Cell Reports Methods

CellPress
OPEN ACCESS

Preview

# Practical machine learning for disease diagnosis

Huw D. Summers[1,*]
[1]Department of Biomedical Engineering, Swansea University, Swansea, UK
*Correspondence: h.d.summers@swansea.ac.uk
https://doi.org/10.1016/j.crmeth.2021.100103

**Deep learning neural networks are a powerful tool in the analytical toolbox of modern microscopy, but they come with an exacting requirement for accurately annotated, ground truth cell images. Otesteanu et al. (2021) elegantly streamline this process, implementing network training by using patient-level rather than cell-level disease classification.**

The advent of automated, high-throughput microscopy has revolutionized cell biology, allowing focus on the particular, e.g., detailed inspection of rare cells, or providing meta-level statistics of population metrics (Ljosa and Carpenter, 2009). The incorporation of imaging capability into flow cytometers has further advanced the field (Blasi et al., 2016), especially in application to clinical diagnostics as hematological image analysis becomes possible (Ogle et al., 2016). However, automated image acquisition and quantification inevitably leads to a requirement for automated image analysis (Caicedo et al., 2017), and this has been provided by evermore sophisticated machine learning approaches (Sommer and Gerlich, 2013).

Early advances in machine learning delivered *expert systems*—computer models based on expert knowledge in which data metrics and the rules that linked them were user defined. Thus, in essence, the machine simulated the analytical steps of the human brain, bringing much enhanced speed and reliability (Buchanan and Smith, 1988). Over time the algorithmic operations of the machine have grown increasingly complex and opaque to the human user. This process has led us to today's deep learning systems in which automated correlation, classification, and decision making is done within artificial neural architectures. Thus, we have progressed from machine learning that aimed to model the decision making processes of the brain to systems that mimic the brain itself. The benefit of this computational development are extremely powerful deep learning networks capable of discovering information on processes and interactions that is hidden within data patterns. The requirement for machine supervision still remains, however, because expert knowledge is needed to define the labeled datasets on which the deep learning networks are trained. It is this aspect that Otesteanu et al. (2021), address in their paper, presenting machine learning for clinical diagnostics on the basis of T cell morphology and implemented by using minimized (*weak*) supervision.

The standard approach to neural network training uses *strong* supervision in which input datasets are individually classified and labeled at the level of individual data entries (Zhou, 2018). For example, in cell-based diagnostics, experts have to spend a great deal of time inspecting cell images and annotating them according to whether they correspond to a phenotype associated with a healthy or diseased patient. This annotated "ground-truth" dataset is then used to train the network to automatically recognize the designated cell types (Doan et al., 2018) (Figure 1). This approach is resource-heavy, requiring expert knowledge, a lot of time, and accuracy in data labeling. It also assumes a-priori knowledge of what cells are important and what they look like, but what about unknown populations? How can we use machine learning in the case of clearly indicated disease, with known physiological symptoms, but no knowledge of the cellular biomarkers of the pathology? Otesteanu et al. (2021) present a *weak* learning approach (*iCellCnn*) that removes the need for cell-level, ground truth annotation by training the neural network with a collection of cells labeled according to patient status. They term this a "bag of cells" approach, and its novelty lies in the use of patient-level classification (Figure 1). Instead of recognizing individual cells whose morphology indicates disease, the network is trained to recognize distributions of image features, collected from a population of cells. In the authors' words they use "weak labeling of a set of inputs, instead of a strong labeling of individual inputs."

The disease considered is a blood cancer, Sézary syndrome. This is a T cell lymphoma that is characterized by anomalous cerebriform (brain-like) morphology of T cell nuclei. In *iCellCnn*, multiple T cell images, obtained from an individual patient blood sample, form the input to a convolutional autoencoder—a feature-extracting neural network. This combines morphological feature information from all cells within a one-dimensional feature vector. This vector, an abstracted description of the blood sample, is used as an input to a random forest classifier which indicates the probability of the presence of diseased cells in the input cell collection (those of cerebriform morphology). Thus, the training of the machine learning model defines morphological patterns of disease at the cellular level in a data-driven manner. As all T cells are presented to the neural network, it learns to ignore the non-disease-specific cells that might confuse classification of patient status. The authors benchmark their approach by comparing diagnoses to those obtained by using a strong learning approach, implemented by prior labeling of individual cells as disease-associated or healthy. Two levels of annotation are adopted:

1. *Naive*, in which the status of the patient is assigned to all of their cells

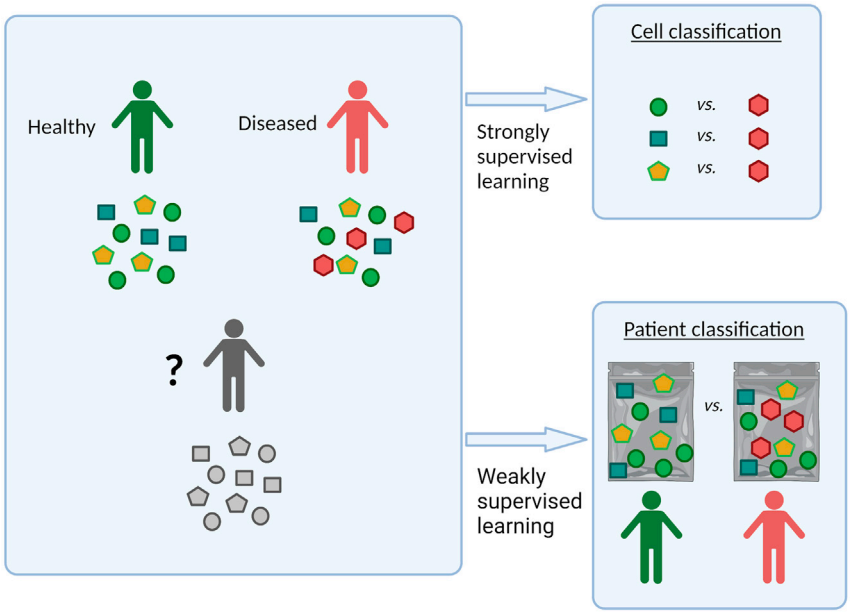


**Figure 1. Training approaches for disease diagnosis by machine learning**

Shown on left, representative cells are harvested from healthy and diseased donors. The aim is to train the artificial neural network to recognize these subsets so that it can determine the status of an undiagnosed patient (in gray). Shown on right, feature extraction from the cell images creates the information set on which the network bases its classifications. This might be ambiguous as some healthy cells (indicated in orange) can present similar features to diseased cells (indicated in red). Supervised training is implemented at cell level to train for recognition of diseased cells (strongly supervised), or with "bags of cells" to train for recognition of a diseased patient (weakly supervised). Figure created with Biorender (https://biorender.com/).

(i.e., 100% of cells from a healthy patient are labeled as healthy and vice versa for a diseased patient). This works on an assumption that patients with Sézary syndrome have an increased frequency of mutated T cell nuclei.

2. *Manual*, in which cells from a healthy patient are again naively annotated as healthy, and 1,000 expertly identified pathological T cells are annotated as diseased. Although disease-associated cells are explicitly identified, this approach still results in morphologically abnormal T cells from healthy individuals being labeled as "healthy," so there is still the potential for distortion of the model predictions.

Although both strong and weak approaches were able to distinguish between healthy and diseased patients, the weakly supervised training produced the most pronounced separation of healthy and diseased classifier scores, estimating ~14% prevalence of diseased cells in healthy and 85% in diseased patients.

In a nice addition to the scope of the paper, the authors use a custom-built microfluidic system to obtain the T cell images. This contains a 45 × 45 μm channel, in a polydimethylsiloxane (PDMS) monomer on a glass substrate, which elasto-inertially focuses cells within the fluid stream and can image >2,000 cells per patient. Image capture is achieved by using a ×60 objective lens and a CMOS camera. The technical simplicity of this device and the streamlined machine learning analysis provide an ideal toolset for ready adoption within clinical laboratories.

Although weakly supervised machine learning has previously been used in conjunction with imaging flow cytometry to analyze blood samples (Doan et al., 2020), this study by Otesteanu et al. (2021) is the first to use the approach to demonstrate disease diagnosis. Its data-driven approach is tailor made for clinical application because the medical determination of a patient's illness becomes the input label when training the neural network and the output decision of the machine. This opens the way to cell-agnostic diagnoses where the presence of disease can be detected but its effect on specific morphological traits of cells remains unknown. The black box nature of machine learning might be seen as an advantage in this situation given that it allows clinicians to bypass the complexity of cell morphology and its alteration by disease, safe in the knowledge that the accuracy of the computational decision making has been verified by comparison to expert medical opinion.

## ACKNOWLEDGMENTS

The author benefitted from stimulating discussions with Paul Rees on machine learning and cell classification.

## DECLARATIONS OF INTERESTS

The author declares no competing interests.

## REFERENCES

Blasi, T., Hennig, H., Summers, H.D., Theis, F.J., Cerveira, J., Patterson, J.O., Davies, D., Filby, A., Carpenter, A.E., and Rees, P. (2016). Label-free cell cycle analysis for high-throughput imaging flow cytometry. Nat. Commun. *7*, 10256.

Buchanan, B.G., and Smith, R.G. (1988). Fundamentals of expert systems. Annu. Rev. Comput. Sci. *3*, 23–58.

Caicedo, J.C., Cooper, S., Heigwer, F., Warchal, S., Qiu, P., Molnar, C., Vasilevich, A.S., Barry, J.D., Bansal, H.S., Kraus, O., et al. (2017). Data-analysis strategies for image-based cell profiling. Nat. Methods *14*, 849–863.

Doan, M., Vorobjev, I., Rees, P., Filby, A., Wolkenhauer, O., Goldfeld, A.E., Lieberman, J., Barteneva, N., Carpenter, A.E., and Hennig, H. (2018). Diagnostic potential of imaging flow cytometry. Trends Biotechnol. *36*, 649–652.

Doan, M., Sebastian, J.A., Caicedo, J.C., Siegert, S., Roch, A., Turner, T.R., Mykhailova, O., Pinto, R.N., McQuin, C., Goodman, A., et al. (2020). Objective assessment of stored blood quality by deep learning. Proc. Natl. Acad. Sci. USA *117*, 21381–21390.

Ljosa, V., and Carpenter, A.E. (2009). Introduction to the quantitative analysis of two-dimensional fluorescence microscopy images for cell-based screening. PLoS Comput. Biol. *5*, e1000603.

Ogle, L.F., Orr, J.G., Willoughby, C.E., Hutton, C., McPherson, S., Plummer, R., Boddy, A.V., Curtin, N.J., Jamieson, D., and Reeves, H.L. (2016). Imagestream detection and characterisation of

circulating tumour cells - A liquid biopsy for hepatocellular carcinoma? J. Hepatol. *65*, 305–313.

Otesteanu, C.F., Ugrinic, M., Holzner, G., Chang, Y.-T., Fassnacht, C., Guenova, E., Stavrakis, S., DeMello, A., and Claassen, M. (2021). A weakly-supervised deep learning approach for imaging flow cytometry based diagnosis of Sézary Syndrome. Cell Reports Methods *1*, 100094-1–100094-9.

Sommer, C., and Gerlich, D.W. (2013). Machine learning in cell biology - teaching computers to recognize phenotypes. J. Cell Sci. *126*, 5529–5539.

Zhou, Z.H. (2018). A brief introduction to weakly supervised learning. Natl. Sci. Rev. *5*, 44–53.