Sequestration of imaging studies in MIDRC: using load factor to minimize algorithm performance overestimation and image reuse

Dylan Tang,[1] Heather M. Whitney,[1] Kyle J. Myers,[2] Maryellen L. Giger[1]
[1]Department of Radiology, The University of Chicago, Chicago, IL
[2]Puente Solutions, LLC, Phoenix, AZ

ABSTRACT

Evaluating medical imaging algorithm performance in a test set may lead to a biased result, especially if the number of images is low. In the case of the Medical Imaging Data and Resource Center (midrc.org) sequestered imaging data commons, developers may seek the evaluation of subsequent iterations of an algorithm using additional test subsets drawn from the sequestered commons, allowing for repeat testing but also possibly resulting in learning the sequestered commons when test samples overlap. We developed a method to measure image reuse in test subsets and to evaluate the impact of degree of image reuse on over- or under-estimation of performance by using the load factor, a metric from hash-table methodology that can be used to summarize the average test subset pairings per image. We established a relationship between the standard error of the area under the receiver operating curve (AUC) and load factor, and compared the relationship to interquartile range of AUC for the case of an image-derived predictor for COVID-19 severity on chest radiographs. As expected, AUC variation was inversely related to load factor while image usage increased with load factor, with similar performances between both predicted and actual AUC variation and load factor. Notably, low AUC variation was observed in load factors well above 1, the load factor typically described in the hash-table literature as optimal. These results translate the use of load factor for characterization of stand-alone test sets, supporting future work for operationalizing the use of sequestered test subsets for algorithm evaluation.

Keywords: MIDRC, image reuse, load factor, AUC, algorithm evaluation

1. INTRODUCTION

Medical imaging data commons can facilitate independent testing of algorithms by reserving images in a completely sequestered set, as is conducted by the Medical Imaging and Data Resource Center (MIDRC).[1] When using an independent test set to measure algorithm performance, a common approach is to evaluate the algorithm on some number of test subsets (the selection of N images by some process, such as sampling with replacement), then calculating an average performance and the respective confidence interval. However, it is possible that the average may overestimate or underestimate the true performance. One approach to decrease bias in estimation of performance and increase precision is to select a greater quantity of images per subset, though this requires increased reuse of images especially in small datasets. Image reuse has been linked with as much as a fourteen-fold increase in false-positive rates, as it potentially leads to overfitting to the sequestered dataset.[2,3]

The purpose of our study was to develop a method to measure medical imaging data use in test sets and to use it to (1) measure image reuse in test subsets and (2) evaluate the impact of degree of image reuse on over- or under-estimation of performance. We used hash table principles[4] and load factor (a metric derived from hash table implementation that summarizes the average test subset pairings per image) as a means to evaluate data selection within these two aims. From this, we developed an analytical expression that predicts the variation of the area under the receiver operating characteristic curve (AUC) as a function of load factor, so that extremes of over- and under-estimation can be predicted. Via the use case of an imaging-derived predictor for COVID-19 severity, we investigated the actual variation in AUC as a function of the prevalence of minority class (the class that contained the fewest number of images) for a range of load factors and identified how the load factor can be used to minimize over- or under-estimation of performance while minimizing image reuse.

Medical Imaging 2024: Image Perception, Observer Performance, and Technology Assessment,
edited by Claudia R. Mello-Thoms, Yan Chen, Proc. of SPIE Vol. 12929, 129290G
© 2024 SPIE · 1605-7422 · doi: 10.1117/12.3008264

2. METHODS

2.1 Description of hash table and load factor

The hash table implementation consists of images (tracked by their unique image ID) paired to multiple test subsets. In hash table terminology, the images are the keys (i.e., the unique identifiers used to track the images) of the hash table, while the test subsets are the values. Each test subset is created by randomly selecting N images from the dataset. Sampling can be done with or without replacement. For each image ID that the test subset selects, a test subset ID value is appended to the image ID key using a linked list data structure. Figure 1 shows the conceptual overview of hash tables.

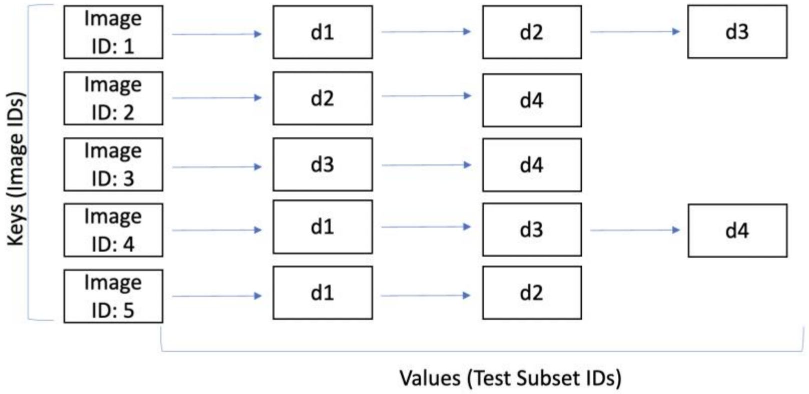

Figure 1: Conceptual overview of a hash table for the example of mapping five images (the image IDs) to four test subsets (labeled d1 through d4) of size three (the test subset IDs).

2.2 Analytical relationship between performance variation and load factor

The load factor is a ratio of values (i.e., test subsets) to keys (i.e., image IDs) and is a metric that describes the number of images drawn relative to the image population.[5] Given the number of images per test subset (sample size N), number of test subsets (number of samples β), and number of images in the dataset (ϕ), load factor (α) can be defined as: $\alpha = \frac{N \cdot \beta}{\phi}$ (Eq. 1). Increases to sample size and number of samples increase the number of values in the hash table, thus leading to a greater load factor. Conversely, an inverse relationship exists between load factor and number of images in the dataset, as an increase in the latter also increases the number of keys in the hash table.

To investigate the relationship between load factor and potential for over/underestimation of performance, the standard error in AUC was evaluated from specified load factors. This was done by analytically determining the standard error in AUC as a function of both AUC and load factor as an adaptation to the expression by Hanley et al.,[6] such that

$$SE(\theta, \alpha) = \sqrt{\frac{\hat{\theta}\left(1 - \hat{\theta}\right) + \left(\frac{\alpha \cdot \phi}{\beta} \cdot i_A\right)\left(\frac{\hat{\theta}}{2 - \hat{\theta}} - \hat{\theta}^2\right) + \left(\left(\frac{\alpha \cdot \phi}{\beta} - \frac{\alpha \cdot \phi}{\beta} - i_A\right) - 1\right)\left(\frac{2\hat{\theta}^2}{1 + \hat{\theta}} - \hat{\theta}^2\right)}{\left(\frac{\alpha \cdot \phi}{\beta} \cdot i_A\right)\left(\frac{\alpha \cdot \phi}{\beta} - \frac{\alpha \cdot \phi}{\beta} \cdot i_A\right)}}$$

(Eq. 2), where β is the sample size, ϕ is the number of images, i_A is the prevalence of images in the minority class, $\hat{\theta}$ is the estimated AUC, and α is the load factor. Fixing sample size and number of images constant and letting the estimated AUC be the AUC performance of the entire dataset, we thus arrive at a relationship between standard error of AUC and load factor. Two image prevalences were studied, with one being that of the dataset and the other exaggerated substantially.

2.3 Clinical data used in the study

A dataset of 1047 chest radiograph (CXR) images from COVID-19+ patients had been acquired between March and September of 2020, along with their status as admitted to the intensive care unit (ICU) or not within 24 hours of imaging. There was one image per patient. 12.2% of the patients had been admitted to the ICU; thus, the prevalence of images in the minority class was 12.2%. A previously-trained deep learning/artificial intelligence model using DenseNet121 architecture was used to predict patients' need for intensive care within 24 hours of imaging.[7] Each image in the dataset was assigned a unique image ID, which respectively formed the keys of the hash table. The AUC of the dataset in the task of predicting ICU admission had been previously determined to be AUC = 0.81.

2.4 Comparison between expected and actual variation in AUC

The distribution of AUC at particular load factors was analyzed with the following investigation: (1) N images were randomly selected per test subset by image ID according to two minority class prevalences of interest: (a) that of the dataset (12.2% ICU admission) and (b) 70% ICU admission. (2) A hash table mapping image IDs to test subset IDs was subsequently implemented, as shown in Figure 1. (3) The hash table was inverted, resulting in a mapping of test subset IDs to image IDs. (4) AUC was calculated per test subset for ICU classification by the deep learning model and clinical outcome. (5) Steps 1 - 4 were repeated for 100 iterations. (6) The preceding steps were repeated at various test subset sizes of interest, resulting in different load factors. The 95% confidence interval of AUC at various load factors was then evaluated for the entire dataset. The confidence interval was calculated with *a posteriori* bootstrapping and then taking the 2.5 and 97.5 AUC percentiles of the resulting samples. Figure 2 gives an overview of the procedure.

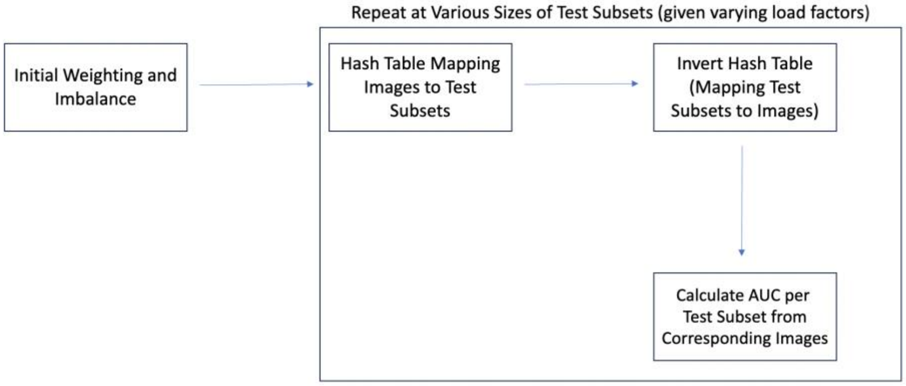
Repeat at Various Sizes of Test Subsets (given varying load factors)
Initial Weighting and Imbalance
Hash Table Mapping Images to Test Subsets
Invert Hash Table (Mapping Test Subsets to Images)
Calculate AUC per Test Subset from Corresponding Images

Figure 2: Overview of calculation of area under the receiver operating characteristic curve (AUC) for each test subset from hash table iterations at various load factors, following a specified class imbalance.

To evaluate the potential for the load factor to be used in optimizing the number of images used in algorithm evaluation, AUC distributions were evaluated across variations in load factor. Load factors varied from 0.5 to 5 in increments of 0.5 and 10 to 210 in increments of 20. The expected standard error of AUC as a function of load factor was evaluated using Eq. 2, assuming 100 test subsets and 1048 images with load factors of 1 to 320 in increments of 1. Interquartile ranges (IQRs) of AUC distributions from the clinical data were then compared with the expected variation of AUC with respect to load factor. The IQR was chosen as a measure comparable to the standard error without assuming that the AUCs from the clinical data were normally distributed. The analytical relationship between standard error of AUC and the load factor as a predictor for the optimal load factor threshold that reduces over or under-estimation of performance and data reuse was considered in the context of actual performance from the clinical dataset. Standard errors of AUC were calculated using the analytical relationship from Eq. 2, and assumed 100 test subsets and 1048 images. Sampling was conducted without replacement for small test subset sizes and with replacement when the size of the test subset exceeded the size of the dataset.

3. RESULTS

Figure 3 provides boxplots that display the AUC performance at various load factors, both with and without replacement in sampling. As expected, the variation in AUC decreased as load factor increased at both prevalences, due to the increase in number of images used as load factor increases. The median AUC of the *a posteriori* bootstrap is also shown.

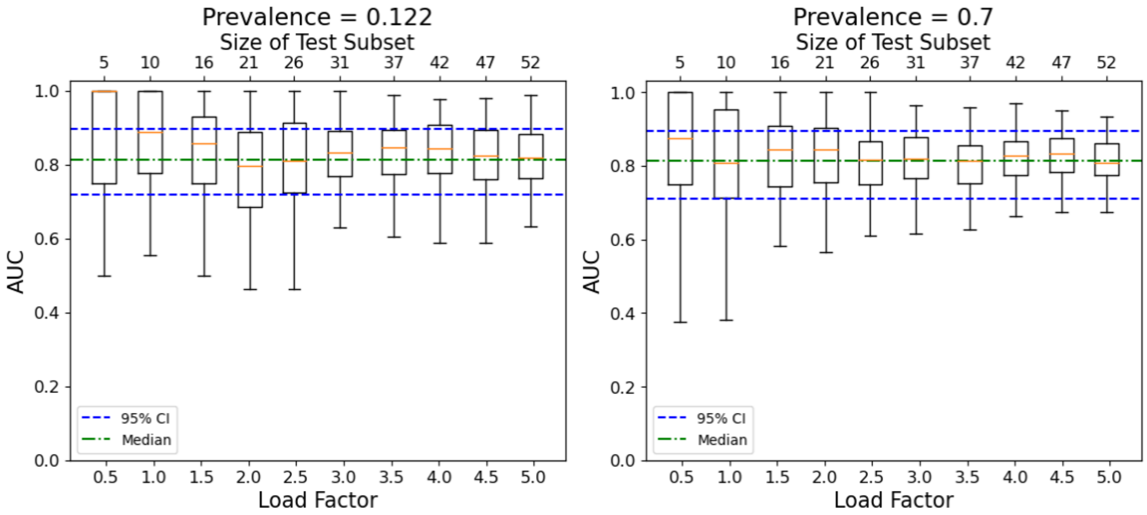

Figure 3: Boxplots of AUC from load factors 0.5-5.0: (left) prevalence of the original data set (12% admitted to the ICU) , (right) prevalence different from the original dataset (70% admitted to the ICU). Images were selected by test subsets *without* replacement (i.e., small subset test sizes). The *a posteriori* 95% CI is shown in blue, and the median AUC of the entire dataset is shown in green.

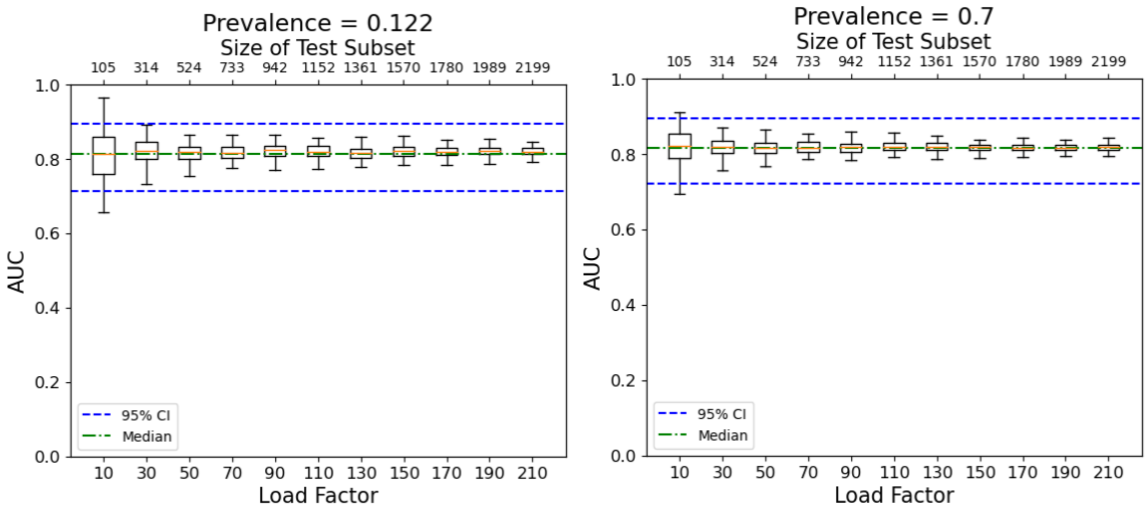

Figure 4: Box plots of AUC from load factors 10-210: (left) prevalence of the original data set (12% admitted to the ICU) , (right) prevalence different from the original dataset (70% admitted to the ICU). Images were selected by test subsets *with* replacement (i.e., large subset test sizes). The *a posteriori* 95% CI is shown in blue, and the median AUC of the entire dataset is shown in green.

Figure 5 shows results from the dataset (IQR) and expected results from Eq. 2 (standard error of AUC with respect to load factor). Labels in Figure 5b provide preliminary insight into an optimal range of load factors. The inverse relationship between load factor and AUC standard error may suggest a lower bound, where further decreasing the load factor leads to increasing the standard error beyond an acceptable point. Increasing load factor past a potential upper bound results in increased image reuse, as load factor measures the average amount of times an image in the dataset has been selected.

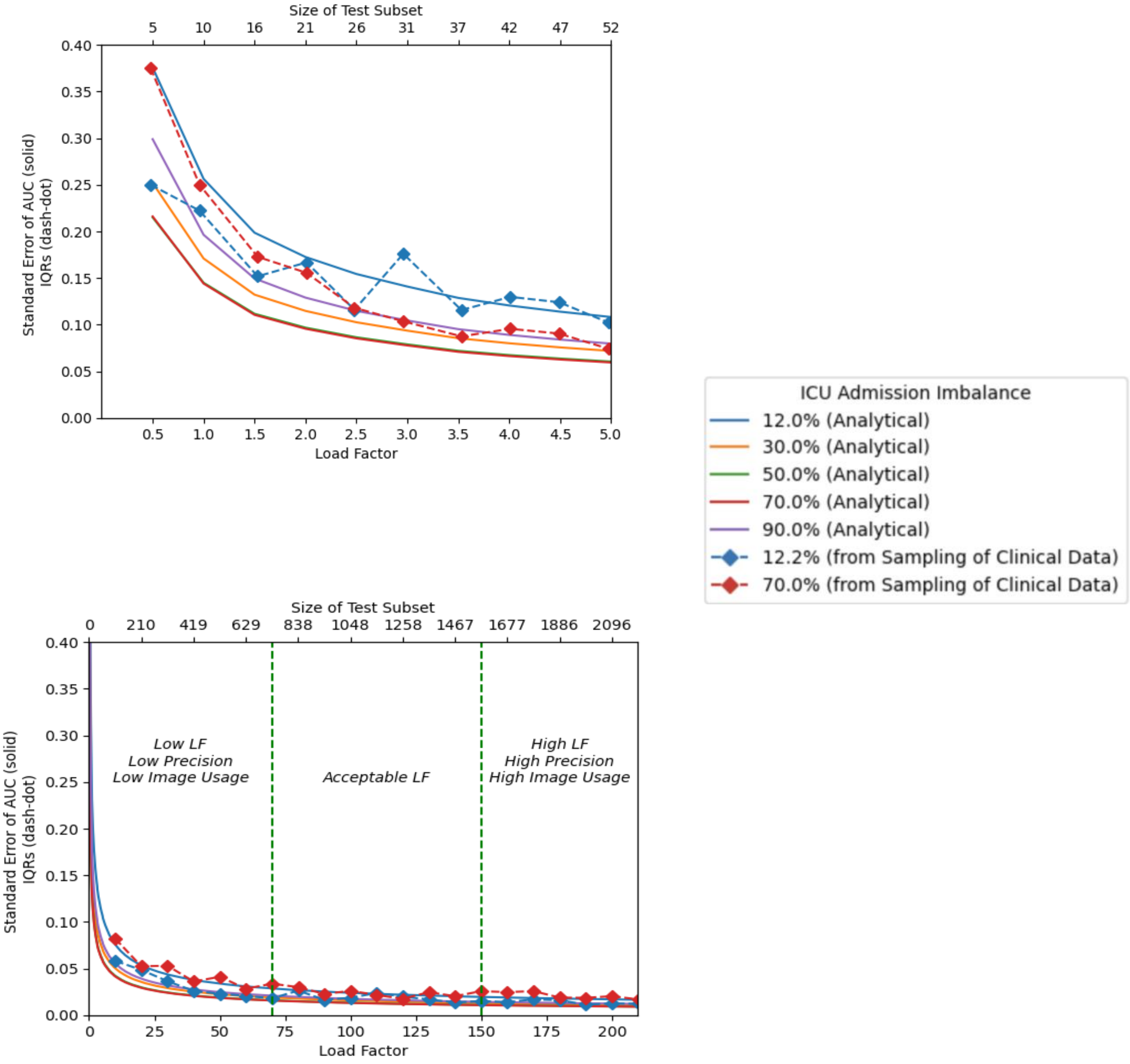

Figure 5a (top): Standard error (from analytical expression) or interquartile range (IQR, from dataset) with respect to load factors 0-5. The solid lines are representative of derived standard error from Eq. 2. As a comparison, interquartile ranges using the same prevalence parameters as Figures 1a and 1b are displayed by diamond markers connected by dashed lines. Images were selected by test subsets *with* replacement. **Figure 5b (bottom):** Standard error or IQR for load factors 10-210. Labels suggesting low precision and image reuse may point to establishing an optimal load factor threshold in future studies. Images were selected by test subsets *with* replacement. LF: load factor.

4. DISCUSSION

This study demonstrates the potential for a hash table implementation, which assigns images to test subsets for the purpose of sequestered test set design, as a method for studying data reuse in algorithm evaluation. The results show, through mathematical derivation and clinical data, a relationship between load factor and AUC standard error. The load factor can then be bounded by a lower and upper threshold, with the former representing the greatest probability of AUC over/underestimation and the latter for image resampling. Specifying such boundaries will be the topic of future work. We note that an optimal load factor is typically around 0.75 to 1 in the field of computer science, but as shown by the differences in AUC distributions between Figures 1 and 2, the optimal load factor in the context of algorithm development in medical imaging can be much greater. This study, while performed in the context of patient images, can be generalized to a variety of use-cases.

5. CONCLUSION

Through the mapping of images to test subsets via a hash table implementation, load factor can be a simple metric that supports the establishment of task-based test subsets from a sequestered data set such that the extent of algorithm performance overestimation and data reuse is minimized. This work contributes to the use of sequestered datasets by providing organizations with a greater understanding of the overall state of their dataset while continuing to uphold the integrity of sequestration.

ACKNOWLEDGEMENTS

MIDRC is supported by the National Institutes of Health, National Institute of Biomedical Imaging and Bioengineering contract 75N92020D00021.

REFERENCES

[1] Baughan, N., Whitney, H. M., Drukker, K., Sahiner, B., Hu, T., Kim, G. H., McNitt-Gray, M., Myers, K. J. and Giger, M. L., "Sequestration of imaging studies in MIDRC: stratified sampling to balance demographic characteristics of patients in a multi-institutional data commons," JMI **10**(6), 064501 (2023).

[2] Thompson, W. H., Wright, J., Bissett, P. G. and Poldrack, R. A., "Dataset decay and the problem of sequential analyses on open datasets," eLife **9**, e53498 (2020).

[3] Gossmann, A., Pezeshk, A., Wang, Y.-P. and Sahiner, B., "Test Data Reuse for the Evaluation of Continuously Evolving Classification Algorithms Using the Area under the Receiver Operating Characteristic Curve," SIAM Journal on Mathematics of Data Science **3**(2), 692–714 (2021).

[4] Tapia-Fernández, S., García-García, D. and García-Hernandez, P., "Key Concepts, Weakness and Benchmark on Hash Table Data Structures," Algorithms **15**(3), 100 (2022).

[5] Maurer, W. D. and Lewis, T. G., "Hash Table Methods," ACM Comput. Surv. **7**(1), 5–19 (1975).

[6] Hanley, J. A. and McNeil, B. J., "The meaning and use of the area under a receiver operating characteristic (ROC) curve," Radiology **143**(1), 29–36 (1982).

[7] Li, H., Drukker, K., Hu, Q., Whitney, H. M., Fuhrman, J. D. and Giger, M. L., "Predicting intensive care need for COVID-19 patients using deep learning on chest radiography," JMI **10**(4), 044504 (2023).