Deep Learning in Cervical Cancer: Searchable Catalogs and Smart Data Curation

Daniela Ushizima[1,2,3], Andrea Campos Bianchi[4], Fatima Sombra Medeiros[5], Claudia Carneiro[6]
[1]Computing Sciences, Lawrence Berkeley National Laboratory, [2]Berkeley Institute for Data Science, UC Berkeley, [3]Bakar Institute, UC San Francisco, [4]Computing Department, Federal University of Ouro Preto, [5]Computer Engineer Department, Federal University of Ceara, [6]School of Pharmacy, Federal University of Ouro Preto

According to the World Health Organization, cervical cancer is the fourth-most common cancer in women and the leading cause of cancer death in 42 countries. The adoption of Pap tests, a cytopathological procedure, has reduced incidence and death associated with cervical cancer by 65% in the past 40 years, but inspection continues largely dependent on human vision. Development of Machine Learning approaches to automate cell analysis has been highlighted to scale the analysis of Pap tests, but the absence of high-quality curated datasets has prevented the development of strategies that truly improve cervical cancer screening.

The CRIC Cervix Collection [1] is a promising start toward overcoming this challenge: at 11,534 cell images, it is the world's largest collection of images of cervical cells collected conventionally through Pap smears. It is open source and searchable, with the aim of advancing reproducible research and FAIR (Findable, Accessible, Interoperable, and Reusable) data. Each cell image in the database was manually classified by a team of cytopathologists using the Bethesda System, a standardized nomenclature for cervicovaginal cytology.

As part of this effort, the Center for Recognition and Inspection of Cells (CRIC) has also deployed computational tools to support remote cell screening and development of more efficient and effective methods for cell segmentation and classification, especially for the detection of cervical cancer. CRIC is a consortium of international researchers that has provided algorithms, software and cell collections to the scientific community, delivering digital pathology capabilities under four main efforts: (i) Cell segmentation based on deep learning [2], and smart data curation tools for cell nuclei detection and classification, broadly tested in hematoxylin-eosin stained images (ii) Searchable image database for creation of 'searchable catalogs' [3], with a computational platform with ability to access cell collections with millions of classified examples, and that enables image classification and segmentation of new samples. The most popular set of samples have been deployed under the name CRIC CERVIX collection, which uses seven classification lesion types as considered by the Bethesda System, (iii) CitoFocus, a collaborative tool for cell analysis, and numerous programs for (vi) Continuous education.

The next frontier is to explore new Data Science methodologies, combining both images and respective biomedical metadata to improve Pap test pre-screening. Future developments will also include creating new deep learning models for classification allied to exploring High Performance Computing to provide automated and accurate classification for other cell types, which will increase chances of early cancer diagnosis, and preventative treatment.

[1] Rezende, Silva, Bernardo, Tobias, Oliveira, Machado, Costa, Medeiros, Ushizima, Carneiro, Bianchi, "Cric searchable image database as a public platform for conventional pap smear cytology data", *Nature Scientific Data* 2021.

[2] Araújo, Silva, Resende, Ushizima, Medeiros, Carneiro, Bianchi, "Deep Learning for Cell Image Segmentation and Ranking", *Computerized Medical Imaging and Graphics*, Mar 2019.

[3] Araujo, Silva, Ushizima, Parkinson, Hexemer, Carneiro, Medeiros, "Reverse Image Search for Scientific Data within and beyond the Visible Spectrum", *Expert Systems and Applications* 2018.