Quantifying Differences Between Machine Learning Classification Models Applied to Cancer Microscopy Data

William Franz Lamberti[1], Chongzhi Zang[1*]

[1]Center for Public Health Genomics, University of Virginia, Charlottesville, VA, USA
[*]Corresponding author: zang@virginia.edu

The use of machine learning (ML) to analyze Microscopy data has grown in popularity in recent years [1-3]. However, the complexity of ML models and their resulting analyses have typically resulted in their treatment as "black boxes", which hinders scientific interpretations. The ability to explain and interpret estimated models is vital for the scientific process. To that end, we present techniques to interpret and explain differences between different ML models using results from four separate experimental datasets.

We will use two classification examples to demonstrate our approach. The first example compares two separate models that were built on two independent microscopy datasets obtained from white blood cell samples from separate labs [4, 5]. The second example compares the classification of genetically-relevant features from human colon cancer cell 3D structured illumination microscopy (3D-SIM) data under two different treatments [6]. A generalized schematic of our approach is provided in Figure 1.

The white blood cell cancer data came from two independent labs using different experimental setups [4, 5]. One dataset was segmented by an expert oncologist while the other was not. After applying a custom universal segmentation algorithm, two independent random forest (RF) models were created to distinguish normal from cancer white blood cells. These models used the same shape, color, and texture metrics. However, the two models did not use the variables with the same weight. Variable importance (VI) can be used to help quantify the differences between the models' use of the features. Distribution plots of absolute differences between the pairs of VIs provide useful insights. Values close to zero would indicate that the variables produce similar impacts to the models. Larger non-zero values would indicate that some of the variables produce different impacts to each model. A useful visualization tool for this is a boxplot as exemplified in Figure 2a. For the case study provided, there are a large number of variables that deviate from zero. Thus, there is evidence that the models do not use the variables in the same manner.

The second case study analyses image data obtained using 3D-SIM in an effort to compare different genetic objects of interest, such as open and closed chromatin domains, from one another in a human colon cancer cell and a cell under a treatment [6]. These data are available for download from the original authors in the reference [6]. For this work, two independent logistic regression models were built using the Least Absolute Shrinkage and Selection Operator (LASSO) to select variables for each model [7]. However, each LASSO reported different numbers of variables to classify the genetic objects of interest. The LASSO's results for the control and data reported 4 and 7 variables, respectively. The increase in model complexity indicates that the cancer treatment results a need for a wider range of cell features versus the control set. We can begin to quantify the parameter differences between the two models using an absolute value of the differences of the parameter values. For example, the logistic regression parameter for the relative major axis length differs by a factor of $\frac{0.801}{0.094} = 8.521$. The overlapping features and their associated parameter values are provided in Figure 2b. Thus, the shape of these genetic objects of interest found in human colon cancer cells are fundamentally changing due to the treatment.

CrossMark

However, directly assigning a physical meaning and relative importance to each variable across all possible ML algorithms remains challenging. Where possible, this step could benefit from additional experimentation to reveal these relationships. Another solution is to use ML algorithms, like logistic regression, which produce weights for the variables with probabilistic meanings. Relating ML parameter weighting to the changes observed in a physical, chemical or biological system in a directly interpretable way is an important and emerging area of exploration for ML algorithm developers.
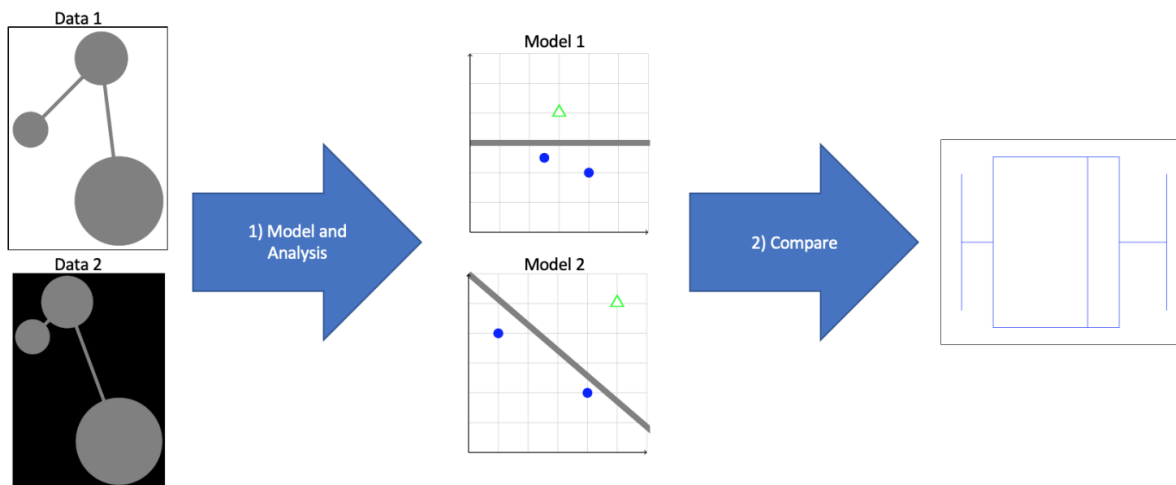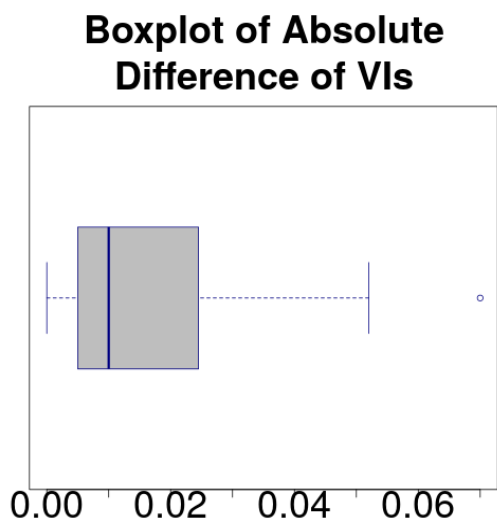

Figure 1. A schematic of quantifying difference between ML models.

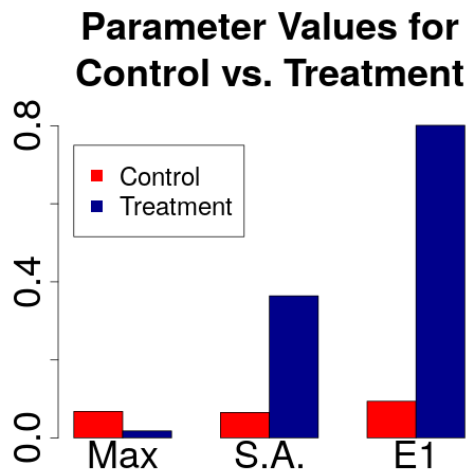(a)   (b)

Figure 2. Examples of methods to visualize the differences between ML model parameters. (a) Boxplot to compare the paired VI values between two RF models. (b) Barplot of different parameter values between a colon cancer control and treatment. "Max" stands for Maximum pixel intensity. "S.A." stands for surface area. "E1" stands for the first eigenvalue.

References:

[1] L. von Chamier, et.al., "Democratising deep learning for microscopy with ZeroCostDL4Mic," Nature Communications, vol. 12, no. 1, p. 2276, Apr. 2021. [Online]. Available: https://www.nature.com/articles/s41467-021-22518-0

[2] L. Waller and L. Tian, "Machine learning for 3D microscopy," Nature, vol. 523, no. 7561, pp. 416–417, Jul. 2015.[Online]. Available: http://www.nature.com/articles/523416a

[3] V. Zinchuk and O. Grossenbacher-Zinchuk, "Machine Learning for Analysis of Microscopy Images: A Practical Guide," Current Protocols in Cell Biology, vol. 86, no. 1, p. e101, 2020. [Online]. Available:https://onlinelibrary.wiley.com/doi/abs/10.1002/cpcb.101

[4] R. D. Labati, V. Piuri, and F. Scotti, "All-IDB: The acute lymphoblastic leukemia image database for image processing," in 2011 18th IEEE International Conference on Image Processing. Brussels, Belgium: IEEE, Sep.2011, pp. 2045–2048. [Online]. Available: http://ieeexplore.ieee.org/document/6115881/

[5] A. Gupta and R. Gupta, Eds., ISBI 2019 C-NMC Challenge: Classification in Cancer Cell Imaging: Select Proceedings, ser. Lecture Notes in Bioengineering. Springer Singapore, 2019. [Online]. Available: https://www.springer.com/gp/book/9789811507977

[6] M. Cremer, et.al. , "Cohesin depleted cells rebuild functional nuclear compartments after endomitosis," Nature Communications, vol. 11, no. 1, p. 6146, Dec. 2020. [Online]. Available: https://www.nature.com/articles/s41467-020-19876-6

[7] G. James, D. Witten, T. Hastie, and R. Tibshirani, Eds., An introduction to statistical learning: with applications in R, ser. Springer texts in statistics. New York: Springer, 2013, no. 103, oCLC: ocn828488009.