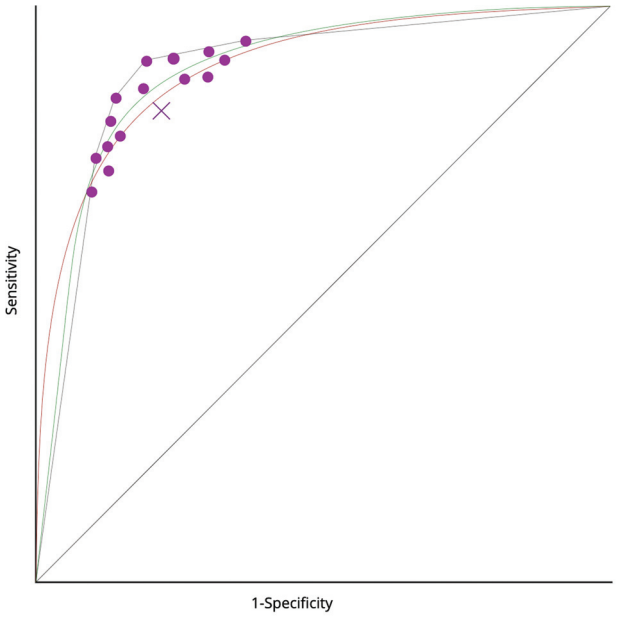

Fig. 1 A set of human performances (*purple dots*) with mean sensitivity and specificity (*purple cross*), the convex hull of optimal human performance (*black*), the estimated ROC of human performance (*green*) and a model's ROC (*red*). All ROCs/ROC estimates dominate the mean performance. The model ROC dominates in high-specificity conditions, and has a higher AUC, but is dominated in the range where humans primarily operate. AUC, area under the curve; ROC, receiver operating characteristic.

Approximating Humans with ROC Curves

A second approach is to approximate humans with ROC curves by constructing the convex hull of human performance, plus the points at (0,0) and (1,1) and considering that curve as the human ROC. Although this does allow one to compute an AUC for humans (and thus compare the same AUC metric between humans and models), there are still several concerns. The primary concern is in comparing a model for which you can choose an operating characteristic with humans who have a single operating characteristic guided by their practical experience in context. This would mean that human performances do not cover the full range of sensitivities and specificities, thus being under sampled especially in the extreme ranges, and rather would be concentrated on a range that is appropriate for the task that they take on in the real world. This can only result in underestimating the ROC of human performance. Due to the construction method of a convex hull, undersampling a region will result in a lower estimation of the ROC.

As medical professionals are generally optimizing for patient outcomes, considering prevalence and relative cost of errors, it is perhaps a reasonable assumption that if there are few human professionals operating at a given sensitivity or specificity region (e.g., high sensitivity), then the operating region may not be clinically valuable. As such, it may not be such an important consideration whether humans or models are better in that region. However, in this form of analysis, human performance will again be underestimated, and the real-world considerations will not be appropriately

considered. Although it is unlikely to miscall results in clear cases, it is not ideal and leaves room for improvement.

Sensitivity and Specificity Matching

The third approach, which is sensitivity/specificity matched performance, takes this into account. Typically, a high-sensitivity and a high-specificity operating point (often chosen to be 0.90 or similar) of a model will be compared with equivalent human performance. This compares operating points with other operating points, which addresses many of the concerns of the above approaches. Of the approaches discussed, this is the most informative and appropriate. However, it still has the issue of requiring the researchers to choose the sensitivity/specificity points to compare, and typically only covers one or two operating points. Ideally, such subjective choices in the analysis by the researchers would not be necessary. Rather, a parameter-free methodology would be preferable, especially one that is able to describe a wider range of performances.

ROC Analysis Variant

An analytical method for ROC was presented by Provost et al in 1998[12] (which forms part of the foundational work of these authors prior to the seminal *An introduction to ROC Analysis*[13]). This method is able to address the problems of comparing ROCs of models with operating points of humans and provide relevant information for assessing real-world performance. They suggest and demonstrate a variation of ROC analysis on the basis of the convex hull of all models. They consider which model dominates in different areas of the full range of cost and prevalence ratios. This results in a table of "slope ranges" where for each prevalence-adjusted cost-gradient slope range (prevalence-cost gradients [CG]), a single model is identified as dominating. In this way, one can identify which model is superior under specific conditions, providing the necessary information for practical applications. This work did not consider the addition of human performances, but as the analysis is on the basis of a finite set of operating points of models, rather than a smoothed, continuous ROC, it is easily generalizable to humans by adding their individual operating points.

One limitation of this approach is that it does not allow the presentation of prevalence-variant metrics such as positive predictive value (PPV), as the prevalence-CGs take into consideration both the prevalence and cost (by definition). To consider such metrics, one can make assumptions about the prevalence in the relevant population and the metrics provided give sufficient information to be able to compute the rest of the relevant information. For example, one might consider a model for use in a diagnostic screening program, and so assume the population-level prevalence of the condition. On that basis, one can compute the 2*2 table of true and false positives and negatives from which all other metrics can be computed.

The cost-gradient ranges are found by identifying which operating point minimizes the equation:

$$CPG = \frac{\pi}{1-\pi} * \frac{C_{fp}}{C_{fn}}$$

The cost prevalence gradient (CPG) is the ratio of positive to negative cases multiplied by the ratio of costs for false positive and false negative results. If one makes an assumption about the prevalence, then one can derive the CG from the CPG as the following:

$$CG = CPG * \frac{1-\pi}{\pi}$$

Similarly, one can adjust for assumed cost ratios and examine the ranges according to prevalence.

Worked Example

Consider the comparison of two models for detecting diabetic retinopathy (DR) against the performance of 15 specialists. The models are trained to detect DR from retinal fundus images taken using a custom attachment for smartphone cameras. The performance of this model is compared against a group of 15 specialists. Assume all testing and data collection methodologies were sound. Consider the application in a two-step diagnosis with population screening followed by an in-clinic assessment for at-risk patients.

The ROC of the model and human results are shown in ►Fig. 1, with the dominance table in ►Table 1. A 17.7% in-population prevalence was assumed to create the cost-gradient table in ►Table 1.

The false-positive and false-negative costs for the screening program are assumed to be the same, resulting in a cost-gradient of 1. Examining ►Table 1, Model A dominates under these conditions and would be the best choice. For the screening program, one might consider the (0.93, 0.60) sensitivity/specificity operating point which would yield a PPV of 0.41 and NPV of 0.97. For those patients requiring in-clinic follow-up, assume a 41% post-screening prevalence and re-evaluate which is the best model (►Table 1). Assuming a 5:1 cost ratio of false positive to false negative, we can see that it falls in the range where Specialist C is dominating. Human performance is better under these prevalence and cost conditions, and so specialist review is still preferred after screening

Discussion

This ROC analysis variant using dominance tables addresses the majority of the concerns raised. It compares operating points between models and humans in a way that provides all the relevant information necessary to evaluate the appropriate choice in differing cost-prevalence conditions. Further, it does this using only small extensions to existing techniques.

It is possible to consider such dominance tables applicable for any ROC analysis, even for model to model comparisons without human performances. First, it applies more generally to any comparison of a group of sensitivity–specificity pairs against models. Second, it can apply when comparing AUCs of models, where one might have partial dominance

Table 1 General model performances and two worked prevalence examples

General model performance			17.7% Prevalence assumption			41.0% Prevalence assumption		
Cost-prevalence gradient range	Dominator	(Sensitivity, specificity) range	Cost-gradient range	PPV range	NPV range	Cost-gradient range	PPV range	NPV range
0.00–0.31	Model A	(1.00, 0.00)–(0.93, 0.60)	0.00–1.43	0.18–0.41	1.00–0.97	0.00–0.44	0.41–0.70	1.00–0.91
0.31–0.50	Specialist A	(0.93, 0.60)–(0.89, 0.73)	1.43–2.32	0.41–0.54	0.97–0.96	0.44–0.72	0.70–0.79	0.91–0.88
0.50–1.14	Specialist B	(0.89, 0.73)–(0.83, 0.85)	2.32–5.31	0.54–0.67	0.96–0.94	0.72–1.64	0.79–0.87	0.88–0.84
1.14–4.33	Specialist C	(0.83, 0.85)–(0.75, 0.92)	5.31–20.14	0.67–0.73	0.94–0.92	1.64–6.24	0.87–0.90	0.84–0.78
4.33–Inf	Model B	(0.75, 0.92)–(0.00, 1.0)	20.14–Inf	0.73–1.00	0.92–0.82	6.24–Inf	0.90–1.00	0.78–0.59

Note: Regular reporting would only present the first three columns as all others can be calculated on different assumptions on the basis of the first three.

between the two models.[12] The usual guidance is to look at the ROC plot, but the dominance table provides all of the information in a quantified and detailed manner that could be understood from analyzing the plot. This level of detail and transparency around the performance of the models can only be positive for ongoing research.

It would be necessary for additional tooling to be created to support the proposed methodology. Relevant statistical packages for ROC analysis would need to be extended to make this accessible.

This method so far does not take into consideration how to manage variance for ROCs and human performance. This is an important and non-trivial aspect of ROC analysis and is treated in depth in a study by Fawcett.[13] Further, this does not address extensions to multiple-class classifiers where we would argue that the concerns are greater due to the higher complexity and increased risk of pitfalls.

Conclusion

It is likely that some claims of models beating human performance are wrong due to flawed ROC analysis methodologies for comparing human and model performances. A variation on a method presented by Provost et al for comparing different models can be adapted for comparing models and humans (or models and models) using dominance tables. These tables, combined with cost-prevalence gradients allow for accurate claims of model or human superiority under a range of assumptions, allowing for clearer presentation of results and more accurate claims. This should be necessary for any of the papers comparing human and model performance.

Conflict of Interest

None declared.

References

1 Andersson S, Heijl A, Bizios D, Bengtsson B. Comparison of clinicians and an artificial neural network regarding accuracy and certainty in performance of visual field assessment for the diagnosis of glaucoma. Acta Ophthalmol 2013;91(05): 413–417
2 Steiner DF, MacDonald R, Liu Y, et al. Impact of deep learning assistance on the histopathologic review of lymph nodes for metastatic breast cancer. Am J Surg Pathol 2018;42(12): 1636–1646
3 Liu Y, Gadepalli K, Norouzi M, et al. Detecting cancer metastases on gigapixel pathology images. arXiv 2017:1–13
4 Mueller M, Almeida JS, Stanislaus R, Wagner CL. Can machine learning methods predict extubation outcome in premature infants as well as clinicians? J Neonatal Biol 2013;2:1000118
5 Haenssle HA, Fink C, Schneiderbauer R, et al; Reader study level-I and level-II Groups. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. Ann Oncol 2018;29(08):1836–1842
6 Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017; 542(7639):115–118
7 Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic

retinopathy in retinal fundus photographs. JAMA 2016;316 (22):2402–2410
8 Hinton G. Deep learning—a technology with the potential to transform health care. JAMA 2018;320(11):1101–1102
9 Stead WW. Clinical implications and challenges of artificial intelligence and deep learning. JAMA 2018;320(11): 1107–1108
10 Yu KH, Beam AL, Kohane IS. Artificial intelligence in healthcare. Nat Biomed Eng 2018;2(10):719–731
11 de Berg M, van Kreveld M, Overmars M, Schwarzkopf O. Computational geometry. Berlin, Heidelberg: Springer Berlin Heidelberg; 1997:1–17
12 Provost F, Fawcett T, Kohavi R. The Case Against Accuracy Estimation for Comparing Induction Algorithms. Paper presented at: Proceedings of the Fifteenth International Conference on Machine Learning (IMLC-98), Madison, WI, 1998
13 Fawcett T. An introduction to ROC analysis. Pattern Recognit Lett 2006;27:861–874