Open-set based single cell identification in microfluidics

David Dannhauser
Interdisciplinary Research Centre on Biomaterials (CRIB) and Dipartimento di Ingegneria Chimica, dei Materiali e della Produzione Industriale, Università degli Studi di Napoli "Federico II", P.le Tecchio 80, 80125 Naples, Italy
david.dannhauser@unina.it

Paolo Antonio Netti
Interdisciplinary Research Centre on Biomaterials (CRIB) and Dipartimento di Ingegneria Chimica, dei Materiali e della Produzione Industriale, Università degli Studi di Napoli "Federico II", P.le Tecchio 80, 80125 Naples, Italy
nettipa@unina.it

Filippo Causa
Interdisciplinary Research Centre on Biomaterials (CRIB) and Dipartimento di Ingegneria Chimica, dei Materiali e della Produzione Industriale, Università degli Studi di Napoli "Federico II", P.le Tecchio 80, 80125 Naples, Italy
causa@unina.it

2024 IEEE 8th Forum on Research and Technologies for Society and Industry Innovation (RTSI) | 979-8-3503-6213-8/24/$31.00 ©2024 IEEE | DOI: 10.1109/RTSI61910.2024.10761694

***Abstract*— Neural networks are commonly used for image classification in life sciences. However, classifying images of unknown cells poses challenges as existing knowledge cannot guide classification, leading to misclassification. To address this issue, open-set recognition was introduced but has not been extensively tested in single-cell applications. In this work we applied open-set recognition to scattering snapshots of living cells to distinguish known from unknown cells. We examined the impact of neural network parameters to improve unknown cell detection accuracy. Our open-set neural network approach highlights measurement uncertainty in cell prediction, offering potential for diverse single-cell classifications.**

Keywords—Single Cell Analysis, Open-set Recognition, Scattering Snapshot, Microfluidics, Light Scattering

I. Introduction

The human body comprises diverse cell types, each with unique biophysical properties distinguishing them.[1,2] However, obtaining single cell information continuously and cost-effectively is challenging. Microfluidics aids high-throughput single cell analysis, yet data processing remains demanding. Therefore, deep learning (DL) neural networks offer a promising solution for data processing. Moreover, label-free classification methods are desirable for therapeutic analysis and patient monitoring, leveraging biophysical cell signatures. Peripheral blood mononuclear cells (PBMCs) are viable candidates for therapeutic analysis due to their accessibility, despite differing from intestinal tissue cells. In fact the liquid biopsy, revolutionizing clinical oncology, allowing a more frequent patient monitoring.[3,4] Traditional flow cytometry methods are resource-intensive. DL has been introduced to address these limitations by classifying cells in flow conditions using various input data, including also light scattering pattern snapshots, which provide valuable cellular information.

One challenge in cell classification via DL is handling "unknown unknowns," where the model may inaccurately classify novel cell types. In fact today, the majority of image classification models are trained with the closed-set assumption, where the testing data is assumed to be drawn from the same distribution of training data.[5,6] At these unknown cells are forced to choose a class label from one of the known classes, which limits their applicability in cell diagnosis applications. Worth mentioning in this context, that thresholding the classification score value for unknown cells in a closed-set scenario is performing impractical.[7,8]

Open-set recognition introduces the concept of detecting and classifying unknown cells, while accurately identifying known ones. Various methods, have been proposed to address this challenge.[8,11] Extending classifier abilities to identify unknown classes is essential for dynamic cell diagnosis applications.

In this work, we propose a microfluidic-based single cell classification approach using label-free light scattering snapshots and an open-set DL classifier based on an open-set approach (see Fig. 1). The classifier, trained initially with cell classes, successfully predicts both known and unknown cell types. Our method demonstrates the application of out-of-distribution classification in reducing uncertainty in DL models in the biomedical field.

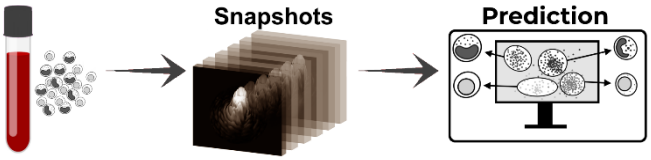

Fig. 1. Single cell identification via scattering snapshots.

II. Material and Methods

A. Sample preparation

Cells were recovered from healthy donors after obtaining informed consent, in accordance with relevant guidelines and regulations. For PBMC a standard density gradient separation was performed together with a cell class isolation based on a negative selection procedure with specific Ab-coupled magnetic beads [13,14]. Whereas the "unknown" cell line THP-1 (ATCC, Manassas, VA, USA) was directly cultured in RPMI-1640 medium, supplemented with 10% FBS, 1% L-Glu and 1% penicillin/streptomycin.

B. Microfluidics

Continuous measurement of single-cell properties was achieved using a microfluidic device consisting of a 3D-printed supporting geometry and two glass channels (see Fig. 2). A round-shaped glass channel was inserted into a hollow square channel, allowing in-flow scattering snapshot readout of cells. By applying pressure, cells were pushed through the alignment channel before entering the readout channel. The sample liquid contained cells in an alignment medium, consisting of a viscoelastic polymer (Polyethylene oxide MW = 4MDa) diluted in PBS at 0.4 wt%. Due to viscoelastic fluid forces, cells were aligned to the channel centreline.

Cell alignment is achieved if the following relationship $3Wi\beta^2\, L/2R > -\ln(3.5\beta)$ is satisfied, where $Wi = 2\lambda\bar{U}/2R$, λ is the relaxation time of the viscoelastic fluid, $\bar{U}$ is the average fluid velocity, R is the channel radius, $\beta = r_l/R$ is a nondimensional geometrical channel parameter, with r_l being the cell radius, and L is the channel length.[12-14] To ensure continuity between the alignment and readout channel, the

alignment section is collinearly inserted into the readout section and sealed with a soft ferrule. At the end of the readout channel, cells can be recovered for further cell studies.

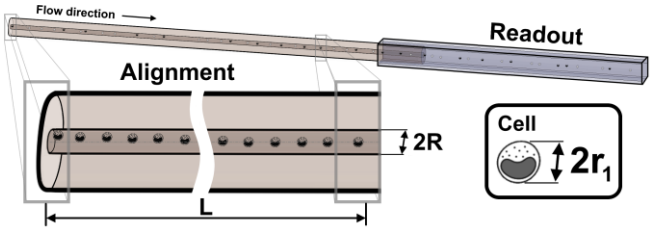

Fig. 2. Cells are aligned to the centreline during their passage in the round shaped capillary, which is inserted into a squared channel to allow maximum readout performance. The soft ferrule, which seals the channel between the two different shapes is not shown for easier readability.

C. Experimental Setup

We used a small-angle light scattering apparatus combined with the microfluidic single-cell alignment device to obtain biophysical single-cell information (see Fig. 3). In more detail, during their passage through the readout channel, cells sequentially encounter the collimated linearly polarized light beam, revealing optical snapshots of living cells. The resulting scattering information is recorded in a continuous angular range from 3° to -30° with an angular resolution of circa 0.1° using a set of optical lenses and a camera sensor with an exposure time of 4 ms, pixel number of 700x700, and pixel size of 6.5 µm. More detailed information about the scattering snapshot recording is provided elsewhere.[13,14]

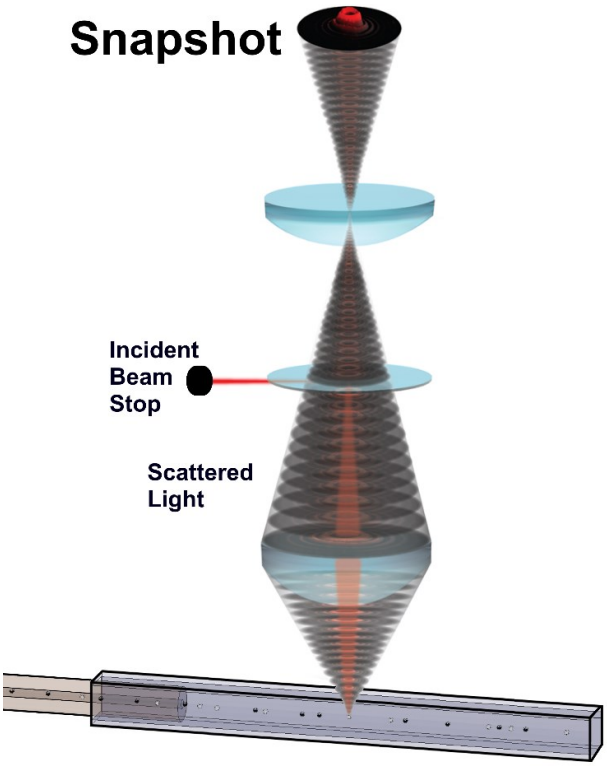

Fig. 3. Experimental setup with typical scattering pattern snapshot on top. Light passes vertically through the microfluidic device and produces scattering pattern when a cell hits the incident light beam (incident light enters from below the microfluidic device, which not shown for easier readability).

D. Data Preparation

The recorded scattering snapshots are pre-processed using a custom MATLAB routine (R2022b, MathWorks), which standardizes snapshots by automatically detecting the scattering pattern centroid (0°) and cropping unwanted scattering regions (< 3° and ≥ 33°).[15] For centroid detection, a binary mask is calculated through an interplay of image processing filters and functions applied to the raw scattering snapshot. In more detail, a 2D Gaussian low-pass filter is iteratively applied four times, followed by spot enhancement using a low-light image enhancement technique based on inversion, haze removal, and further inversion of the enhanced snapshot image. Subsequently, snapshot intensity values are mapped to new values before a global threshold minimizes intra-class variance between low and high-intensity pixels. This process connects separated snapshot components in a unique binary mask, enabling segmentation of the raw scattering snapshot. The centroid detection function is then applied to determine the 0° coordinate of the scattering snapshot. From this position, a donut mask ranging from 3° to 33° is created and overlaid on the original snapshot to select the region of interest (ROI). Following this, snapshots are cropped to 650x650 pixels and resized to 224x224 pixels before being passed to the DL neural network.

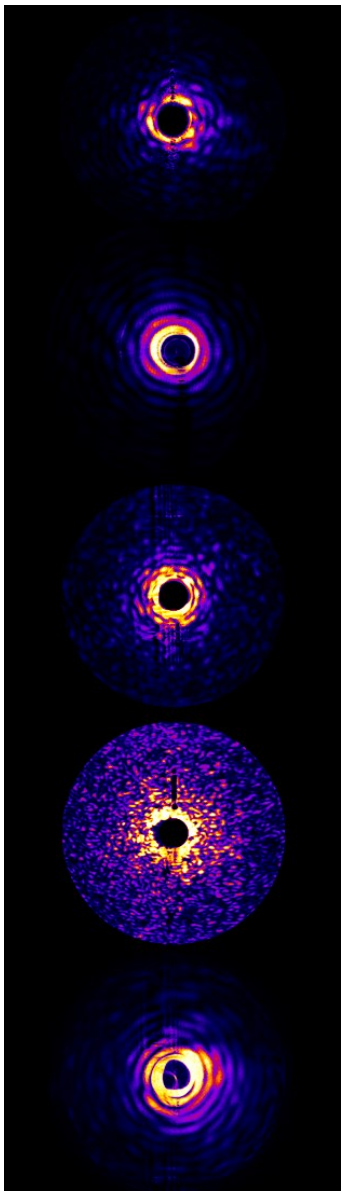

Fig. 4. Typical scattering snapshots for PBMC are illustrated to highlight the difference in cells. All illustrative snapshots were recorded with constant camera exposure time.

E. CNN Classification Framework

In general, CNN frameworks consist of three main types of layers. The input layer receives the scattering snapshot and conveys image information to the hidden layers, where actual feature extraction processing occurs using convolutional filters. Typically, a SoftMax function (in a closed-set environment) is employed as an activation function in the output layer to perform the classification task. The network is trained through backward propagation by adjusting the values of weighted connections to optimize the loss function, such as cross-entropy, representing the difference between the output of the SoftMax function and the desired output, to achieve low classification errors in the training data.

For the closed-set assumption, test data is drawn from the same distribution as the training data. Many types of CNNs can be utilized for such purposes, employing different numbers of filters and network architectures. In our case, due to the intrinsic nature of the experimental scattering data, transfer learning of existing CNNs is challenging, as scattering data is monochromatic and presents significant speckle information as snapshot features. Therefore, we decided to develop a CNN architecture from scratch, optimized for the used scattering pattern range (3-33°), considering the high dynamic range of snapshot feature intensities. The closed-set CNN architecture was designed to classify cell classes, corresponding to training labels. In more detail, the CNN architecture is composed by an input layer, alternating convolution and max pooling layers depicted for features extraction, followed by a sequence of fully connected layers, while the last layer is a SoftMax activation function layer, needed to perform the closed-set classification. To avoid network overfitting, we developed, trained, and tested several CNN architectures with different numbers and types of layers. The most suitable CNN architecture strikes the best compromise between accuracy, loss and computational time.

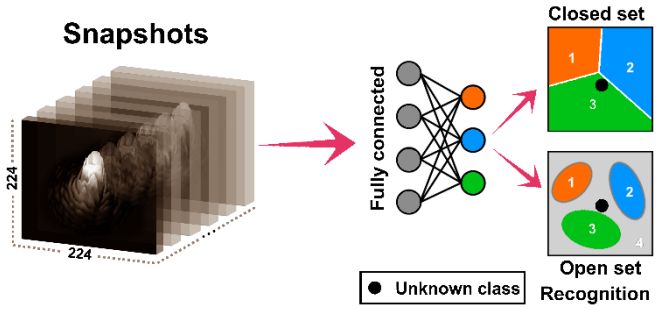

Fig. 5. Schematic overview of the snapshots and neural network. A possible classification outcome for open-set *vs.* closed-set assumption is illustrated, where the closed-set configuration misclassifies a unknown sample (black dot), while the open-set approach correctly handles such measurement condition due to its different neural network architecture (not shown).

On the other hand, open-set recognition for scattering snapshots of living cells was implemented based on the auxiliary open-set risk (AOSR) approach[16], which utilizes an instance weighting strategy to align training samples and auxiliary samples, aiming to recognize unknown cell classes by minimizing the auxiliary open-space risk. In detail, AOSR first defines the label space of known cell classes, while the remaining space is designated as unknown class. Therefore, we train the closed-set CNN architecture to classify the known cell classes. Then, we use the last layer before the SoftMax function as the encoded feature vector to train the AOSR algorithm. Next, we initialize the auxiliary domain of the network architecture by randomly generating samples in the encoded feature space and estimate the weights between the new encoder and SoftMax. The higher the estimated weight, the more likely a sample belongs to known classes. Therefore, the main tuning parameter is $beta_{AOSR}$, which is crucial for defining an ideal auxiliary domain distribution and thereby tuning the feature space to correctly classify unknown samples (see Fig. 5).

III. Results

First, we performed a closed-set prediction for macrophages subtypes (M0, M1, M2) as well as monocytes using different number of epochs for the training process. We trained different classifier models using more than 300 snapshots for each cell class. We found that higher epoch numbers show a significant increase of the model performance, albeit a high epoch number does not automatically ensure a good open-set performance and does not guarantee closed-set overfitting avoidance.

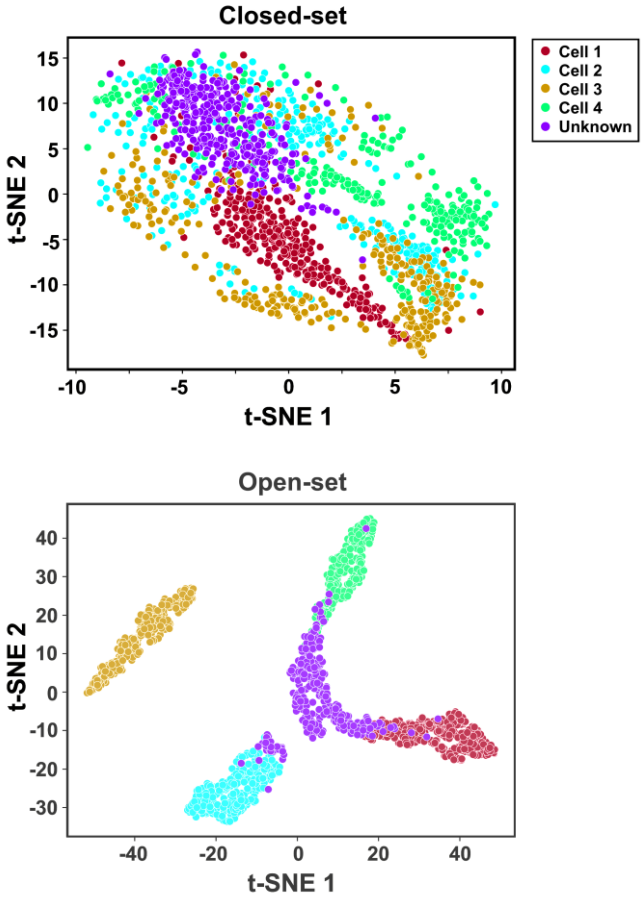

Fig. 6. The t-SNE visualization of scattering snapshots feature inputs for open-set *vs.* closed-set.

However, neural network-based image classification model techniques must have a robust performance that requires trading-off between maximizing the recognition rate and minimizing the inclusion of novel data. Therefore, the open-set prediction goal is to minimize the risk of capturing unknown cells as known. Therefore, we had a closer look at the snapshot input data using t-SNE for dimensionality reduction of snapshot features. Figure 6

shows t-SNE visualization results of scattering patten features of all investigated cell classes (320 cells respectively for M0-, M1-, M2- macrophages and monocytes, as well as 410 unknown cells). As unknown cells, we used acute monocytic leukemia cells, which were not seen by the closed-set architecture during its training phase.

However, the number and feature space of all classes to predict is not known. Therefore, we trained a standard DL neural network, which was then adapted to detect also unknown classes (AOSR approach). The training of the AOSR configuration requires the setting of a double number of epochs, since it relies on the combination of two loss functions. We chose epoch number 1 = 2 and epoch number 2 = 40 at the end of an optimization process.

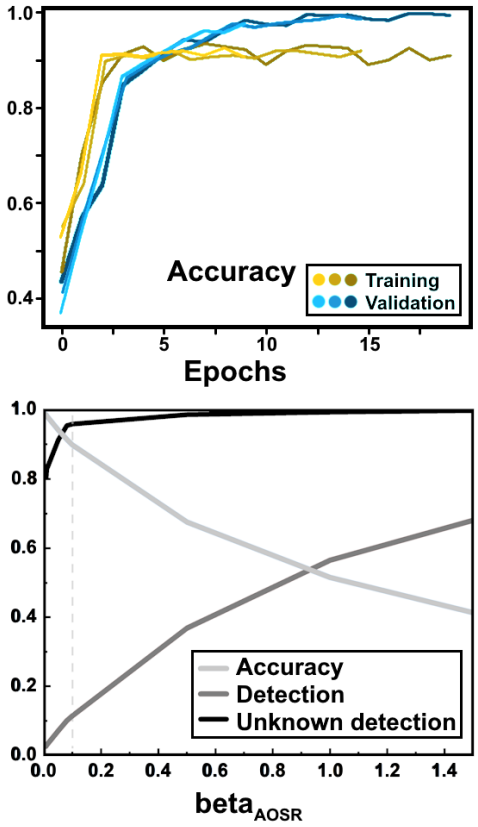

Fig. 7. Training and validation process for the closed-set CNN architecture (top). Cell class prediction accuracy and unknown cell class detection rate are presented for alternating beta$_{AOSR}$.

We tested the network performance as a function of beta$_{AOSR}$ (0.007-1.5), studying the ability to detect an unknown class. For higher beta$_{AOSR}$, the open-set architecture accuracy significantly drops down, while unknown cell detection increases. Therefore, a compromise between neural network accuracy and unknown cell detection precision must be established by selecting the best performing beta$_{AOSR}$. We found beta$_{AOSR}$ = 0.1, the best compromise of known cell prediction and unknown cell detection (see Fig 7). Note, that, higher model accuracy results in an automatic higher unknown detection, which results in a higher misclassification of known cells. On the other side, lower model accuracy leads to higher unknown detection, which results in a higher misclassification of known cells.

IV. Discussion

Our new measurement method is simpler and more cost-effective than classical flow cytometric approaches, enabling cell subtype classification without extensive labeling or large cell populations. It utilizes living cells in suspension, which can be reused for further analysis. We explore how uncertainty impacts single cell classification, particularly in predicting unknown cell types, an area underexplored in life science applications using neural network-based models. While achieving high accuracy for known cell classes with a closed-set architecture, we address misclassification of unknown cell types by implementing the open-set based AOSR architecture, maintaining accuracy without sacrificing prediction quality. Our optimized approach shows promising results for scattering snapshots across various cell classes, with potential for adaptation to other single-cell picture data. We believe such uncertainty studies will enhance cell classification confidence, aiding in the identification of circulating tumor cells.

Acknowledgment

We thank Gaia Cioffi for valuable discussions and support with the open-set architecture.

References

[1] A. Merino, L. Puigví, L. Boldú, S. Alférez, and J. Rodellar, "Optimizing morphology through blood cell image analysis." *Clin. Lab. Haematol.*, 40, 54-61, 2018.

[2] N. Tatsumi, and R. V. Pierre, "Automated image processing: past, present, and future of blood cell morphology identification." *Clinics Lab. Med.*, 22(1), 299-315, 2002.

[3] H. Chen, Z. Zhang, and B. Wang, "Size-and deformability-based isolation of circulating tumor cells with microfluidic chips and their applications in clinical studies." *AIP Advances*, *8*(12), 120701, 2018.

[4] T. Blasi, H. Hennig, H. D. Summers, F. J. Theis, J. Cerveira, J. O. Patterson, D. Davies, A. Filby, A. E. Carpenter, and P. Rees, "Label-free cell cycle analysis for high-throughput imaging flow cytometry." *Nat. Commun.*, 7(1), 10256, 2016.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks." *Commun. ACM*, 60(6), 84-90. 2017.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers:Surpassing human-level performance on imagenet classification," *ICCV*, 2015.

[7] W. J. Scheirer, A. de Rezende Rocha, A. Sapkota, and T. E. Boult, "Toward open set recognition," *TPAMI*, 2013.

[8] A. Bendale, and T. Boult, "Towards open world recognition," *CVPR*, 2015.

[9] T. E. Boult, S. Cruz, A. R. Dhamija, M. Gunther, J. Henrydoss, and W. J. Scheirer, "Learning and the unknown: Surveying steps toward open world recognition," *AAAI*, 2019.

[10] C. Geng, S. J. Huang, and S. Chen, "Recent advances in open set recognition: A survey," *TPAMI*, 2020.

[11] A. Mahdavi, and M. Carvalho, "A survey on open set recognition," *arXiv*, 2109.00893, 2021.

[12] D. Dannhauser, D. Rossi, V. De Gregorio, P. A. Netti, G. Terrazzano, and F. Causa, "Single cell classification of macrophage subtypes by label-free cell signatures and machine learning." *Roy. Soc. Op. Sci.*, 9(9), 220270, 2022.

[13] D. Dannhauser, D. Rossi, M. Ripaldi, P. A. Netti, and F. Causa, "Single-cell screening of multiple biophysical properties in leukemia diagnosis from peripheral blood by pure light scattering," *Sci. Rep.* 7(1), 1-13, 2017.

[14] D. Dannhauser, G. Romeo, F. Causa, I. De Santo, and P. A. Netti, "Multiplex single particle analysis in microfluidics," *Analyst* 139, 5239-5246, 2014.

[15] G. Cioffi, D. Dannhauser, D. Rossi, P. A. Netti, and F. Causa, "Unknown cell class distinction via neural network based scattering snapshot recognition." *Biomed. Opt. Expr.*, 14(10), 5060-5074 2023.

[16] Z. Fang, J. Lu, A. Liu, F. Liu, and G. Zhang, "Learning bounds for open-set learning," International conference on machine learning. *PMLR*, 3122-3132, 2021.