A Hazard-Aware Metric for Ordinal Multi-Class Classification in Pathology

David Jin[1], **Ariel Kapusta**[2], **Patrick A. Minot**[2], **Niels H. Olson**[3], **Joseph H. Rosenthal**[4], **Jansen N. Seheult**[5], **Michelle Stram**[6]

[1] DoD Chief Digital and Artificial Intelligence Office
[2] MITRE Corporation
[3] Defense Innovation Unit
[4] Henry M. Jackson Foundation for the Advancement of Military Medicine
[5] Mayo Clinic
[6] NYU Grossman School of Medicine
seheult.jansen@mayo.edu

Abstract

Artificial Intelligence (AI) for decision support and diagnosis in pathology could provide immense value to society, improving patient outcomes and alleviating workload demands on pathologists. However, this potential cannot be realized until sufficient methods for testing and evaluation of such AI systems are developed and adopted. We present a novel metric for evaluation of multi-class classification algorithms for pathology, Error Severity Index (ESI), to address the needs of pathologists and pathology lab managers in evaluating AI systems.

Introduction

The domains in which Artificial Intelligence (AI) and Machine Learning (ML) are applied are varied and each has its own specific challenges and idiosyncrasies. Pathology is a field that commonly deals with multi-class classification problems for diagnosis of tissue samples. Such problems include semi-quantitative immunohistochemistry (IHC) analysis, such as PD-L1 stain quantitation (Reisenbichler et al. 2020), tumor risk group stratification, such as prostate Gleason group grading, Nottingham grading or grading of cervical dysplasia, as well as semi-quantitative cell differentials, such as myeloid blast count thresholds in bone marrow aspirates. AI has been applied to this problem in an effort to improve patient outcomes and alleviate workload demands (Litjens et al. 2016; Nagpal et al. 2020; Perincheri et al. 2021; Bulten et al. 2022; Chen et al. 2019). Although there are many methods to evaluate algorithms, there are some limitations in the existing metrics in literature for capturing the severity of hazards associated with certain errors in multi-class classification problems.

The US government's Clinical Laboratory Improvements Act of 1988 (42 CFR § 493.1253) requires laboratories to verify or establish the accuracy of clinical tests or procedures. The accuracy of any unmodified, FDA-cleared or approved test system must be comparable to that established by the manufacturer. For any modified FDA-cleared or approved test system or test system not subject to FDA clearance or approval (including methods developed in-house or laboratory-developed test procedures), the laboratory must establish the accuracy of the test system prior to implementation. While laboratorians are familiar with methods and tools for evaluating accuracy claims for binary classification tasks, there is a need for more informative hazard-aware evaluation metrics for multi-class and ordinal tasks.

We present a novel metric for evaluation of multi-class classification algorithms for pathology, Error Severity Index (ESI), to address the needs of pathologists and pathology lab managers in evaluating AI systems.

Authors ordered alphabetically.

Method and Results

Error Severity Index

Our proposed metric, error severity index (ESI), can be calculated as follows for a multi-class classification problem.

Let C be an n by n confusion matrix with $n > 0$. Let W be an n by n error severity weight matrix with values ranging from 0 to 1. The weights should correspond to the severity of the corresponding error in a confusion matrix. If there are no misclassification errors, ESI $= 0$, otherwise

$$\text{ESI} = 10 \left(\sum_{i,j=1}^{n} C_{ij} W_{ij} \right) \left(\sum_{i,j=1,\ j \neq i}^{n} C_{ij} \right)^{-1}$$

Demonstrative Example

An example is used to demonstrate the partially-informative nature of naive accuracy applied to ordinal tasks and the potential value of ESI.

Evaluating heart biopsies for acute allograft rejection in heart transplant recipients is a complex task that involves microscopic examination of the tissue by a pathologist with histologic grading for acute cellular and humoral/antibody mediated mechanisms of rejection. Grading heart biopsies for acute cellular rejection (ACR) using the International Society for Heart and Lung Transplantation (ISHLT) scoring system is an example of a multi-class ordinal problem encountered in the practice of pathology where the results can alter the clinical management of the patient. Importantly, the ISHLT ACR score is only one variable in determining the clinical treatment of a transplant patient.

The ISHLT ACR grading system is as follows:

1. Grade 0 – No rejection

Weight Severity Matrix

Inferences	Ground Truth G0	G1R	G2R	G3R
G0	0	0.3	0.6	1.0
G1R	0.3	0	0.3	0.6
G2R	0.6	0.3	0	0.3
G3R	1.0	0.6	0.3	0

Vendor 1

Inferences	Ground Truth (%) G0	G1R	G2R	G3R
G0	20	0	0	0
G1R	5	20	5	0
G2R	0	5	20	0
G3R	0	0	0	25

Vendor 2

Inferences	Ground Truth G0	G1R	G2R	G3R
G0	20	2	3	0
G1R	2	20	2	0
G2R	3	3	20	0
G3R	0	0	0	25

Vendor 3

Inferences	Ground Truth G0	G1R	G2R	G3R
G0	20	0	5	0
G1R	0	20	0	0
G2R	0	0	20	0
G3R	5	5	0	25

Figure 1: (Left) Weight Severity Matrix. (Right) Confusion matrices for the three example vendors.

2. Grade 1 R, mild – Interstitial and/or perivascular infiltrate with up to one focus of myocyte damage.
3. Grade 2 R, moderate – Two or more foci of infiltrates with associated myocyte damage.
4. Grade 3 R, severe – Diffuse infiltrate with multifocal myocyte damage, with or without edema, hemorrhage, or vasculitis.

Three algorithms are developed by three vendors to classify the whole slide images of the heart biopsies as: Grade 0 (no rejection), Grade 1 R (mild rejection), Grade 2 R (moderate rejection) and Grade 3 R (severe rejection). A pathologist is presented with the three algorithms and must select between them.

The pathologist is provided a simple weight severity matrix and the confusion matrices shown in Fig. 1. The weight matrix might come from the vendors, it might be created by pathologists/clinicians, or it might come from literature or a standards body.

The severity matrix used in this example is simple, with clear ordinal weighting. A correct classification is assigned a misclassification severity weight of 0, while a misclassification that is one category away from the true class is assigned a misclassification severity weight of 0.3, a misclassification that is two categories away from the true class is assigned a misclassification severity weight of 0.6 and a misclassification that is three categories away from the true class is assigned a misclassification severity weight of 1.0.

While all three algorithms have an accuracy of 85%, they have different distributions of errors and different ESI. While accuracy remains relevant, ESI provides additional information on the severity of the errors made by the algorithm.

The error severity index (ESI) was calculated for each of the algorithms:

Vendor	Accuracy	Error Severity Index
1	85%	3.0
2	85%	4.2
3	85%	7.3

The ESI reveals that the misclassification errors made by the algorithms from the three vendors would likely have dramatically different clinical consequences even though they all share the same accuracy.

Discussion

More complex severity weight matrices Real world applications likely do not have equally spaced severity of errors. Unfortunately, determining the severity of error can be difficult. It requires evaluation of the clinical impact that each error will have on patient management, and this evaluation is very complex. For example, the difference between the clinical severity of misclassifying a case of mild rejection (Grade 1 R) and no rejection (Grade 0) changes depending on the presence or absence of hemodynamic compromise (Ludhwani, Abraham, and Kanmanthareddy 2022). If the patient has no hemodynamic compromise, there is little difference in the patient's management with a misclassification error calling Grade 1 R as Grade 0, but if the patient has hemodynamic compromise, then the misclassification error indicating there is no rejection is a more severe error, because the classification of Grade 1 R would have resulted in the patient receiving pulse dose orally or intravenously, whereas the Grade 0 would not.

How ESI can help Providing the ESI alongside the conventional metrics improves the explainability and transparency of the degree of diagnostic errors that are hidden in an overall accuracy score or other conventional metrics. Combined with other common metrics, ESI can help pathologists and other clinicians understand the performance and limitations of algorithms and enable better decision-making. Algorithm developers should ideally provide their misclassification/confusion matrices along with common metrics so ESI could be evaluated by the clinician on their own. There is no one singular metric which will provide an end-user with all of the information they need to evaluate and select an algorithm among an increasing number of options.

Limitations The ESI assumes that ground truth is established, which is not necessarily the case for all tasks in medicine, especially when the clinical problem does not have a gold standard method for comparison. This metric is not intended to handle challenges such as inter-rater reliability, which also contributes error to establishing ground truth.

Future Work This work holds the promise of allowing for more informed use, with a more appropriate level of trust and awareness of underlying hazards associated with a product. We propose that in order to better align the output from the ESI method with real-world harm, a survey of the pathology community should be conducted. The survey would solicit inputs from the community to allow for the ESI values to better represent collective agreement on the severity levels associated with different forms of ordinal misclassification.

References

Bulten, W.; Kartasalo, K.; Chen, P.; et al. 2022. Artificial intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge. *Nature Medicine*, (28): 154—-163.

Chen, P.-H. C.; Gadepalli, K.; MacDonald, R.; Liu, Y.; Kadowaki, S.; Nagpal, K.; Kohlberger, T.; Dean, J.; Corrado, G. S.; Hipp, J. D.; et al. 2019. An augmented reality microscope with real-time artificial intelligence integration for cancer diagnosis. *Nature medicine*, 25(9): 1453–1457.

Litjens, G.; Sánchez, C. I.; Timofeeva, N.; Hermsen, M.; Nagtegaal, I.; Kovacs, I.; Hulsbergen-Van De Kaa, C.; Bult, P.; Van Ginneken, B.; and Van Der Laak, J. 2016. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Scientific reports*, 6(1): 26286.

Ludhwani, D.; Abraham, J.; and Kanmanthareddy, A. 2022. *Heart Transplantation Rejection*. StatPearls Publishing, Treasure Island (FL).

Nagpal, K.; Foote, D.; Tan, F.; Liu, Y.; Chen, P.-H. C.; Steiner, D. F.; Manoj, N.; Olson, N.; Smith, J. L.; Mohtashamian, A.; et al. 2020. Development and validation of a deep learning algorithm for Gleason grading of prostate cancer from biopsy specimens. *JAMA oncology*, 6(9): 1372–1380.

Perincheri, S.; Levi, A. W.; Celli, R.; Gershkovich, P.; Rimm, D.; Morrow, J. S.; Rothrock, B.; Raciti, P.; Klimstra, D.; and Sinard, J. 2021. An independent assessment of an artificial intelligence system for prostate cancer detection shows strong diagnostic accuracy. *Modern Pathology*, 34(8): 1588–1595.

Reisenbichler, E. S.; Han, G.; Bellizzi, A.; Bossuyt, V.; Brock, J.; Cole, K.; Fadare, O.; Hameed, O.; Hanley, K.; Harrison, B. T.; et al. 2020. Prospective multi-institutional evaluation of pathologist assessment of PD-L1 assays for patient selection in triple negative breast cancer. *Modern pathology*, 33(9): 1746–1752.