Automated Cell Phenotyping for Imaging Mass Cytometry

Sindhura Thirumal[1], Amoon Jamzad[1], Tiziana Cotechini[2], D. Robert Siemens[3] and Parvin Mousavi[1]

***Abstract*— Imaging mass cytometry (IMC) is a new advancement in tissue imaging that is quickly gaining wider usage since its recent launch. It improves upon current tissue imaging methods by allowing for a significantly higher number of proteins to be imaged at once on a single tissue slide. For most analyses of IMC data, determining the phenotype of each cell is a crucial step. Current methods of phenotyping require sufficient biological knowledge regarding the protein expression profile of the various cell types. Here, we develop a deep convolutional autoencoder-classifier to automate the cell phenotyping process into four basic cell types. Biopsy tissue from bladder cancer patients is used to evaluate the efficacy of the classification. The model is evaluated and validated through feature importance, confirming that the significant features are biologically relevant. Our results demonstrate the potential of deep learning to automate the task of cell phenotyping for high-dimensional IMC data.**

I. INTRODUCTION

With recent technological advances in tissue imaging, biologists can now obtain an in-depth understanding of the *in situ* organization of tissues at the cellular level. Traditional immunodetection techniques in microscopy, such as immunofluorescence or immunohistochemistry, have limitations due to spectral overlap of fluorophores and few colours of chromagens, respectively. To obtain highly multiplexed imaging using these methods, staining and imaging must be repeated iteratively, leading to extensive acquisition times.

Imaging mass cytometry (IMC) using the Hyperion Imaging System (Fluidigm, Markham) is an emerging technology that improves upon traditional methods by using mass measurements rather than emission spectra [1]. IMC enables detection of 37 protein signals at a time and provides a high-throughput visualization of cells and proteins *in situ* [2]. This is a compelling improvement over traditional methods since it enables the identification of numerous, diverse cell types.

An important step in IMC analysis is phenotyping individual cells, which is the identification of a cell's type based on its protein composition. Traditionally, single cell phenotyping relies on the use of manual, hierarchical gating strategies - an iterative process where the user manually selects cell populations based on expression of proteins using a *channel x* by *channel y* scatter plot. Gating has a major drawback due to the manual nature, resulting in both user bias and difficulty scaling when many channels are involved. In lieu of this, automated approaches using unsupervised learning have become more common for cell phenotyping of IMC data [3]. Although these methods improve upon manual gating, the end-user is still required to have enough biological knowledge to be able to phenotype cells based on the pattern of protein expression of various clusters. Additionally, when there are a large number of tissue samples, it is time-consuming for the user to have to annotate each one separately. Thus, it would be of great benefit to be able to automate the cell annotation process entirely.

Deep learning methods lend themselves well to automated phenotyping of cellular data as IMC is of high dimensionality. Currently, there are very few studies using deep learning for analysis of IMC data, and even less in regard to cell phenotyping specifically. Studies have focused mainly on the application of deep learning to either the cell segmentation process [4] or clinical outcome prediction [5]. Deep learning for cell phenotyping has been proposed in the past for similar imaging methods, but never for IMC data specifically [6].

In this paper, for the first time we apply deep learning to the task of cell phenotyping for IMC data. We develop and test a deep joint autoencoder-classifier model for phenotyping cells into four basic cell types. The model is trained using biopsy tissue collected from patients with bladder cancer. In addition, we identify and evaluate the feature (protein expression) significance for predicting the various phenotypes in order to confirm our model's performance is biologically relevant. With this model, we demonstrate the prospect of automating the cell phenotyping process through the use of deep learning, and interpreting the significant features for biological relevance.

II. MATERIALS & METHOD

An overview of the workflow is outlined in Figure 1. Once the tissues are imaged, the cells are segmented and their protein expression information for each channel marker is obtained. The cells are annotated with their type, to be used as the gold standard label for training. The network is a joint autoencoder-classifier trained on both the reconstruction of the input data as well as prediction of class labels.

A. Data and pre-processing

For this study, we use 30 biopsy samples collected from 15 bladder cancer patients at Kingston Health Sciences Center. Tissue collection was approved by the Queen's University Research Ethics Board. Each tissue sample is stained with 22 protein markers[1], and protein channel data are acquired with

*This work was supported by Bladder Cancer Canada.

[1]School of Computing, Queen's University, Kingston, Canada sindhura.thirumal@queensu.ca

[2]Department of Biomedical and Molecular Sciences, Queen's University, Kingston, Canada

[3]Department of Urology, Queen's University, Kingston, Canada

[1]alpha actin, CD66b, vimentin, CD14, CD163, pan-cytokeratin (pan-CK), CD11b, GATA3, CD45, TIM3, FoxP3, CD4, CD11c, CD68, CD20, CD8, PD-1, granzyme B, Ki67, DC-LAMP, CD3, and HLA-DR

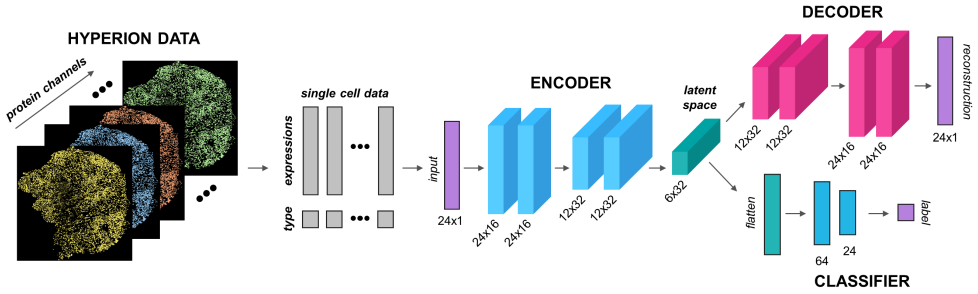

Fig. 1. Overview of convolutional autoencoder architecture. Network is trained on protein expressions of individual cells to classify the cell phenotype.

the Hyperion^{TM}. We generate an image for each channel and segmented the nucleus stains into cell masks for each ROI using Titan, an IMC analysis software [7]. Using the cell masks, we extract the protein expressions of each pixel and calculate the mean intensity values for each individual cell, which represent the overall expression of a given protein for each cell. We then use PhenoGraph [3] to cluster cells with matching protein expressions and used domain knowledge to annotate these clusters with their cell type by human expert. The cells are annotated based on their general type as follows: immune, stromal, tumour, and other. These four cell types are chosen since they represent the majority of cell populations in our tissue samples and will provide a general understanding of the overall spatial distribution of cells within the tumour microenvironment. The resulting single-cell dataset contains 139,093 data samples with 22 features, where each feature represents a protein expression and each data sample is an individual cell. The feature dimension is then increased to 24 by zero-padding for better handling of dimensionality change within the autoencoder network. The data is split into training and testing sets using stratified random sampling based on the cell type. We over-sample the training set using Synthetic Minority Over-sampling Technique (SMOTE) in order to balance the number of data points in each class, and further split it into training and validation sets using the same stratification method. The entire data stratification process is repeated 20 times, each generating different, randomized sets of the data to increase generalization during training.

B. Network structure

The network we developed is a convolutional autoencoder with joint classifier as shown in Figure 1. The convolutional autoencoder is an unsupervised modification of a convolutional neural network (CNN) that is trained to reconstruct its input from an encoded latent space. They are most commonly used for image reconstruction tasks and tailored towards two-dimensional data samples, however we modified this architecture to be compatible with one-dimensional single-cell data. They are unique in that they use multiple layers to train and typically perform with high accuracy for tasks of pattern recognition, which makes it an ideal choice for cell phenotyping [8]. The conventional autoencoder consists of an encoder, which constructs a lower dimensional representation of the original data in its latent space, and a decoder, which reconstructs the input from the latent space. In our structure, the latent space is also passed through fully-connected layers in order to reconstruct the input. The joint training of autoencoder and classifier assures a reconstructable latent space, in which the data representation between class labels are distinguishable. The convolution layers in the encoder all used a Rectified Linear Unit (ReLU) activation function, with the final encoding and decoding layers using sigmoid activation. A dropout layer with a rate of 0.7 is incorporated in the final layer to prevent overfitting. The final parameters of the architecture are determined using an ablation study describe in the next subsection.

C. Experiments

We perform an ablation study to identify the optimum values of structural parameters such as number of layers and convolutional filters, based on the classification accuracy and reconstruction loss in the validation set. To reinforce the training optimization, we first conduct an unsupervised pre-training of the convolutional autoencoder alone, disregarding class label, for 20 epochs to initialize the network weights. The network is then trained jointly for reconstruction loss and classification accuracy. Early stopping based on the validation set loss is used to prevent overfitting or underfitting with patience of 10. Adam optimizer with a learning rate of $1e^{-5}$ for pre-training and $5e^{-5}$ for joint training is used. Mean squared error and sparse categorical cross-entropy are used as autoencoder and classifier loss functions, with weights of 1 and 10, respectively. This weight ratio is used since the autoencoder was pre-trained before joint training and outcome prediction is the main target of this network. The joint-error is back-propagated through the network to update the weights during training. The model is then evaluated on the independent testing set. To increase the generalization of evaluation, the process is repeated for the 20 randomly generated sets, mentioned in Data subsection, and the average of the performance metrics are reported as the performance measure. In addition to the proposed convolutional autoencoder-classifier, a conventional CNN network is trained and evaluated as the baseline for this study, in a same manner. The parameters of the baseline network are also optimized in the ablation study.

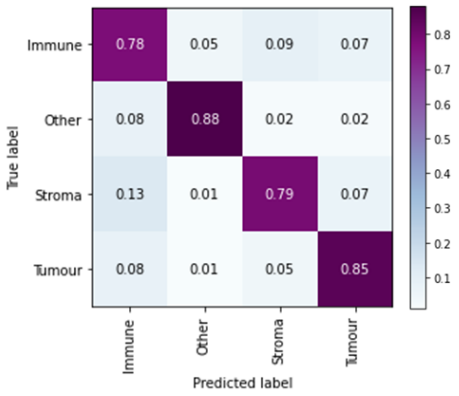

Fig. 2. Confusion matrix of network predictions.

In order to interpret the outcome of the model and its failure modes, and to further evaluate its effectiveness, we analyze the importance of the features for predicting cell types. To achieve this, we utilize SHapley Additive exPlanations (SHAP), a method coined by Lundberg and Lee which scores the impact of each feature for each prediction [9]. Using the scores, we can identify the overall importance of the various features for classifying each label, as well as the important features for prediction of a single data point.

III. RESULTS & DISCUSSION

Table 1 summarizes the classification accuracy results for both the baseline CNN model as well as the final convolutional autoencoder-classifier network. The overall training and test accuracy of the baseline model over 20 random runs is 79.9%±0.7 and 81.5%±0.6 respectively. For the autoencoder-classifier - the optimum model - the testing and training accuracy of cell type classification in 20 random runs is 81.9%±0.3 and 83.4%±0.4 respectively. It can be seen that the proposed autoencoder-classifier outperforms the baseline, and the improvement is statistically significant (p-value$<$0.0001 in one-tailed Wilcoxon Signed-Rank test).

As mentioned above, the optimal structural parameters of both baseline and proposed model are determined in an ablation study. For the baseline, a model with three layers of double convolution kernels (each with max pooling and filter sizes of 16, 32, and 64 respectively), followed by two dense layers of size 64 and 24, and a 4-neuron output layer results in the highest performance. For the autoencoder-classifier model, the optimum encoder consists of two layers of double convolution kernels, each followed by max pooling, as shown in Figure 1. The convolution with kernel size of 3 and filter size of 16 and 32 is used for first and second layer respectively. The decoder has a matching structure that results in a symmetric autoencoder. The classifier network consists of the same two dense layers of size 64 and 24, and the final 4-neuron output layer. Training and testing per set of samples took on average 2.5 hours and 15 minutes respectively, using an Intel® Core™ i5-3570 CPU @ 3.40 GHz processor with 12 GB of RAM. This is an improvement on the traditional method of gating which can take 1-2 hours per sample. The reported performance metrics in Table 1 are calculated for models with these parameters.

TABLE I
PERFORMANCE METRICS OF MODELS

	Autoencoder	**CNN (baseline)**
	ACC (%)±SD	ACC (%)±SD
Testing	81.9 %± 0.3	79.9 ± 0.7
Training	83.4 %± 0.4	81.5 ± 0.7

To further explore inter-class performance, the prediction results of one run are illustrated as a confusion matrix in Figure 2. The confusion matrix depicts the percentage of each label that is classified both correctly and incorrectly. As seen in the matrix, a significant portion of cells are classified accurately, specifically for “tumour” and “other” cell types, indicating high performance of our network. Approximately 13% of stromal cells are misclassified as immune cells, which is higher than the other cell type misclassifications. Biologically, this specific misclassification is not unexpected; vimentin is a protein predominantly expressed by stromal cells, but can also be expressed by macrophages, an immune cell subtype. Thus, there is some difficulty in resolving the two populations. The same applies for the protein GATA3, which can be expressed by tumour cells but can also be expressed by T cells (also an immune cell type) resulting in errors in distinguishing tumor from immune cells. In addition to misclassifications occurring due to expression of non-discrete markers, errors in segmentation may also contribute to the label misclassifications. When we examine this further, we observe that the misclassified cells are generally spatially localized in close proximity to the cell type as which they are misclassified. Figure 3 illustrates an example of this. The cells in this sample are coloured by their true cell type, with an overlay of the stromal cells that were misclassified as immune, shown in red. As seen, the misclassified stromal cells are mostly close in proximity to immune cells, indicating the likelihood of segmentation error contributing to the misclassification.

We investigate the importance of each protein channel in determining the cell types. The results are shown in Figure 4. The channels identified to be important for predicting the various cell types (Figure 4a) conform with what would be expected biologically. The features identified indicate channels where either positive or negative expression are essential to classifying a given cell type. For tumour cell classification, expression of pan-CK and GATA3 are important for recognizing a cell as tumour [10], [11]. As seen in Figure 4a, pan-CK and GATA3 are the top features recognized by the network for determining if a cell is of tumour type. Similarly, the channels identified as highly important for classifying stromal and immune cells are in accordance with biological expectations for their protein expressions.

We also explore the contribution of each feature in the calculation of probabilities of output nodes per each data

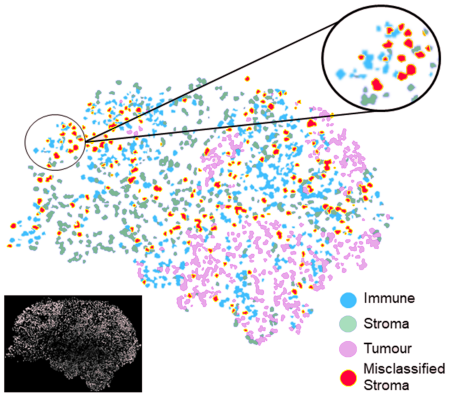

Fig. 3. Visualization of tissue showing prediction misclassification. Cells in tissue are coloured by their true labels. Red coloured cells depict the stromal cells that are misclassified as immune cells, zoomed in selection shows their proximity to immune cells.

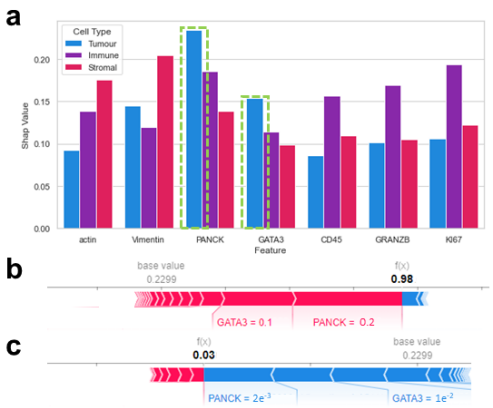

Fig. 4. Feature importance for classifying the different cell types. a) Overall importance ranking of top features for predicting immune, stromal, and tumour cell types. The highlighted bars indicate the top relevant features for predicting a tumour cell: pan-CK and GATA3. b) Importance of features for predicting a tumour cell as tumour versus c) an immune cell as tumour. Each arrow represents a feature that is influencing the model to classify the cell as either tumour (red) or not tumour (blue). The arrow's magnitude indicates its level of importance. As shown, the level of pan-CK and GATA3 expression influences whether a cell is tumour or not.

sample. A sample visualization for the "tumor" output node, for two cells with labels of "tumor" and "immune" is illustrated in Figures 4b and Figure 4c, respectively. The base value represents the mean prediction value for the tumour class, while the bold number indicates the prediction value of tumour class membership for these two data samples. Prediction values higher than the base value are expected if the data belongs to the class of interest (0.98 in Figure 4b for classifying a tumour cell as tumour), while a data sample from a different class results in values lower than the base value (0.03 in Figure 4c as probability of an immune cell belonging to the tumor class). Each arrow in the plot represents a protein feature, and the magnitude of the arrow indicates its level of importance for that prediction. Features in red are those that are indicative of the cell being tumour, while blue features suggest that the cell is not a tumour. For this example, we chose to focus on pan-CK and GATA3 as shown. For the tumour cell that is correctly predicted as such (Figure 4b), it is clear that the high pan-CK (0.2) and GATA3 (0.1) expression affected the prediction of this cell to be tumour. In contrast, the immune cell (Figure 4c) has low pan-CK ($2e^{-3}$) and GATA3 ($1e^{-2}$) expressions which influenced the prediction to shift towards not being a tumour cell. The results suggest that the proposed model maintains biological relevance in its classification.

IV. CONCLUSIONS

Phenotyping cells is an important step in the analysis pipeline of IMC data. We have developed a deep convolutional autoencoder-classifier that demonstrates the potential of deep learning to be used for cell type annotation tasks. We ensured our model's performance was biologically relevant by evaluating the feature significance for classifying the different cell types and examined its modes of failure. The features shown to have high importance adhered to what would be expected biologically. We developed the model using bladder cancer tissue data, but this approach can be extended to other IMC data of various tissue types. Generalizability of this model is a crucial next step that can be explored as IMC data becomes more available. With a more robust dataset to train the network, there is potential to use this methodology in a clinical setting to investigate biological outcome such as biomarker-based cancer detection, or further stratification of cancer sub-types.

REFERENCES

[1] C Giesen et al., "Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry," *Nature methods*, vol. 11, no. 4, pp. 417–422, 2014.

[2] Q Chang, OI Ornatsky, I Siddiqui, A Loboda, VI Baranov, and DW Hedley, "Imaging mass cytometry," *Cytometry part A*, vol. 91, no. 2, pp. 160–169, 2017.

[3] JH Levine et al., "Data-driven phenotypic dissection of aml reveals progenitor-like cells that correlate with prognosis," *Cell*, vol. 162, no. 1, pp. 184–197, 2015.

[4] Y Zhu et al., "Sio: A spatioimageomics pipeline to identify prognostic biomarkers associated with the ovarian tumor microenvironment," *Cancers*, vol. 13, no. 8, pp. 1777, 2021.

[5] Z Hu, A Tang, J Singh, S Bhattacharya, and AJ Butte, "A robust and interpretable end-to-end deep learning model for cytometry data," *Proceedings of the National Academy of Sciences*, vol. 117, no. 35, pp. 21373–21380, 2020.

[6] H Li, U Shaham, KP Stanton, Y Yao, RR Montgomery, and Y Kluger, "Gating mass cytometry data by deep learning," *Bioinformatics*, vol. 33, no. 21, pp. 3423–3430, 2017.

[7] S Thirumal, A Jamzad, T Cotechini, CT Hindmarch, CH Graham, DR Siemens, and P Mousavi, "Titan: An end-to-end data analysis environment for the hyperion imaging system," *Cytometry Part A*, 2022.

[8] Y Guo, Y Liu, A Oerlemans, S Lao, S Wu, and MS Lew, "Deep learning for visual understanding: A review," *Neurocomputing*, vol. 187, pp. 27–48, 2016.

[9] SM Lundberg and S Lee, "A unified approach to interpreting model predictions," in *Proceedings of the 31st international conference on neural information processing systems*, 2017, pp. 4768–4777.

[10] RK Jepsen et al., "Digital image analysis of pan-cytokeratin stained tumor slides for evaluation of tumor budding in pt1/pt2 colorectal cancer: Results of a feasibility study," *Pathology-Research and Practice*, vol. 214, no. 9, pp. 1273–1281, 2018.

[11] MK Najafabadi, E Mirzaeian, SM Montazerin, AR Tavangar, M Tabary, and SM Tavangar, "Role of gata3 in tumor diagnosis: A review," *Pathology-Research and Practice*, p. 153611, 2021.