Cancer detection for white urban Americans

Kajsa Møllersen[1], Lars Ailo Bongo[2], Masoud Tafavvoghi[1]
[1]Department of Community Medicine, [2]Department of Computer Science
UiT - The Arctic University of Norway

Keywords: Machine learning, Bias, Public repositories

Introduction: Development, validation and comparison of machine learning methods require access to data, sometimes lots of data. Within health applications, data sharing can be restricted due to patient privacy, and the few publicly available data sets become even more valuable for the machine learning community. One such type of data are H&E whole slide images (WSI), which are stained tumour tissue, used in hospitals to detect and classify cancer, see Fig. 1. The Cancer Genome Atlas (TCGA) has made an enormous contribution to publicly available data sets. For breast cancer H&E WSI they are by far the largest data set, with more than 1,000 patients, twice as many as the second largest contributor, the two Camelyon competition data sets [1] with 399 + 200 patients.

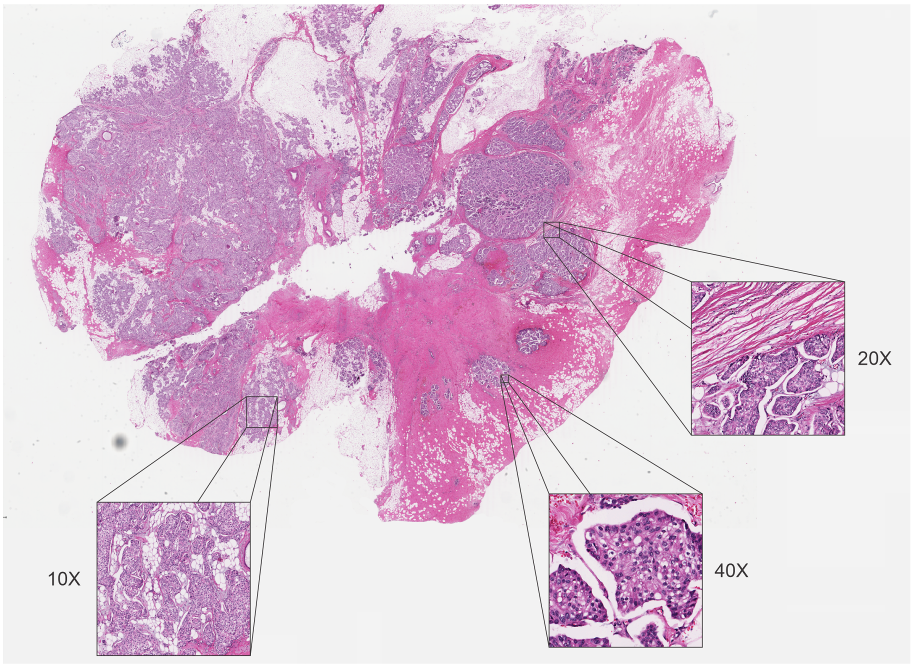

Figure 1: An example of breast H&E WSI from the TCGA-BRCA data set.

Objective: Large data sets can get a too large place, frequently used by researcher, gaining fame and ending up dominating a field. In this study, we have investigated the use of TCGA-BRCA[2] and 11 other publicly available breast cancer H&E WSI data sets for machine learning, with the objective of revealing possible overuse of one data set.

Contribution: Whereas the use of a particular data set in itself is not problematic, lack of awareness regarding what this data set represents can introduce severe bias into a field. Our contribution is shining a light on this bias, and suggests a solution to how it can be handled by the researchers in their next publication.

Findings: Among over 100 articles, nearly half of them use TCGA-BRCA, making this data set clearly dominating in its field. It is barely mentioned in any of the publication that the TCGA-BRCA patients are Americans, predominantly white, from urban areas, and much younger than the average breast cancer patient.

Future work: Future publications should include a description of the patients, so that the readers and authors are aware of the limited population from which the data is drawn, and thus avoid making non-valid general claims about the findings. A systematic review of available data sets will be published to make it easier for researchers to have more diversity in their studies.

[1] https://camelyon17.grand-challenge.org/

[2] https://portal.gdc.cancer.gov/projects/TCGA-BRCA