Interpretable Hierarchical Bayesian Modeling of Cell-Type Distributions in COVID-19 Disease

Sarah Parsons, Nathan P. Whitener, Sapan Bhandari and Natalia Khuri*
Department of Computer Science
Wake Forest University, Winston-Salem, North Carolina, 27109
* Corresponding author: natalia.khuri@wfu.edu

2022 56th Annual Conference on Information Sciences and Systems (CISS) | 978-1-6654-1796-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/CISS53076.2022.9751177

***Abstract*—High-throughput sequencing of ribonucleic acid molecules is used increasingly to understand gene expression in organs, tissues, and therapies, at a single-cell level. To facilitate the discovery of the heterogeneity and cell-specific factors of the COVID-19 disease, we use an interpretable computational approach that derives cell mixtures from peripheral blood mononuclear cells of healthy donors, and influenza, asymptomatic, mild and severe COVID-19 patients. Cell mixtures are generated using hierarchical Bayesian modeling and are subsequently used as features in the gradient boosting tree classifier. Balanced accuracy of five-fold cross-validation was 68%, significantly higher than expected by random chance. Moreover, 11 out of 19 donors' samples were classified accurately. The main advantage of the mixture-based approach compared to the traditional feature-based classification, is its ability to capture associations between genes as well as between cells.**

***Index Terms*—COVID-19, extreme gradient boosting tree, hierarchical Bayesian modeling, single-cell gene expression**

I. Introduction

Cells are the basic building blocks of all organisms and, due to the technological advances, they can now be studied individually. Single-cell sequencing of ribonucleic acid (RNA) molecules (scRNA-seq), for example, has been used to examine gene expression in single cells from cancer [1], kidney [2], brain [3] and autoimmune [4] diseases, as well as the coronavirus disease 2019 (COVID-19) [5]. These studies provided insights into cell-specific changes that occur in different tissues, patients and conditions, as well as into the heterogeneity and stochasticity of gene expression. The majority of these studies are descriptive, that is they mainly aim to elucidate the diversity of cells' populations within different samples. However, the overarching aim of scRNA-seq is to be able to predict patient's disease and response to treatments. To create predictive models, supervised machine learning (ML) may be used to automatically discover the relationships between gene expression and types of individual cells. Then, an entire sample of individual cells from a donor may be classified into healthy or diseased, based on the relative proportions of the diverse cell types, for example.

Data sets of scRNA-seq experiments have several challenging characteristics, such as zero-valued inflation, high dimensionality and sample biases. To overcome these challenges, extensive data preprocessing is used to normalize, scale, reduce and transform gene expression into low-dimensional representation [6]. However, such transformations render the data non-interpretable, because the expressions of the individual genes are lost during these manipulations. Therefore, to address the loss of interpretability resulting from data preprocessing, we represent cells as cell-type mixtures, derived from the distributions of gene expressions within them. We show that these cell mixtures can be used as features in the classification of peripheral blood mononuclear cells (PBMCs). Furthermore, we demonstrate that classifiers trained with cell mixtures and with the standard cell embeddings, perform similarly in cross-validation and in sample-out validation experiments. However, cell mixtures preserve the information about gene expressions within them, allowing for ease of interpretation and fine-grained analyses.

II. Prior and Relevant Work

Here, we focus on COVID-19 disease and its prediction from the scRNA-seq data. As the primary data source, we use the immunophenotyping study of healthy donors, influenza, asymptomatic, mild and severe COVID-19 patients [5]. In prior work, cluster analysis was used to partition scRNA-seq data into 22 clusters, followed by the manual annotation of marker genes, differentially expressed in each cluster compared to all other clusters, and 13 known immune cell-types were derived. To identify immune genes correlating with the disease status, different visualizations were constructed and analyzed. Other scRNA-seq studies that followed, used the same workflow to analyze their data.

Here, we extend generative mixture modeling into a five-class predictor to identify the healthy, influenza, asymptomatic, mild, and severe COVID-19 samples. As the backbone learning algorithm, we adopt the decision tree induction, which builds a tree-like graph structure with a root node having no incoming edges. All other nodes have exactly one incoming edge. A node with outgoing edges is referred to as a test node, and the leaves of the tree are known as the decision nodes. In the decision tree methodology, the classification process is modeled using a set of hierarchical decisions about the features. These decisions are known as the split criterion, a test of one or more features in the training data, which divides the data into two or more parts. Finally, each decision node is assigned to one class, depending on the response of the tests at the internal nodes [7].

Among different decision tree classifiers, gradient boosting trees have been shown to achieve high performance in numerous application domains. Gradient boosting is the process

of combining weak-learning decision trees into an optimal model [8]. A weak decision tree is the one that predicts class labels slightly better than random chance, and thus, it is not useful by itself. However, when weak learners are combined, the model improves its performance. Moreover, the final model may include features that have been ignored in one of the weak-learning decision trees, also resulting in higher classification performance [9]. To make ensemble learning of decision trees more efficient and flexible, a parallel tree boosting system, called XGBoost, was designed and ported to a variety of programming languages and platforms.

Among the many advantages of XGBoost, is its insensitivity to data normalization and built-in estimation of feature importance [10]. While boosting trees can be trained using high-dimensional data sets, for efficiency, feature selection and reduction are typically performed first. In scRNA-seq, preprocessing consists of the selection of 700 to 2000 highly variable genes, followed by the principal component analysis and the harmonization of principal components to eliminate experimental biases [6]. At the end, 2 to 20 harmonized features are derived from over 10,000 original genes, and these features are used in clustering and classification tasks. Unfortunately, such data manipulations lead to the loss of interpretability in the predictive models because gene expression values are encoded into low-dimensional harmonized embedding vectors.

In this work, we make two contributions. Firstly, we show the utility of ML in the classification of PBMC samples into multiple conditions. Secondly, we show that XGBoost classifiers can be successfully trained using cell-type mixtures derived via generative hierarchical modeling. Rather than masking original information about gene expression, our method automatically extracts and highlights important gene expression patterns.

III. Data and Methods

A. Data Acquisition and Preprocessing

We used publicly available scRNA-seq data from a recent immunophenotyping study [5]. In the study, PBMCs were extracted from healthy donors, and from patients with severe influenza and COVID-19 disease. Single cells were isolated from each sample and their active RNA molecules were sequenced.

We downloaded three files from the Gene Expression Omnibus (GEO) at the National Center for Biotechnology Information, using accession number GSE149689. These files contained cell identification, gene names, and transcript counts, respectively. We used the Seurat package [6] to combine these three files into one matrix comprising transcript counts. Next, we filtered out all genes that were detected in fewer than 5% or greater than 95% of the cells. Using GEO information, we annotated cells with their donor information, such as age, sex, disease type and status. We also created a categorical label for each cell, denoting one of five conditions. Specifically, cells of healthy donors were assigned a categorical label 0, influenza cells were assigned 1, and asymptomatic, mild and severe COVID-19 cells were assigned labels 2, 3, and 4, respectively.

For comparison, we also executed the standard Seurat preprocessing of the count matrix [6]. Specifically, the count matrix was log-transformed, and 720 highly variable genes were selected. Next, principal component analysis and data harmonization were performed to create harmonized cell embeddings of size 20.

B. Hierarchical Bayesian Model

To represent each cell as a cell-type mixture, we applied a three-level hierarchical Bayesian modeling. Specifically, we modeled each cell as a mixture of distinct populations or clusters of identifiable cell-types (n=20), and each cluster was modeled as an infinite mixture over an underlying set of cluster probabilities, characterized by the distribution of transcript counts within them.

Assuming that there are K latent clusters within the data, V unique genes and D cells in the data set, we first assign a gene distribution to each cell, denoted as β_k for each cluster k. Here, each β_{kw} is the likelihood that gene w will be expressed in cluster k. Next, a cluster distribution, θ_d, will be generated for each cell d, where θ_{dk} is the likelihood that cell d will be assigned to cluster k. At last, we assign each gene w in cell d to cluster k with a probability θ_{dk}, and draw gene w' from distribution β_k. The generative modeling process is summarized as follows.

1) Draw gene distributions $\beta_k \sim$ Dirichlet(η).
2) Draw cluster distributions $\theta_d \sim$ Dirichlet(α).
3) For each gene w in cell d:
    a) Draw cluster assignment $z_{dn} \sim$ Cat(θ_d).
    b) Draw gene $w_{dn} \sim$ Cat($\beta_{z_{dn}}$).

Cat($\cdot$) represents a categorical distribution, η and α are two hyper-parameters of Dirichlet distributions, where η has a length of V, and α has a length of K. The generative process described above is known as the Latent Dirichlet Allocation, a popular method for identifying latent topics within a collection of texts [11]. Unstructured text data has characteristics similar to scRNA-seq data, such as data sparsity and high dimensionality.

C. Experimental Design

The practical task of our problem domain is to develop a system for the classification of scRNA-seq samples into different classes. Given this task and the prior knowledge in the form of paired observations of cells' features and cell types, a classifier can be trained using XGBoost. We used categorical accuracy in the assessment of XGBoost's learning performance. Specifically, we encoded categorical class labels (healthy, influenza, asymptomatic, mild and severe COVID-19) into one-hot vectors and calculated the percentage of predicted labels that matched with the actual values of the encoded labels.

We compared our mixture-based classifier with a classifier trained with harmonized cell embeddings, and performed two computational experiments. In the first experiment, we

estimated the accuracy of each classifier using five-fold cross-validation. In five-fold cross-validation, we divided the data into k partitions of the same size [12]. For each partition, we trained the model on the $k-1$ partitions, and evaluated the model on the ith partition. Then, the average of the k performance scores and their standard deviations were computed. Cross-validation experiments are beneficial for determining if the model's performance had notable variance depending on the data splits. For our experiments, we have chosen the value of k to be 5, such that each training and validation set were large enough to be representative of the entire data set.

To estimate the performance of the trained classifiers, we computed two metrics, namely, normalized mutual information (NMI) and balanced accuracy adjusted for chance. To compute NMI, let A = $\{a_1, a_2, a_3...a_k\}$ and B = $\{b_1, b_2, b_3...b_k\}$ designate the ground truth labels and the cluster labels across k classes, respectively. In our application, k is 5. Then, NMI can be calculated as follows: $NMI = (2 \times I(A,B))/(H(A)+H(B))$, where $I(A,B)$ is the mutual information of A and B, computed as $I(A,B) = \sum_{i=1}^{K}\sum_{j=1}^{K}(|a_i \cap b_j|/N) \times (log(N \times (|a_i| \cap |b_j|))/(|a_i| \times |b_j|))$, and H(A) and H(B) are the entropy of partitions A and B: $H(A) = -\sum_{i=1}^{K}(a_i/N) \times log(a_i/N)$, $H(B) = -\sum_{i=1}^{K}(b_i/N) \times log(b_i/N)$. N is the total number of cells. NMI is a metric that is independent of the absolute values of the labels, that is a permutation of the labels in A and B does not change the score. NMI scores range between 0 and 1, and higher scores indicate better performance.

Because our data set was imbalanced in class label distribution, we also used balanced accuracy, which avoids inflated performance estimates. This metric weighs raw accuracy according to the inverse prevalence of true class labels. Specifically, given the predicted class label, $\hat{y_i}$ for sample i, we compute balanced accuracy as $\frac{I}{\sum \hat{w_i}} \sum 1(\hat{y_i} == y_i)\hat{w_i}$, where $I(x)$ is an indicator function. The range of balanced accuracy scores is $\frac{1}{1-num_{classes}}$ to 1. When adjusted for chance, balanced accuracy reports relative increase from $\frac{1}{num_{classes}}$, and the higher the score of the balanced accuracy, the better the performance.

In the second experiment, we studied the performance of each classifier using sample-out validation. Specifically, we removed all cells belonging to one donor, trained a classifier using remaining cells and predicted cell types of the withheld donor's sample. To assign a single class label to each withheld sample, we find the cell-type with the largest proportion and use it as a sample label. This process was repeated for each donor in our data set.

IV. Results

A. Cell Type Classification

We trained two gradient boosting tree classifiers using the XGBoost package [10]. The first (proposed) classifier used as features, cell-type mixtures from our hierarchical Bayesian modeling and the second classifier used harmonized cell embeddings computed by the Seurat package [6].

In the five-fold cross-validation experiments, performance was similar. The first classifier, built using cell-type distributions, achieved the average NMI score of 0.4813 ± 0.0046 and the average balanced accuracy of 0.6834 ± 0.0040. The NMI score of the second classifier was 0.4788 ± 0.0075 and its balanced accuracy score was 0.6926 ± 0.0067, respectively.

Next, we examined cross-validated performance for each class separately (Table I). The proposed classifier had high accuracy in predicting healthy (0.8148), influenza (0.8144) and mild COVID-19 (0.7838). However, classification of severe COVID-19 cells was less accurate, with a balanced accuracy of 0.7320, and even lower balanced accuracy of 0.5886 was seen for asymptomatic cells.

The standard classifier had a similar performance, except for the influenza and asymptomatic COVID-19 samples. Balanced accuracy of classifying healthy, mild and severe COVID-19 samples differed from the proposed classifier, by 1% to 2%. However, approximately 5% decrease was noted in the classification of influenza cells and almost 8% increase in the classification of asymptomatic COVID-19 samples, when harmonized cell embeddings were used to train XGBoost.

TABLE I
PER-CLASS CLASSIFICATION ACCURACY. FOR EACH CELL TYPE, SHOWN ARE THE NUMBER OF CELLS AND THE NUMBER OF DONORS, CLASSIFICATION ACCURACY OF THE PROPOSED AND STANDARD APPROACHES, AS WELL AS THE DIFFERENCE BETWEEN THE PROPOSED AND STANDARD CLASSIFICATION ACCURACY.

Label	Cells	Donors	Proposed	Standard	Difference
Healthy	16,147	4	0.8148	0.8288	0.0140
Influenza	9,054	5	0.8144	0.7597	-0.0548
Asymptomatic	3,505	1	0.5886	0.6719	0.0833
Mild COVID-19	14,772	5	0.7838	0.8006	0.0168
Severe COVID-19	6,742	5	0.7320	0.7078	-0.0242

Similarly, misclassification patterns did not differ between the two classifiers (Table II). Notably, about 14% of healthy cells were misclassified as mild COVID-19, and a similar proportion of mild COVID-19 cells were misclassified as healthy cells. About 7% of the severe COVID-19 cells were assigned mild status, and a similar percentage of mild cells was assigned to the severe COVID-19 class. Misclassified influenza and asymptomatic cells were distributed equally across all other classes.

In the sample-out validation experiments, we repeatedly removed all cells of a single donor, trained classifiers using remaining cells and predicted class labels of the withheld cells. Majority rule was then applied to assign a unique class label to the entire sample. Overall, both classifiers predicted correctly 11 out of 19 samples (Fig. 1).

Standard classifier correctly identified all healthy samples (3, 9, 10, and 12), while our proposed classifier incorrectly assigned influenza status to sample 10. All mild COVID-19 samples (1, 2, 6, and 19) were correctly predicted by the standard classifier. Our proposed approach assigned erroneously a healthy status to sample 6, while correctly classifying the remaining three samples. Both classifiers missed asymptomatic patient 7 and labeled this sample as mild COVID-19. There

TABLE II
CONFUSION MATRICES OF THE PROPOSED AND THE STANDARD XGBOOST CLASSIFIERS. PROPORTION OF CELLS PREDICTED FOR EACH CLASS ARE SHOWN IN ROWS. COLUMNS DENOTE TRUE CLASS LABELS.

	Healthy		Influenza		Asymptomatic COVID-19		Mild COVID-19		Severe COVID-19	
Labels	Proposed	Standard	Proposed	Standard	Proposed	Standard	Proposed	Standard	Proposed	Standard
Healthy	0.8148	0.8288	0.0577	0.0585	0.075	0.0696	0.1325	0.1147	0.0366	0.0438
Influenza	0.0748	0.0945	0.8144	0.7597	0.0234	0.0108	0.0171	0.0296	0.0206	0.026
Asymptomatic COVID-19	0.0355	0.0237	0.0093	0.0041	0.5886	0.6719	0.0424	0.0379	0.0233	0.0252
Mild COVID-19	0.1393	0.1213	0.0179	0.0205	0.085	0.0856	0.7838	0.8006	0.0719	0.0743
Severe COVID-19	0.0307	0.0403	0.0179	0.0221	0.0534	0.0676	0.0652	0.0598	0.732	0.7078

were several disagreements in the prediction of severe COVID-19 samples (5, 8, 11, 15, and 16). Surprisingly, both classifiers labeled sample 5 as asymptomatic, and the standard classifier mislabeled sample 8 as a mild COVID-19 sample. Samples 15 and 16 were correctly labeled by both classifiers. Finally, influenza samples proved to be challenging to classify as well. Samples 4, 17 and 18 were predicted as healthy by both classifiers, and sample 14 was incorrectly labeled by the standard classifier.

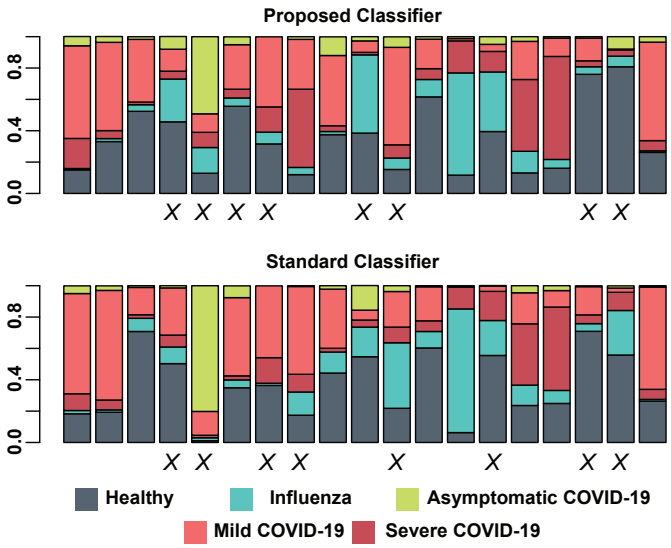

Fig. 1. Results of sample-out validation experiments. Shown are stacked bar plots depicting predicted cell type within each sample. Proportions of the predicted cell types are shown on the y-axis and each sample is represented by a bar. X symbols denote misclassified samples.

In summary, both classifiers had high NMI scores and balanced accuracy greater than the chance performance of 0.20. However, standard encoding of the data in the form of harmonized embeddings provided no interpretable insights. On the other hand, our proposed approach is amiable to in-depth analyses post-classification, as shown next.

B. Analysis of Cell Mixtures

For brevity, we refer to cell-type mixtures as clusters. Cell clusters comprised unique distributions of cells (Fig. 2). For example, clusters 6, 16, 18, 19 and 20 comprised mostly cells from healthy samples, and clusters 8, 11 and 17 had influenza cells. Cells from mild COVID-19 were overrepresented in clusters 7, 10 and 13, while severe COVID-19 cells were found in clusters 4, 5, 14, and 15. Finally, asymptomatic COVID-19 cells were detected in clusters 1, 13 and 15, and they were completely absent from several clusters, such as cluster 8, 12, 14, and 18, for example.

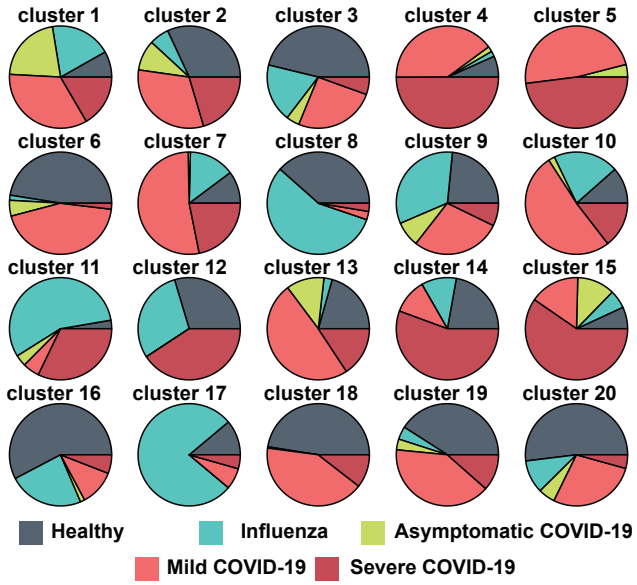

Fig. 2. Composition of cell clusters. Shown are pie charts for each cluster (cell-type mixture) comprised cells from healthy, influenza, asymptomatic, mild and severe COVID-19 samples.

Next, we examined distributions of cell clusters in different conditions. Specifically, for each cluster of cells, we computed their representation in healthy, influenza, and asymptomatic, mild and severe COVID-19 samples (Fig. 3). Several patterns emerged. First, influenza cluster distribution differed from distributions of other conditions. Cluster distributions were similar between healthy and mild COVID-19 samples, as well as between severe and asymptomatic COVID-19 samples, respectively. Healthy samples comprised cells from all clusters, with the exception of cluster 5. Cluster 5 was also not detected in influenza nor asymptomatic samples. However, cells from cluster 5 were abundant in mild and severe COVID-19 samples. Similarly, cells from clusters 4 were found in small quantities in healthy, influenza, and asymptomatic samples, and in greater proportion in mild and severe COVID-19 samples. Cells from cluster 1, 11 and 15 were also underrepresented in healthy samples. On the other hand, cells from cluster 3, 6, 18, 19 and 20, were abundant

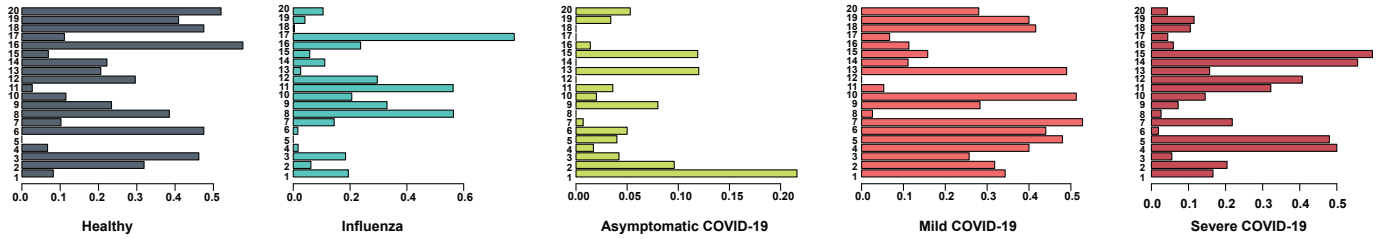

Fig. 3. Distribution of clusters in cell samples. Shown are bar plots of the proportions of cells in each cluster, detected in healthy, influenza, asymptomatic, mild and severe COVID-19 samples. Cell-type mixture or cluster identification are on the y-axes and their proportions are on the x-axes.

in healthy and mild COVID-19 samples, yet were barely detectable or absent in all other samples.

As mentioned, influenza samples had distinct cluster distributions. About 70% of cluster 17 cells were found among influenza samples, followed by over 50% of cells from clusters 8 and 11, and in a much lesser proportion, from clusters 9 and 12.

Asymptomatic cells were all derived from a single donor and, hence, had a very small proportion in each cluster. The most dominant cluster was cluster 1, also found in similar abundance in mild COVID-19 patients, followed by clusters 13 and 15. Clusters 8, 12, 14, 17 and 18 were not detected among asymptomatic cells.

Although similar to healthy samples, yet distinct from severe COVID-19 samples, mild COVID-19 cells were represented by over 50% of cells from clusters 7 and 10, and over 40% of cells from clusters 4, 5, 6, 13, 18 and 19. Similar to the asymptomatic sample, cluster 12 was completely absent in mild COVID-19 samples, and clusters 8, 11, and 17 were found in small quantities.

Interestingly, over 40% of cluster 12 cells, which were absent among asymptomatic and mild COVID-19 samples and detected in moderate quantities, were found in severe COVID-19 samples. Moreover, clusters 15 and 16 were dominant in severe COVID-19 and underrepresented in all other samples.

To better understand the contribution of each cluster to sample type, we computed proportions of each cluster in samples of individual donors and performed a two-way clustering of donors and cell clusters (Fig. 4). Interestingly, most of the clusters were present in all donors' samples. We hypothesize that cell types in these clusters may not carry information important for the prediction of the disease. On the other hand, inter-donor differences were observed in 9 out of 20 clusters. These differences led to four distinct groupings of donors.

The first group encompasses 3 influenza and 1 severe COVID-19 donors (donors 13, 14, 15, and 18), characterized by a large proportion of cells from clusters 1, 3, 9 and 11. The second group contains cells from three donors with mild COVID-19, two influenza and one healthy donor. These cells come from clusters 1, 2, 3, and 19. Three severe COVID-19 and one asymptomatic donor form the third group of donor cells. Finally, the last group contains mixed cell populations, with cells of healthy, mild and severe COVID-19 donors. Notably, two outlier donors were observed in the first (influenza donor 18) and fourth (severe COVID-19 donor 15) groups, each possessing donor-specific distributions of cells.

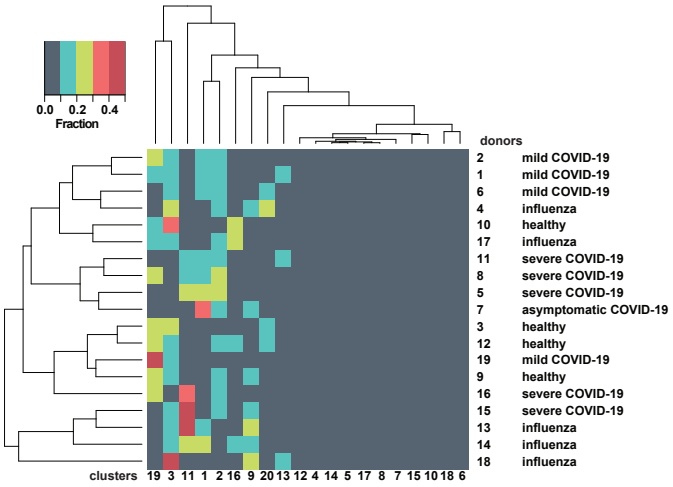

Fig. 4. Two-way clustering of cells. Shown is a heatmap of two-way clustering of cells of 19 donors and cells of 20 mixtures or clusters. Donor-wise (rows) and mixture-wise (columns) dendrograms are drawn. Color scale represents proportions of each cluster in donors' samples, with gray and red colors denoting low and high proportions, respectively.

C. Interpretation of Cell Mixtures

To demonstrate a more advanced analysis of interpretable insights from the data set, we computed XGBoost feature gains of cell mixtures.

Consider for example, cluster 4, comprising 60 cells derived mostly from mild and severe COVID-19 donors (Fig. 5). When used as a feature in the XGBoost classifier, cell mixture of cluster 4 had the largest average gain in the five-fold cross-validation experiments. We examined the top 30 genes deemed most important to the cell mixture of cluster 4. Among these genes, were thyroid peroxidase (TPO), prostaglandin D2 receptor 2 (PTGDR2), and TNFRSF4, for example. TPO is an enzyme that plays an important role in the production of thyroid hormones, and thyroid dysfunction has been reported in around 15% of patients with mild to moderate COVID-19 [13]. PTGDR2 is preferentially expressed in T helper cells, where it mediates the pro-inflamatory response and lymphopenia in COVID-19 disease, and predicts disease morbidity and mortality [14]. Finally, TNFRSF4 is a known

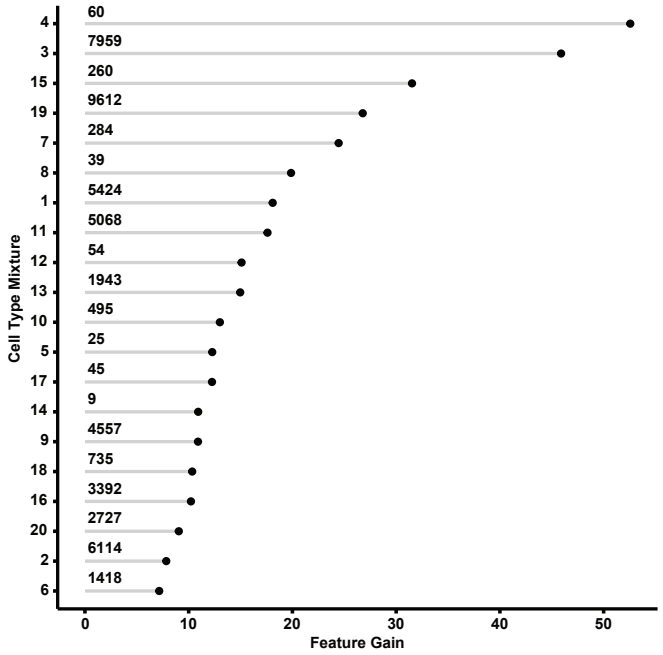

Fig. 5. Feature importance of cell type mixtures expressed as feature gain. Cell type mixture or cluster identifications are shown on the y-axis and gain scores on the x-axis, respectively. The number of cells within each cluster is listed.

immunoregulatory gene, involved in immunoresponse during COVID-19 disease [15].

Similar analyses can be performed for the remaining 20 clusters, resulting in an additional evidence for prior reports about COVID-19 disease or in new insights.

The main limitation of our approach is that, similarly to cluster analysis, the number of mixtures must be determined by the user. Although some metrics, such as perplexity and coherence, had been proposed to find the number of mixtures automatically, they fail to guide the modeling process. For instance, we observed that as the number of mixtures increases, they become less populated (contain fewer cells).

V. Conclusion

Analysis of gene expression in single cells has become an important tool for the study of cell-specific changes that occur in different patients, disorders and diseases. Current data transformations, such as harmonized cell embeddings, may result in noninterpretable models. In this work, cell-type assignment is made using hierarchical Bayesian modeling. These cell-type mixtures are then used as features in the gradient boosting tree classifier that achieved, in five-fold cross-validation, a normalized mutual information of 48% and a balanced accuracy of 68%. In sample-out classification, our approach predicted correctly 11 out of 19 single cell samples. These results are on par with the baseline gradient boosting tree classifier that uses harmonized cell embeddings as features. The main advantage of our approach is its interpretability and transparency, and its application is not limited to PBMC data. Future work will focus on designing and implementing a user interface to support the fine-grained examination of genes expressed in each cell mixture.

Acknowledgment

This research was partially supported by the Pilot Grant from Wake Forest Center for Biomedical Informatics. The authors thank Tian (Simon) Yun and Cody Stevens for technical assistance and acknowledge the Distributed Environment for Academic Computing (DEAC) at Wake Forest University for providing HPC resources that have contributed to the research results reported within this paper. URL: https://is.wfu.edu/deac

References

[1] M. Bartoschek, N. Oskolkov, M. Bocci, J. Lövrot, C. Larsson, M. Sommarin, C. D. Madsen, D. Lindgren, G. Pekar, G. Karlsson *et al.*, "Spatially and functionally distinct subclasses of breast cancer-associated fibroblasts revealed by single cell RNA sequencing," *Nature Communications*, vol. 9, no. 1, pp. 1–13, 2018.

[2] J. Liao, Z. Yu, Y. Chen, M. Bao, C. Zou, H. Zhang, D. Liu, T. Li, Q. Zhang, J. Li *et al.*, "Single-cell RNA sequencing of human kidney," *Scientific Data*, vol. 7, no. 1, pp. 1–9, 2020.

[3] D. Ofengeim, N. Giagtzoglou, D. Huh, C. Zou, and J. Yuan, "Single-cell RNA sequencing: unraveling the brain one cell at a time," *Trends in Molecular Medicine*, vol. 23, no. 6, pp. 563–576, 2017.

[4] M. Zhao, J. Jiang, M. Zhao, C. Chang, H. Wu, and Q. Lu, "The application of single-cell RNA sequencing in studies of autoimmune diseases: a comprehensive review," *Clinical Reviews in Allergy & Immunology*, pp. 1–19, 2020.

[5] J. S. Lee, S. Park, H. W. Jeong, J. Y. Ahn, S. J. Choi, H. Lee, B. Choi, S. K. Nam, M. Sa, J.-S. Kwon, S. J. Jeong, H. K. Lee, S. H. Park, S.-H. Park, J. Y. Choi, S.-H. Kim, I. Jung, and E.-C. Shin, "Immunophenotyping of COVID-19 and influenza highlights the role of type I interferons in development of severe COVID-19," *Science Immunology*, vol. 5, no. 49, 2020.

[6] T. Stuart, A. Butler, P. Hoffman, C. Hafemeister, E. Papalexi, W. M. M. III, Y. Hao, M. Stoeckius, P. Smibert, and R. Satija, "Comprehensive integration of single-cell data," *Cell*, vol. 177, pp. 1888–1902, 2019.

[7] V. A. Dev and M. R. Eden, "Formation lithology classification using scalable gradient boosted decision trees," *Computers & Chemical Engineering*, vol. 128, pp. 392–404, 2019.

[8] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of Statistics*, pp. 1189–1232, 2001.

[9] Z. Zhang, Y. Zhao, A. Canes, D. Steinberg, O. Lyashevska *et al.*, "Predictive analytics with gradient boosting in clinical medicine," *Annals of Translational Medicine*, vol. 7, no. 7, 2019.

[10] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[11] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[12] D. Berrar, "Cross-validation," *Encyclopedia of bioinformatics and computational biology*, vol. 1, pp. 542–545, 2019.

[13] D. T. W. Lui, C. H. Lee, W. S. Chow, A. C. H. Lee, A. R. Tam, C. H. Y. Fong, C. Y. Law, E. K. H. Leung, K. K. W. To, K. C. B. Tan *et al.*, "Thyroid dysfunction in relation to immune profile, disease status, and outcome in 191 patients with COVID-19," *The Journal of Clinical Endocrinology & Metabolism*, vol. 106, no. 2, pp. e926–e935, 2021.

[14] F. Zhou, T. Yu, R. Du, G. Fan, Y. Liu, Z. Liu, J. Xiang, Y. Wang, B. Song, X. Gu *et al.*, "Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study," *The Lancet*, vol. 395, no. 10229, pp. 1054–1062, 2020.

[15] B. Kalfaoglu, J. Almeida-Santos, C. A. Tye, Y. Satou, and M. Ono, "T-cell hyperactivation and paralysis in severe COVID-19 infection revealed by single-cell analysis," *Frontiers in Immunology*, vol. 11, p. 2605, 2020.