ANNALEE CORCORAN, Dept of Computer Science, Mathematics, and Engineering, Shepherd University, Shepherdstown, WV 25443, and JASON R MILLER, Dept of Computer Science, Mathematics, and Engineering, Shepherd University, Shepherdstown, WV 25443.
Analyzing Lung Cancer Data for Machine Learning.

Data preparation is a critical step for any machine learning experiment. We have analyzed a dataset derived from images of human male lung cancer tumors. These tumors had been analyzed with genetic markers to identify Y-chromosome loss, which was the case in about half of the samples. Whole slide images (WSI) had been collected and H&E stained by collaborators. We had processed the images with the CellProfiler software to extract numeric features. In this study, we analyzed the data in preparation for training a convolutional neural network to predict Y-chromosome loss from the extracted features, thereby recapitulating the genetic marker analysis. Using Excel and Python, we identified uninformative features and missing data. We predict that data cleaning, informed by these results, will improve the chances of successful machine learning.