An Analysis of Machine Learning Practices in Medical Imaging

A Technical Report
presented to the faculty of the
School of Engineering and Applied Science
University of Virginia

by

William Belcher

May 3, 2024

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

William Belcher

Technical Advisor: Rosanne Vrugtman

An Analysis of Machine Learning Practices in Medical Imaging

CS4991 Capstone Report, 2024

William Belcher
Computer Science
The University of Virginia
School of Engineering and Applied Science
Charlottesville, Virginia USA
wpb7nd@virginia.edu

ABSTRACT
Machine learning analysis in medical imaging can improve the accuracy, speed, and cost of specialists' diagnoses, but is hindered in practice by poor performance and misleading results. To mitigate differences between testing and real-world performance, I propose a greater degree of transparency in data collection and algorithm development. Biases in training data and model evaluation can misrepresent the effectiveness of a proposed analysis tool. Data and algorithm transparency allows researchers, medical specialists, and developers to make more informed decisions on ways to publicize and use machine learning models. Future steps include examining the regulatory and legal implications of proposed methods and formalizing specific improvements in evaluating medical imaging models.

1. INTRODUCTION

Growing interest in recent machine learning advancements has promised solutions to a wide range of problems. In the medical imaging field, machine learning may help experts more accurately detect certain conditions by highlighting features of a scan that a professional may have otherwise missed. Conversely, a poorly performing model may yield false results, interfering with professionals' reasoning. Algorithms based on deep neural networks are a popular approach to computer vision problems, but inherently lack transparency in decision making and depend on large training datasets.

Labeling medical imaging datasets depends on the opinion of medical professionals, therefore the process of curating accurate and unbiased data is both expensive and nuanced. Furthermore, regulatory and cost constraints may lead to the creation of datasets which fail to cover a representative spectrum of patients and conditions. These dataset biases may lead to better model performance in testing than in the real world. The black box, or unclear, nature of how deep neural networks and similar algorithms make decisions limits the degree to which medical professionals can trust algorithm recommendations.

2. RELATED WORKS

Several studies cover issues concerning dataset bias and its implications for model generalization. Varoquaux and Cheplygina (2022) examine practices concerning dataset bias and model evaluation as it relates to medical imaging and provide some recommendations for improving current standards. Researchers at Stanford propose a set of guidelines for standardizing labeling procedures in preparing medical imaging datasets (Willemink, et al, 2020). They emphasize establishing a set criterion for labeling images such that models can make accurate distinctions between conditions. These publications do not address transparency between dataset aggregators and

users, or the implications of black box algorithms on decision making.

A group of Canadian researchers examine the nation's current guidelines on the use of machine learning models in healthcare and provide recommendations for how these guidelines may be improved or clarified (Da Silva, et al, 2022). They focus primarily on issues concerning public safety, data privacy, and algorithmic bias. While furthering discussion on how regulators may enforce best practices for machine learning in healthcare, they do not propose improvements for algorithm development and transparency.

3. PROPOSAL DESIGN

Practices in machine learning as applied to medical imaging must address crucial aspects including dataset and labeling bias, a focus on metrics during model development, and the role of communication between medical professionals and engineers. Each of these factors play a significant role in the robustness, interpretability, and practical utility of tools in the industry.

3.1 Dataset and Labeling Bias

Dataset bias is a critical consideration in developing any machine learning model, where aggregated data is not sampled in such a way that it is representative of the problem space. These biases, while not obvious during development, severely hinder the generalizability of models in practice. For example, a computer vision model used to identify vehicles in an image may have difficulty classifying a motorcycle if the dataset contained primarily cars. Computer vision algorithms, such as those used in medical imaging, generally rely on large, diverse datasets due to the complexity of making inferences from images. Medical image datasets, however, are generally smaller and geographically isolated, which perpetuates biases in data (Willemink, et al, 2020). Better transparency in the data aggregation process allows developers and medical professionals to identify and rectify potential dataset biases.

Curating raw images into usable datasets requires humans to manually label samples. This, however, introduces individuals' inherent biases into labeled data. In the field of medical imaging, labels can be influenced by medical professionals' own experience and opinions. Rigorous guidelines and standardized protocols may be introduced to help objectively differentiate between labels. Medical diagnoses are nuanced, and professionals' assessments of a patient's state can be subjective. Whenever possible, assignment of labels should be supported by empirical evidence and clinical data associated with each image. To further the integrity of labeled data, cross-validation by different annotators and medical professionals should be used to mitigate subjective bias.

3.2 Overemphasis on Metrics

During model development, focusing on a small set of metrics can lead to poor generalization and practical performance. Optimizing for a small set of performance metrics, such as accuracy or precision, can overstate a model's ability to make inferences on new data. This is exacerbated by the limited size of medical image datasets, which can invite model overfitting. Data analysis challenges are often used to evaluate the potential application of different machine learning algorithms in medical imaging and exemplify this problem. The results of more prestigious challenges receive significant attention and can influence the direction of further work in the area.

The ranking system for model performance in these challenges is not standardized, and in a sample of 383 challenge tasks, minor changes in the evaluation metric radically changed model rankings (Maier-Hein, et al, 2018). This

suggests that winning models may be highly optimized for the conditions of the challenge, rather than demonstrating generalizable performance on the problem. Careful cross-validation of model performance, transparency in competition ranking methods, and the use of multiple performance metrics should improve the robustness of models and the reproducibility of results.

3.3 Communication and Feedback with Medical Professionals

Effective communication between medical professionals and engineers throughout the model development process is necessary to create relevant and useful tools for real-world practice. Specifically, feedback from medical professionals should ensure the outputs provided by the model are interpretable and improve the quality of care. This also establishes channels for users to suggest areas of refinement, improving the iterative development cycle of tools. Medical professionals should also work in conjunction with engineers to develop comprehensive criteria to evaluate a model's ability to generalize performance to a clinical setting.

4 ANTICIPATED RESULTS

While direct results are dependent implementation, more engagement of medical professionals in the development and evaluation of medical imaging models should yield more effective tools and heightened trust between users and developers. Insightful and interpretable model analysis of medical images can improve both the accuracy of diagnoses and quality of care. Through their involvement in the development process, medical professionals will have a better understanding of the benefits and limitations of tools, building trust and increasing adoption.

Standardization and transparency in the data aggregation, labeling, and model evaluation processes should provide several benefits to model generalization and the reproducibility of results. Clear documentation of data aggregation and labeling processes allows professionals to validate methods, identify potential biases, and ensure data is sufficient to develop a robust model. Detailed guidelines for the labeling process, combined regular review of labels by medical professionals and clinical data, yield more consistent datasets and mitigate labeling bias. While this improves the quality of labeled data, it would also increase the complexity and cost of dataset development which limits sample size. Using multiple evaluation metrics in data challenges and model development provides better insight into applicable algorithms and a stronger analysis of generalizability.

5 CONCLUSION

The advancement of machine learning in medical imaging can improve the accuracy of diagnoses and the quality of patient care, but is hindered by dataset and labeling bias, insufficient validation criteria, and a lack user trust in machine learning tools. The proposed approach aims to address these challenges through emphasizing transparency in data collection, labeling, and model development processes. Proposals include methods to mitigate biases, increase user trust, and improve the reliability of models used in medical diagnostics. Application of these strategies provides a framework for better model generalization in medical imaging, benefiting patients, medical professionals, and engineers.

6 FUTURE WORK

Exploring the current state of medical professional training and interaction with machine learning in medical imaging would help inform further strategies for increasing user trust. Additionally, analysis legal and regulatory restrictions on health data aggregation may raise challenges in mitigating

dataset bias. Further development may include formalizing methods for evaluating classifier performance based on specific metrics.

REFERENCES

Da Silva, M., Flood, C. M., Goldenberg, A., & Singh, D. (2022). Regulating the Safety of Health-Related Artificial Intelligence. *Healthcare policy = Politiques de sante*, *17*(4), 63–77. https://doi.org/10.12927/hcpol.2022.26824

Maier-Hein, L., Eisenmann, M., Reinke, A., Onogur, S., Stankovic, M., Scholz, P., Kopp-Schneider, A. (2018). Why rankings of biomedical image analysis competitions should be interpreted with care. Nature Communications, 9(1), 5217. doi:10.1038/s41467-018-07619-7

Varoquaux, G., & Cheplygina, V. (2022). Machine learning for medical imaging: methodological failures and recommendations for the future. *Npj Digital Medicine*, *5*(1), 48. doi:10.1038/s41746-022-00592-y

Willemink, M. J., Koszek, W. A., Hardell, C., Wu, J., Fleischmann, D., Harvey, H., … Lungren, M. P. (2020). Preparing Medical Imaging Data for Machine Learning. *Radiology*, *295*(1), 4–15. doi:10.1148/radiol.2020192224