CYTOSET: A DEEP LEARNING MODEL FOR PREDICTING CLINICAL OUTCOMES FROM CYTOMETRY DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Flow and mass cytometry technologies are being increasingly applied in clinical settings, as they enable the simultaneous profiling of multiple proteins across millions of cells within a multi-patient cohort. In this work, we introduce CytoSet, a deep learning model that can directly predict a patient's clinical outcome from their collection of single cells profiled with technologies, such as, flow and mass cytometry. Unlike previous work, CytoSet explicitly models the collection of cells profiled in each patient as a set, allowing for the use of recently developed permutation invariant architectures. We show that CytoSet achieves state-of-the-art classification performance on a flow cytometry benchmark dataset, where the task is to classify > 300 infants who were exposed to HIV and remained uninfected (HEU) from control patients who were unexposed (UE). The strong classification performance achieved in this dataset suggests the great potential to interpret and study clinical cytometry data as sets.

1 INTRODUCTION

High-throughput single-cell technologies, such as flow and mass cytometry have a wealth of applications, ranging from systems immunology (Finak et al., 2016; Davis et al., 2017) to clinical monitoring (Hartmann et al., 2019). For use in clinical settings, mass cytometry in particular can simultaneously measure ~ 40 proteins at single-cell resolution (Spitzer & Nolan, 2016), across multiple patient samples. The rich information related to the diverse cell-populations and activated signaling pathways stored in each multi-dimensional single-cell measurement can be used to predict a patient's clinical outcome, or classification. Some previously studied examples of cytometry in clinical settings include studying the response to human immunodeficiency virus (HIV) (Aghaeepour et al., 2013) and in understanding why certain women experience pregnancy complications (Han et al., 2019).

Recently, a number of methods have been proposed to link the sets of profiled single cells from each patient to their clinical outcome or classification. These methods fall into one of two main categories. The class of 'gating-based' methods (Bruggner et al., 2014; Lun et al., 2017; Weber et al., 2019; Stanley et al., 2020) first clusters cells across all patient samples into homogeneous subsets according to the expression of the measured proteins. The abundances or relative proportion of each sample's cells assigned to each cluster is then used as an engineered feature vector that can be used to predict clinical outcomes.

Alternatively, the second class of methods are gating-free (Arvaniti & Claassen, 2017; Hu et al., 2019), and therefore operate on the single-cell level. CellCNN (Arvaniti & Claassen, 2017) uses a convolutional neural network (CNN) as an end-to-end model to learn the associated phenotype from multi-cell input. CytoDx (Hu et al., 2019) uses a two-level linear model to predict clinical outcome across individual cells. Although these techniques can successfully identify and achieve strong classification accuracy in small datasets, these methods often do not adapt well to larger datasets with batch effects and noisy measurements.

In this paper, we propose a new deep learning model called CytoSet to predict clinical outcomes from single-cell flow and mass cytometry data. The novelties of CytoSet are as follows: 1) CytoSet is an end-to-end model that can predict clinical outcomes directly on a set of cells profiled for each patient, rather than from a feature vector engineered through clustering or 'gating'; 2) CytoSet uses the permutation invariant network architecture inspired by Deep sets (Zaheer et al., 2017) to extract

the information from set-structured cytometry data; 3) CytoSet achieves state-of-the-art classification performance on a benchmark flow cytometry dataset and outperforms two baselines by a large margin.

2 BACKGROUND

2.1 SET MODELING

For set-structured data, each sample is a collection of unordered data points denoted by $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m\} \in \mathcal{X}^m$, where m is the cardinality of the set $\boldsymbol{X}$ and $\mathcal{X}$ denotes the domain of each element $\boldsymbol{x}_i$. For example, a set of multiple 3d points is often used to represent object shape. Since different sets are equal only if they have same elements, models that map a set to some target values must preserve the following permutation invariance property:

Definition 1 (Permutation Invariant) *Let $f : \mathcal{X}^m \to \mathcal{Y}$ be a function, then f is permutation invariant iff for any permutation $\pi(\cdot)$, $f(\boldsymbol{X}) = f(\pi(\boldsymbol{X}))$.*

One natural way to encourage permutation invariance is to cluster the elements of a set and output statistics, such as histograms as the feature vectors. Although clustering can preserve permutation invariance, it cannot be trained in a model using back-propagation because of the non-differentiability. To solve this problem, Zaheer et al. (2017) and Edwards & Storkey (2017) propose permutation invariant neural network architectures in which set-pooling layers play a key role in preserving permutation invariance and aggregating information over elements. Additionally, Zaheer et al. (2017) also showed the general form of functions that have permutation invariant properties. Recently, Lee et al. (2019) proposed a new transformer-based architecture that uses attention mechanisms for both encoding and aggregating features in a set.

2.2 FLOW AND MASS CYTOMETRY DATA

Let $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m\}$ denote a collection of cells profiled with mass or flow cytometry. Here, m is the number of cells and each $\boldsymbol{x}_i \in \mathbb{R}^p$ encodes the expression of p proteins in cell i. In cytometry experiments, the order in which cells are profiled has no biological relevance, so a collection of cells, $\boldsymbol{X}$, can be viewed as a set rather than a data matrix.

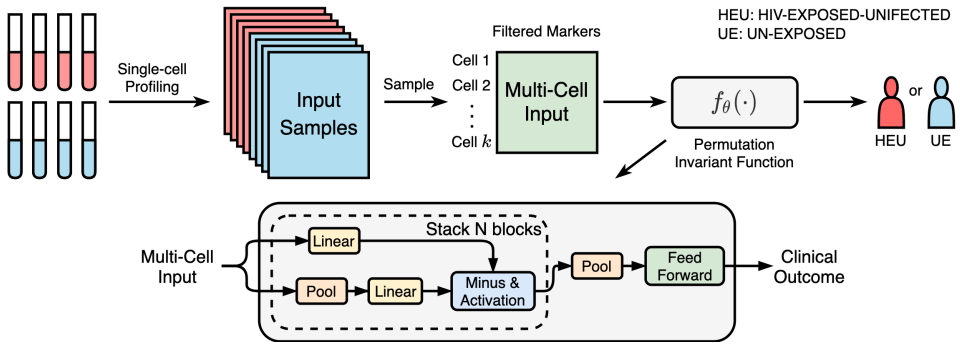

Figure 1: An illustration of the model architecture of CytoSet. On the FlowCap HIV benchmark dataset, the task is to classify each patient sample as exposed but uninfected with HIV (HEU) from unexposed (UE).

3 METHODS

Here, we introduce our CytoSet model, with the corresponding architecture illustrated in Figure 1. The multi-cell input of k measured cells is randomly drawn with replacement across all of the individual patient samples. This step is necessary because: (1) the original input samples often have different numbers of cells; (2) the number of cells in the original input samples is too large

to fit into a GPU. In addition, the diversity across sampled subsets can also increase the number of cells involved in training and help the model scale to larger datasets. Our experimental results demonstrate that our CytoSet model is robust to the total number of sampled cells, k (see Section 4.2).

Inspired by Zaheer et al. (2017), the network $f_\theta(\cdot)$ contains several permutation equivalent blocks that independently transform the elements in the set. In each block, we introduce a residual connection to stabilize the network training. A pooling operation is then performed within the multi-cell input and thereby maps the set into an embedding vector. In this step, we can either use max or mean pooling. As was previously described in CellCNN, max pooling can measure the presence of cells yielding high response while mean pooling can approximate the frequency of particular cell subsets. Finally, the embedding vector is connected to a feed forward network (fully connected layers) to predict the clinical outcome y, which refers to the classification of the sample (e.g. healthy or sick).

4 Experimental Results

4.1 HIV Exposure Dataset

We tested CytoSet along with the baseline methods CellCNN and CytoDX on the HEUvsUE flow cytometry dataset [1], which was first introduced in the FlowCAP benchmarking challenge (Aghaeepour et al., 2013). The HEUvsUE dataset consists of 308 blood samples from African infants who were either exposed to HIV in *utero* but remained uninfected (HEU) or who were unexposed (UE). The features measured for each single cell in the HEUvsUE are comprised of 10 measured proteins. The raw measurements were transformed using an `arcsinh` transformation (Azad et al., 2016) $f(x) = \text{arcsinh}(x/5)$ before training. From the results previously reported in the FlowCap Challenge (Aghaeepour et al., 2013), the accurate prediction of patient clinical outcomes (HEU vs UE) was shown to be quite challenging in the HEUvsUE dataset.

4.2 Results

We randomly chose 80% of the samples from the HEUvsUE dataset for training and validation and left out the remaining 20% for testing. We used both the classification accuracy (ACC) and the area under the receiver operator curve (AUC) to quantify the performance quality. The test classification results are shown in Table 1. Varying k, or the number of cells sampled from each patient's set of cells, CytoSet consistently outperforms CellCNN and CytoDX by a large margin. Furthermore, our model also performs well even for smaller values of k, such as $k = 1024$. The promising performance of CytoSet on the HEUvsUE dataset suggests great potential to think about clinical flow and mass cytometry data as sets.

Model	$k = 1024$		$k = 2048$		$k = 4096$	
	ACC	AUC	ACC	AUC	ACC	AUC
CellCNN (Arvaniti & Claassen, 2017)	0.630	0.778	0.674	0.763	0.689	0.785
CytoDx (Hu et al., 2019)	0.587	0.587	0.600	0.588	0.597	0.590
CytoSet (Ours)	**0.858**	**0.936**	**0.842**	**0.935**	**0.881**	**0.933**

Table 1: The testing ACC and AUC with different size of subsets on the HEUvsUE dataset. The number reported is averaged over 5 different runs.

5 Conclusion

In this paper, we propose a new deep learning model called CytoSet for predicting clinical outcomes from single-cell flow and mass cytometry data. Considering the set-structure of these data, CytoSet uses a stackable permutation invariant network architecture to build the classification model. Our preliminary results show that CytoSet greatly outperforms other baselines and is also robust to the

[1] `https://flowrepository.org/id/FR-FCM-ZZZU`

k, or the number of subsampled cells. In future work, we will apply CytoSet to other clinical flow and mass cytometry datasets, and perform ablation studies of different model architectures.

REFERENCES

Nima Aghaeepour, Greg Finak, Holger Hoos, Tim R Mosmann, Ryan Brinkman, Raphael Gottardo, and Richard H Scheuermann. Critical assessment of automated flow cytometry data analysis techniques. *Nature methods*, 10(3):228–238, 2013.

Eirini Arvaniti and Manfred Claassen. Sensitive detection of rare disease-associated cell subsets via representation learning. *Nature communications*, 8(1):1–10, 2017.

Ariful Azad, Bartek Rajwa, and Alex Pothen. flowvs: channel-specific variance stabilization in flow cytometry. *BMC bioinformatics*, 17(1):1–14, 2016.

Robert V Bruggner, Bernd Bodenmiller, David L Dill, Robert J Tibshirani, and Garry P Nolan. Automated identification of stratifying signatures in cellular subpopulations. *Proceedings of the National Academy of Sciences*, 111(26):E2770–E2777, 2014.

Mark M Davis, Cristina M Tato, and David Furman. Systems immunology: just getting started. *Nature immunology*, 18(7):725, 2017.

Harrison Edwards and Amos Storkey. Towards a neural statistician. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.

Greg Finak, Marc Langweiler, Maria Jaimes, Mehrnoush Malek, Jafar Taghiyar, Yael Korin, Khadir Raddassi, Lesley Devine, Gerlinde Obermoser, Marcin L Pekalski, et al. Standardizing flow cytometry immunophenotyping analysis from the human immunophenotyping consortium. *Scientific reports*, 6(1):1–11, 2016.

Xiaoyuan Han, Mohammad S Ghaemi, Kazuo Ando, Laura S Peterson, Edward A Ganio, Amy S Tsai, Dyani K Gaudilliere, Ina A Stelzer, Jakob Einhaus, Basile Bertrand, et al. Differential dynamics of the maternal immune system in healthy pregnancy and preeclampsia. *Frontiers in immunology*, 10:1305, 2019.

Felix J Hartmann, Joel Babdor, Pier Federico Gherardini, El-Ad D Amir, Kyle Jones, Bita Sahaf, Diana M Marquez, Peter Krutzik, Erika O'Donnell, Natalia Sigal, et al. Comprehensive immune monitoring of clinical trials to advance human immunotherapy. *Cell reports*, 28(3):819–831, 2019.

Zicheng Hu, Benjamin S Glicksberg, and Atul J Butte. Robust prediction of clinical outcomes using cytometry data. *Bioinformatics*, 35(7):1197–1203, 2019.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International Conference on Machine Learning*, pp. 3744–3753. PMLR, 2019.

Aaron TL Lun, Arianne C Richard, and John C Marioni. Testing for differential abundance in mass cytometry data. *Nature methods*, 14(7):707, 2017.

Matthew H Spitzer and Garry P Nolan. Mass cytometry: single cells, many features. *Cell*, 165(4): 780–791, 2016.

Natalie Stanley, Ina A Stelzer, Amy S Tsai, Ramin Fallahzadeh, Edward Ganio, Martin Becker, Thanaphong Phongpreecha, Huda Nassar, Sajjad Ghaemi, Ivana Maric, et al. Vopo leverages cellular heterogeneity for predictive modeling of single-cell data. *Nature communications*, 11(1): 1–9, 2020.

Lukas M Weber, Malgorzata Nowicka, Charlotte Soneson, and Mark D Robinson. diffcyt: Differential discovery in high-dimensional cytometry via high-resolution clustering. *Communications biology*, 2(1):1–11, 2019.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep sets. In *Advances in Neural Information Processing Systems*, 2017.