Research Article | Vol. 14, No. 10 / 1 Oct 2023 / Biomedical Optics Express 5060
Biomedical Optics EXPRESS

Unknown cell class distinction via neural network based scattering snapshot recognition

GAIA CIOFFI,[1,†] DAVID DANNHAUSER,[1,*,†] DOMENICO ROSSI,[2] PAOLO A. NETTI,[1,2] AND FILIPPO CAUSA[1]

[1] *Interdisciplinary Research Centre on Biomaterials (CRIB) and Dipartimento di Ingegneria Chimica, dei Materiali e della Produzione Industriale, Università degli Studi di Napoli "Federico II", Piazzale Tecchio 80, 80125 Naples, Italy*
[2] *Center for Advanced Biomaterials for Healthcare@CRIB, Istituto Italiano di Tecnologia, Largo Barsanti e Matteucci 53, 80125 Naples, Italy*
[†] These authors contributed equally to this work
[*] *david.dannhauser@unina.it*

Abstract: Neural network-based image classification is widely used in life science applications. However, it is essential to extrapolate a correct classification method for unknown images, where no prior knowledge can be utilised. Under a closed set assumption, unknown images will be inevitably misclassified, but this can be genuinely overcome choosing an open-set classification approach, which first generates an in-distribution of identified images to successively discriminate out-of-distribution images. The testing of such image classification for single cell applications in life science scenarios has yet to be done but could broaden our expertise in quantifying the influence of prediction uncertainty in deep learning. In this framework, we implemented the open-set concept on scattering snapshots of living cells to distinguish between unknown and known cell classes, targeting four different known monoblast cell classes and a single tumoral unknown monoblast cell line. We also investigated the influence on experimental sample errors and optimised neural network hyperparameters to obtain a high unknown cell class detection accuracy. We discovered that our open-set approach exhibits robustness against sample noise, a crucial aspect for its application in life science. Moreover, the presented open-set based neural network reveals measurement uncertainty out of the cell prediction, which can be applied to a wide range of single cell classifications.

1. Introduction

The human body is composed of a variety of types of cells, where each one has its own biophysical properties, such as dimension, structure, and function. In other words, cells contain a cell specific biophysical signature, which allows to distinguish them from each other. However, the acquisition of useful single cell information in a straightforward continuous and cost-effective manner remains challenging. To this end, microfluidics enables high-throughput rates at single cell level, but data elaboration persists being demanding. A key enabling technology for next generation cell discrimination in flow can be given by deep learning (DL) based neural network classifiers. Nowadays, cell discrimination is routinely performed by cytometric analysis, based on surface receptor expression, which is a costly and time intensive procedure [1]. Beyond that, monoclonal antibodies that bind on cell surface molecules during the cytometric analysis, may interfere with cell functions, which can complicate cell reuse. Alternatively, a label-free classification, which preserves the integrity of cell functions is desirable, when cells are analysed for a therapeutic purpose. In fact, biophysical cell signatures can be useful for their label-free classification in microfluidics as well as for patient monitoring during a therapy, or to promptly identify recurrence (e.g., leukaemia or generic bone marrow disease).

#492028 https://doi.org/10.1364/BOE.492028
Journal © 2023 Received 31 Mar 2023; revised 25 Jul 2023; accepted 31 Jul 2023; published 6 Sep 2023

Biomedical Optics EXPRESS

For therapeutic purposes, human peripheral blood mononuclear cells (PBMCs) represent an appropriate candidate, due to their simple availability. Despite PBMCs having a different composition, phenotype, and activation status from cells found in intestinal tissue, they are considered as a liquid biopsy of our body. This latter has drastically revolutionised the field of clinical oncology, with its possibility of continuous patient monitoring and circulating tumour cell analysis [2]. Recent advancements in screening of label-free cell information have shown the possibility to distinguish fractions of cell class and/or states, especially when dealing with sparsely present cells [3–10]. Beside the single cell recognition, the relative peripheral cell class count alterations, lower or higher than physiologic one, as well as anomalous cell shapes or cytoplasm complexity are useful indicators of dysregulated responses to inflammation stimuli [11,12]. Such investigations are generally performed with classical flow cytometry methods, which are known as expensive and resource intensive measurement tools. To overcome the main limitations of classical flow cytometry approaches, DL was recently introduced for the classification of cells in flow condition [13]. In this, DL processes either phase contrast or bright-field images as input, while light scattering pattern snapshots are understudied [14–17]. Nevertheless, scattering snapshots provide invaluable cellular information of cell shape and inner structures, which can significantly assist an individual cell class distinction [14,18–20]. In this study, we aim to clarify and illustrate the enormous potential of pure light scattering snapshots obtained in microfluidics, for accurate DL-based cell classification.

Cells are not homogenous objects: they are composed of membranes, cytoplasm, nuclei, and various organelles which are among the most important contributors to the scattering pattern. Therefore, scattering pattern direction and width are fundamental measurement parameters to be considered. For instance, forward scattering (~0°) reveals significant cell size and inner structure information. At side direction (5°-30°) nucleus and nucleus over cell ratio are predominant information parameters, while small internal organelles contribute more at larger angles to the scattering pattern (~90°). The backward direction (~180°) can be useful for membrane roughness investigations [21,22]. Overall, an intermediate state of forward and side scattering information offers the most promising approach for label-free cell identification.

In addition to the mentioned heterogeneity of intrinsic cell properties, a crucial point for the automatic classification of cells via neural networks is the handling of so-called “unknown unknowns”. In other words, we don’t know what we don’t know [23]. In such cases the classification model is confident about its choice of classifying a never-before-seen cell class but is actually performing wrong. In fact, a classification model should not only produce accurate predictions of known cells, but also detect unknown cells and reject or classify them in a new class of cells [24–27]. Until now, most existing models for image classification are trained based on the closed-set assumption, where the test data is assumed to be drawn from the same distribution of training data [28,29]. In the case of these unknown cells, they must opt for a class label from the existing known classes thus significantly limiting their applicability in dynamic and ever-changing cell diagnosis applications. For instance, thresholding the classification score value for unknown cells in a closed-set scenario proves to be impractical [23,30,31].

This classification problem necessitates the creation of an open-set recognition concept to classify unknown cells that were not encountered during training while, simultaneously, achieving highly accurate classification of known cells. To recognize new cell classes, the neural network-based classifier must project known cell class input in very compact and separated regions of the features space, to finally distinct distant unknown cells. Hereby, samples included and excluded in the label space are referred to as knowns and unknowns, respectively. In other words, open-set based classifiers need to use incomplete knowledge learned from a finite set of accessible cell classes to devise effective representations able to separate known from unknown cells [23]. In more detail, when a classification model is applied on open-set recognition tasks, a semantic shift (e.g., due to the occurrence of new cell classes), or covariate shift emerge in

Biomedical Optics EXPRESS

the label space [32–35]. Beside the open-set recognition [36–38] also outlier detection [39–42], anomaly detection (detect any anomalous cells that are derived from the predefined normality during testing) [43–45] and novelty detection [46–49] are from significant interest for the so called out-of-distribution detection [23]. OpenMax was the first emergent open-set classifier designed to address the out-of-distribution problem [31]. From that point onward, other open-set classifiers have been reported in literature, which apply a minimising of the open-space risk or using an extreme value theory [30,31,50,51]. However, most open-set recognition methods use threshold-based strategies, in which the threshold is selected using the knowledge of the known classes. Thus, without prior knowledge about unknown classes and applying a constant global threshold will lead to a significant open-set recognition risk. For instance, modifying the threshold based on the information of unknown class received at the test phase can improve the robustness of open-set approaches. Another area of investigation would involve the robust selection of the tail's size while applying an extreme value theory to model the data distribution's tail. In fact, extreme value modelling has been increasingly used to analyse post-processing scores and enhance the performance of open-set recognition. For the sake of diversity, open-set classifiers utilise the intuition that new cells to be classified as known or unknown are more likely to be unknown if they are far away from the training data. To address the open-set challenge, investigating the ability of classifier to identify unknown class domain, a so called auxiliary open-set risk (AOSR) DL network was presented recently [52]. The AOSR approach utilise an instance weighting strategy to align training sample and auxiliary sample, which is generated through an auxiliary domain that minimises the auxiliary risk, to learn how to recognize unknown classes. All in all, we define a certain in-distribution, with the target of detection of out-of-distribution cells under the open-set assumption.

Therefore, we present a simple microfluidic based single cell classification approach from label-free obtained light scattering snapshots, which can predict known as well as unknown cell classes thanks to an underlying open-set classifier based on the AOSR principle. In comparison with previous single cell investigation approaches presented in literature, no matching of scattering patterns with adequate scattering models is needed [53–57]. Such straightforward approach significantly reduces the time and computational costs for single cell detection. Moreover, the acquisition of scattering snapshots in s label-free under flow conditions makes this approach highly versatile in the application to multiple cell types and sizes. The mentioned DL-based cell discrimination uses scattering snapshots as input to predict the searched for cell types. Therefore, the open-set prediction model was first trained with monoblasts. We also investigated the effect of experimental sample noise, substituting cell data with 10% or 20% of debris (organic waste left over after a cell dies) snapshots. Note that monoblast subclasses such as monocytes and macrophages play a primary role during the innate immune response, where macrophages represent the resident cells in peripheral tissue derived from blood circulating monocytes that can extravasate from the bloodstream. As proof-of-concept, a monocytic cell line derived from an acute monocytic Leukaemia patient, named THP-1, was added as an unknown tumour cell class for the testing phase of the open-set classifier model. Indeed, the obtained classifier was utilized on a mixed dataset containing both known monoblasts and unknown THP-1 cells allowing the automatic prediction of all the present cell classes. The presented approach shows high versatility and could be applied for circulating tumour cell detection in microfluidics, where, in general, no prior cell knowledge can be used for the model training. In fact, the presented work gives emphasis to the application of an out-of-distribution classification approach in the biomedical field to minimize uncertainty in DL classification models.

Biomedical Optics EXPRESS

2. Materials and methods

2.1. Sample preparation

Cells were recovered from healthy donors after obtaining informed consent, in accordance with relevant guidelines and regulations. For peripheral blood monocytes and macrophages, a standard density gradient separation was performed as followed: first, blood was diluted with an equal volume of phosphate buffered saline (PBS), and then gently layered on with an equal volume fraction of density gradient medium (Histopaque-1077) using a 50 mL centrifuge tube. After that, a centrifugation step was performed at $300\vec{g}$ for 30 min and disabled machine brake, resulting in a visible PBMC ring at the interface between gradient medium and plasma. Cells were collected and washed in the Erythrocyte lysis buffer, to eliminate a possible contamination. Finally, cells were cultured in RPMI-1640 medium, supplemented with 10% fetal bovine serum (FBS), 1% L-Glu and 1% penicillin/streptomycin (Euroclone). Next, PBMC were divided into four culture flasks (T-75, Corning) to obtain monocytes and to transform monocytes in unpolarized (M0), M1-polarised (M1) and M2-polarised (M2) macrophage phenotypes. First cells were incubated for 24 hours at 37°C and 5% CO_2. For the monocyte flask, cells in suspension (lymphocytes) were discarded, while adherent monocytes were diluted in RPMI-1640. Next, in the remaining three flasks cell medium was substituted with RPMI-1640 and specific macrophage phenotype generation media (M0 = C-28057; M1 = C-28055; M2 = C-28056) was added following the manufacturer instructions (Promocell). After 6 days each flask was supplied with a volume of cell medium equal to the 75% of the initial cell volume (day 0). On day 7 a new aliquot of cytokine mix (Promocell) was added following the manufacturer instructions. At day 9 cell medium was aspirated and a fresh medium was added to each flask. At day 10, polarised and not activated macrophages were detached from the flask surfaces using a cell scraper tool and subsequently centrifuged at $200\vec{g}$ for 10 min in 15 mL centrifuge tubes. Whereas the THP-1 cell line (ATCC, Manassas, VA, USA) was directly cultured in RPMI-1640 medium, supplemented with 10% FBS, 1% L-Glu and 1% penicillin/streptomycin. However, all separated cell classes were resuspended into complete RPMI-1640 medium, ready to be analysed with the single cell scattering snapshot approach.

2.2. Microfluidic device and alignment

Continuous measurement of single cell properties was achieved with a microfluidic device, composed of a supporting geometry fabricated with a 3D printer (Objet30 pro, Stratasys) and a series of two glass channels (see Fig. 1). Briefly, a round shaped glass channel (R = 50 µm) is inserted on one side in a hollow square channel (ID = 400 µm), which permits in-flow scattering snapshot readout of cells, and on the other side immersed in the cell sample. By applying a certain pressure on the sample liquid (P-pump, Dolomite), cells are pushed through the centreline of the alignment channel before entering the readout channel. The sample liquid consists of cells immersed in an alignment medium, consisting of a highly biocompatible viscoelastic polymer (polyethylene oxide, M_W = 4 MDa, Sigma Aldrich) diluted in PBS at 0.4 wt%. However, thanks to resulting fluid properties, generated by viscoelastic fluid forces, cells are aligned to the centreline of the alignment channel and subsequently remain aligned at the centreline of the subsequent readout channel [58]. Note that fluid forces were chosen to prevent cell deformation effects, while ensuring sufficient cell alignment to the channel centreline. In more detail, cell alignment was achieved if the following relationship $3Wi\,\beta^2\frac{L}{2R} > -ln(3.5\beta)$ was satisfied. Where $Wi = 2\lambda\bar{U}/2R$, uses λ the relaxation time of the viscoelastic fluid, $\bar{U}$ the average fluid velocity, R the channel radius, $\beta = r_1/R$, a nondimensional geometrical channel parameter, with r_1 being the cell radius, and L the channel length [59]. To ensure continuity between the alignment and readout channel, the alignment section was collinearly inserted in the readout section and sealed with a soft ferrule

Biomedical Optics EXPRESS

(UP-N-123-03X, Idex). At the end of the readout channel, cells can be recovered for further cell studies.

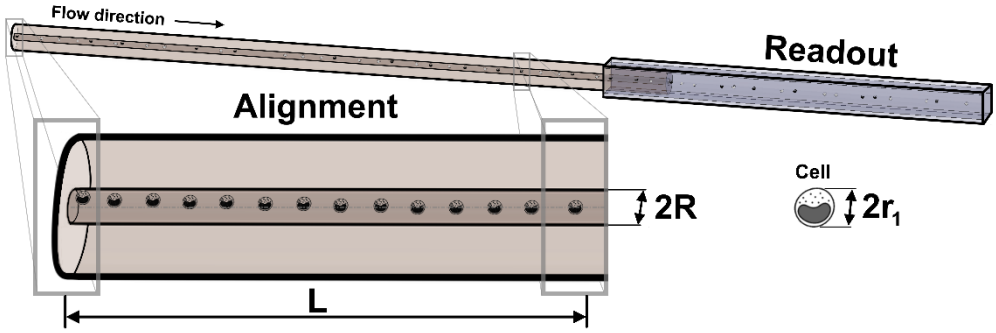

Fig. 1. Microfluidic alignment and readout principle. Cells are aligned to the centreline during their passage in the round shaped alignment channel. This channel is inserted into a squared channel to allow maximum readout performance, while conserving the cell alignment. The soft ferrule, which seals the connection between the two different shaped channels is not shown for easier readability.

2.3. Experimental setup

For this study, we utilised a small angle light scattering apparatus [60], combined with the previously mentioned microfluidic single cell alignment device to obtain biophysical single cell information (see Fig. 2). In more detail, during their passage through the readout channel, cells hit one after another the collimated linear polarised light beam (HeNe with 632.8 nm and 5 mW), that reveals the optical scattering snapshots of a living cell (in-flow records). The resulting scattering information is recorded in a continuous angular range from ~3° to 30° with and an angular resolution of ~0.1° using a set of optical lenses and a camera sensor (ORCA Flash 4.0, Hamamatsu Photonics) with an exposure time of 4 ms, pixel number of 700×700 and pixel size of 6,5 µm (see Fig. 2). More detailed information about the scattering snapshot recording is shown elsewhere [53,60].

2.4. Data preparation

The recorded scattering snapshots are pre-processed by a self-written MATLAB (R2022b, Mathworks) routine, which uniforms snapshots by automatically detecting the scattering pattern centroid (scattering angle of 0°), cropping unwanted scattering regions (scattering angle $\leq 3°$ and $\geq 33°$). For the centroid detection, a binary mask was calculated through an interplay between a sequence of several image processing filters and functions applied on the raw scattering snapshot. In more detail, first a 2D gaussian low-pass filter (size 15×15 pixel and standard deviation of 6) was iteratively applied four times on the snapshot, followed by a spot-enhancement using a low-light image enhancement based on inversion of low-light image, applying a haze removal algorithm and an inversion of the enhanced snapshot image. Then, snapshot intensity values were mapped on new values before a global threshold (Otsu's method) [61] minimised the intra class variance between low and high intensity pixels. Thus, connects separated snapshot components in a unique binary mask, to finally perform a segmentation on the raw scattering snapshot. Subsequently, a centroid detection function returns the 0° coordinate of the scattering snapshot. From such a position a donut mask from 3° (radius of 30 pixel) to 33° (radius of 400 pixel) was created and overlaid on the original snapshot to select the region of interest (ROI). Afterwards, the snapshots were cropped to the ROI, resulting in an image size of 650×650 pixel,

Biomedical Optics EXPRESS

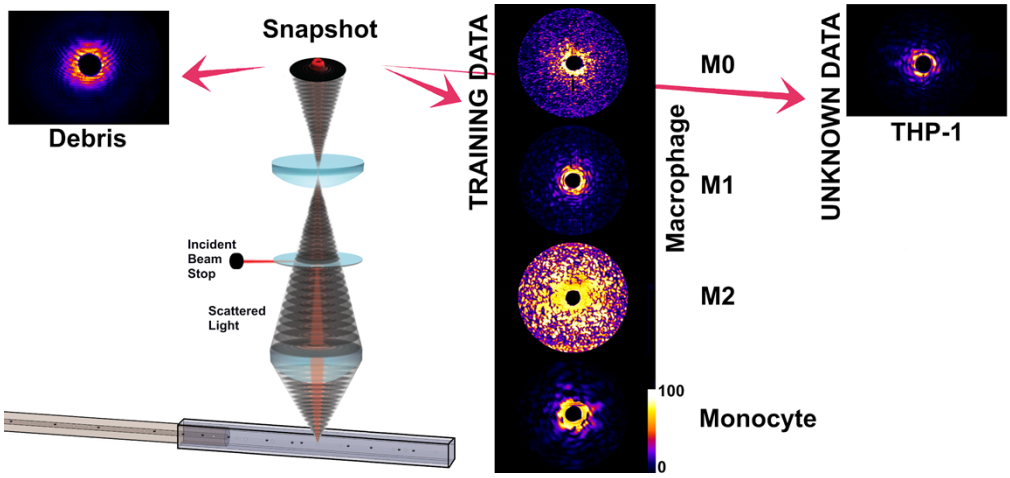

Fig. 2. Experimental setup and scattering snapshots. Light passes vertically through the microfluidic device and generates scattering snapshots when a cell hits the incident light beam. Typical snapshots for training data (debris, monocyte, M0-, M1- and M2-macrophage) as well as unknown cell data (THP-1) are illustrated to highlight the different scattering patterns. All snapshots are recorded with constant camera exposure time.

which subsequently was resized (224×224 pixel) before passing them to the neural network. Note that different resize dimensions were tested for the obtained scattering snapshots, while the mentioned snapshot dimension results in a good compromise between snapshot feature quality and computational costs (see ESI Fig.S1).

2.5. CNN classification framework

In general, convolutional neural networks (CNN) frameworks consist of 3 main types of layers. The input layer, which receives the scattering snapshot and communicates image information to the hidden layers where the actual feature extraction processing is done using the convolutional filters. Usually, a SoftMax function (closed-set environment) is used as an activation function in the output layer to perform classification task. The network is trained in backward propagation by adjusting the value of weighted connections to optimise the loss function, such as cross-entropy that represents the difference between the output of the SoftMax function and the desired output to achieve a low classification error in the training data.

2.5.1. Closed-set classifier

For the closed-set assumption, test data is assumed to be drawn from the same distribution of training data. In other words, the number of input and output class labels is constant. Many types of CNN can be used for such purposes using a different number of filters and network architecture. For the intrinsic nature of the experimental scattering data, transfer learning of existing CNN is challenging since scattering data is monochromatic and presents significant speckle information as snapshot features. Therefore, we decided to develop from scratch a CNN architecture, optimised for the used scattering pattern range (3-33°) considering the high dynamic range of snapshot feature intensities (see Fig. 3).

The close-set CNN architecture was designed to classify four different cell classes, corresponding to four training labels. For this task, along with avoiding network overfitting and guaranteeing structure robustness for further AOSR implementation, we developed, trained, and tested several CNN architectures with different numbers and types of layers (see ESI Fig. S3- ESI Table

Biomedical Optics EXPRESS

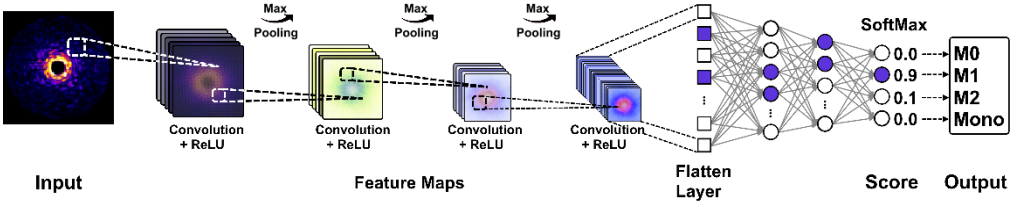

Fig. 3. Architecture of the closed-set CNN. The convolutional layers perform the feature extraction from the input image and are connected to the flatten layer, which is joined to a series of dense layers and dropout layers constituting the fully connected layer needed to perform classification. The last layer (SoftMax layer) normalises the scores and predicts the cell class.

S1). The most suitable CNN architecture represents the best compromise between accuracy, loss (overfitting monitoring) and computational time (see ESI Fig. S4). In more detail, the CNN architecture is composed by an input layer with a dimension of $224 \times 224 \times 1$, 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 (see ESI Fig. S3b). Beside the mentioned monoblast classification a pre-screening convolution neural network -regarding cell versus not cell snapshots- was developed using a binary SoftMax function. For this purpose, a CNN structure similar to the aforementioned closed-set model was used, resulting in circa 5000 labelled monoblasts and ~7000 debris snapshots. The aim of this pre-screening step was to provide a dataset without cell debris content and subsequently, to test model robustness of newly developed open-set models, incorporating a defined content of no-cell material.

2.5.2. Open-set classifier

Open-set recognition for scattering snapshots of living cells was implemented based on the AOSR approach [52], which utilised an instance weighting strategy to align training samples and auxiliary samples, which aims to recognize unknown (not seen during training) cell classes by minimising the auxiliary open space risk. In more detail, AOSR first defines the label space of known cell classes (defined as correct known classes), while the remaining space is allocated as unknown class. Therefore, we train the closed-set CNN architecture to classify the known

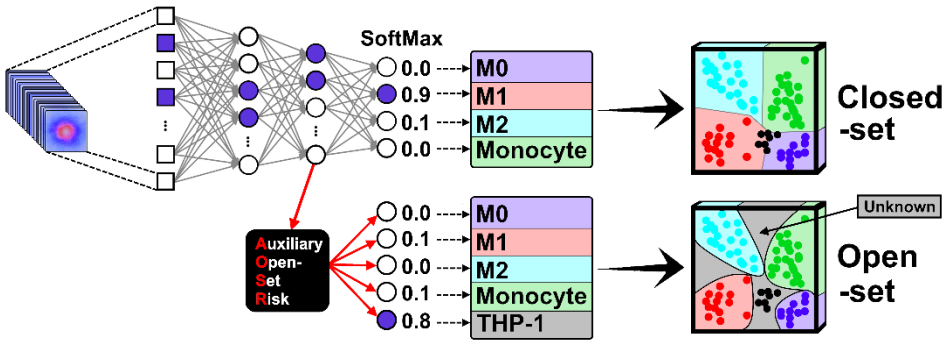

Fig. 4. Closed-set versus open-set classifier (AOSR) architecture. The closed-set environment misclassifies unknown samples (black dots). AOSR uses the penultimate layer (before SoftMax) of closed-set CNN to build an encoder whose outputs are used as a decoder as input to classify unknown cell classes.

Biomedical Optics EXPRESS

cell classes. Then, we use the last layer before the SoftMax function (penultimate CNN layer) as the encoded feature vector to train the AOSR algorithm (see Fig. 4). Next, we initialised the auxiliary domain of the network architecture by randomly generating samples in the encoded feature space and estimated the weights between the new encoder and SoftMax. The higher such an estimated weight, the more likely a sample belongs to known classes. The main tuning parameter is β, which is important to define an ideal auxiliary domain distribution and therefore tuning the feature space to correctly classify unknown samples. More detailed information about AOSR can be seen elsewhere [52].

3. Results and discussion

In this work, we tested monoblast prediction accuracy in a real-world life-science application. We investigated known and unknown cell classes with an open- and closed-set CNN architecture using scattering snapshot with different percentages of experimental sample error (debris) as input to minimise prediction uncertainty. Therefore, one acute monocytic leukaemia cell line (unknown) and four monoblast (known) cell classes were analysed separately with an optical cell investigation tool using a microfluidic cell alignment approach to retrieve the search for single cell scattering snapshots. The separate measurement of different cell classes allowed us to create labelled training samples, which are needed for the classifier model (see Fig. 5). Moreover, in-flow records are classified in cell *versus* no-cell snapshots using a closed-set binary classifier. This pre-screening step permits the adjustment of the amount of experimental sample error by removing cell snapshots and replacing them with debris content.

Fig. 5. Working principle of cell class prediction with a CNN model from a fresh blood sample using single cell scattering snapshots. The classifier model is altered for closed- and open-set application to predict known and unknown cell classes at the same time.

Then, each separately investigated cell class was diluted in viscoelastic cell alignment medium. We prepared living cell samples of ~$1.25 * 10^5$ cells mL^{-1} to ensure a throughput rate of ~2 cells sec^{-1} passing through the readout laser beam. Each scattering event was recorded as a snapshot for further image analysis and cell class prediction. Note that cell throughput -for the used camera sensor- can be increased up to ~50 cell sec^{-1} by simply changing the fluid flow rate, which in this proof-of-concept was not needed.

First, the pre-processing of scattering snapshot was calibrated and tested with polystyrene beads of different dimensions (see ESI Fig.S2). In this process, snapshots are uniformed to reduce scattering centroid misalignment due to experimental variations in the microfluidic cell alignment. In fact, such initial snapshot normalisation processes significantly improve further CNN performances. Even if an excessive pre-processing may lead to a natural distortion of the raw snapshot dataset, a proper balance significantly improves misclassification [62] and therefore speeds-up the classification model training process. However, after calibration, we applied the pre-screening procedure on in-flow records before a binary DL-based cell prediction approach identified cell *versus* not cell snapshots. Thus, initial operations, which involved removing debris or possible cell agglomerates from the sample data, had a significant improvement on the training performance of the classifier model. Note that each scattering event during an in-flow record was considered for this process, which enables an automation of the measurement process. In other words, pre-screening after snapshot normalisation removes the main part of experimental

Biomedical Optics EXPRESS

errors from the sample data. In fact, debris snapshots were used as experimental sample error dataset for the investigation of CNN robustness against noisy data. In other words, sample noise is not considered to improve unknown cell detection accuracy, while it can help to simulate a more realistic life-science measurement scenario, where cell heterogeneity and sample noise are common.

Next, we predicted four different monoblast cell classes (M0-, M1- and M2-macrophages as well as monocytes) using a widely used closed-set prediction architecture, based on a SoftMax activation function. A part of the previously obtained not-cell snapshots was reused for the CNN training to investigate the influence of cell debris on the classification accuracy and overall network robustness. Note that not-cell snapshot samples were randomly selected from the debris dataset to replace in equal parts content of each investigated cell class (dataset number remains constant). Next, we implemented an AOSR open-set approach to the closed-set architecture and optimised it for an unknown (THP-1) cell class. Lastly, we performed a prediction of mixed known and unknown cell classes using a closed-set as well as open-set architecture to show the model performance, regarding prediction uncertainty.

3.1. Closed-set prediction

We performed a closed-set prediction for M0-, M1- and M2-macrophages as well as monocytes using different CNN architecture (see ESI Fig. S3). We trained the classifier model with 10, 15 and 20 epochs using 400 cells (80% training, 20% validation) for each cell class. Moreover, we substituted 10% as well as 20% of cell data with not-cell ones (debris) to investigate the influence of experimental measurement noise on the prediction outcome. We tested and computed confusion matrices for the different CNN architectures with a different testing dataset of 221, 332, 179 and 65 cells for M0-, M1-, M2- macrophages and monocytes respectively. We observed for epoch 10 the highest misclassification for M1-macrophages, which were classified as M2-macrophages for training data without debris content. While for increasing no-cell content M1-macrophages were better predicted, and a higher misclassification for M2-macrophages as M1-macrophages was observed for all investigated model architectures. This misclassification was observed for the CNN - 2, 3 and 4 architectures, while for CNN - 1 the misclassification trend remained constant (see ESI Fig. S4). According to the misclassification performance and Matthew correlation coefficient (MCC) calculations (see ESI Fig. S5) CNN - 2 show the best performance for all investigated measurement conditions.

Therefore, we focused on the CNN - 2 model and investigated the training and validation accuracy, as well as confusion matrices (see Fig. 6). Confusion matrix outcomes with epoch 15 are shown in Fig. 6(b). Here, the misclassification ratio between M1- and M2-macrophages is changing for different amounts of no-cell content. This implies that no-cell content strongly influences M1, moderate influence M2 and not significantly influences M0 as well as monocytes predictions. However, higher epoch numbers show a significant increase of the model performance for only cell data, while for the dataset with no-cell content, the training and validation accuracy remains constant (see Fig. 6(a)). In fact, a higher epoch number could possibly further increase the performance with no-cell data for the presented CNN architecture. Albeit it does not automatically ensure a good open-set performance and does not guarantee closed-set overfitting avoidance. Note that validation behaviour with no-cell content is more unstable compared to only cell content datasets.

3.2. Snapshot features

A DL-based classification model technique must have a robust performance that requires trading off between maximising the recognition rate and minimising the inclusion of novel data. The open-set prediction goal is to minimise the open space risk to capture the risk of labelling the unknown cells as known [23]. Therefore, we had a closer look on the snapshot input data using

Biomedical Optics EXPRESS

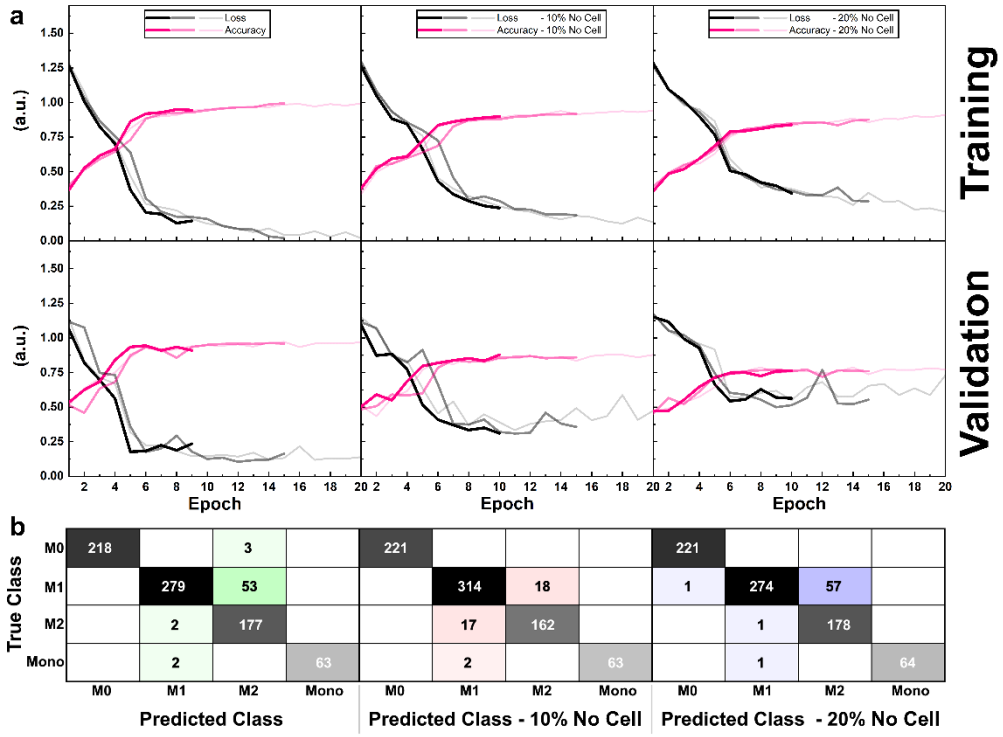

Fig. 6. Training and validation process for the closed-set CNN architecture. (a) Accuracy increase and loss decrease for higher epoch values, while experimental sample error of 10% and 20% of no-cell content in the dataset (size of the training dataset is constant), respectively decrease accuracy and *vice versa* increase the obtained loss outcome. Each training process was performed for a fixed epoch number of 10, 15 and 20, which is indicated with increasing transparency for higher epoch numbers. (b) Confusion matrix (for a testing dataset composed of 221, 332, 179 and 65 cells for M0-, M1-, M2- macrophages and monocytes respectively) outcomes for the closed-set model with different percentages of no-cell content for an epoch number of 15 are presented.

t-SNE for dimensionality reduction of snapshot features. Figure 7 shows t-SNE visualisation results of cell snapshot features of all investigated cell classes (320 cells respectively for M0-, M1-, M2- macrophages and monocytes, as well as 410 cells for THP-1). Latter cells are acute monocytic leukaemia cells, which are used as proof-of-concept for an unknown cell class, which was not seen by the closed-set architecture during the training phase.

No evident clustering or separation of an individual cell class was noticed for raw data features (input for closed-set CNN), while a strong clustering is present for input features for the AOSR algorithm. This outcome indicates that THP-1 cells share significant snapshot features with all other cell classes except M2-macrophages, which would inhibit a correct closed-set prediction. Note that M2-macrophages show significantly different scattering snapshot images compared to the other cell classes, which is in-line with the t-SNE representation. In fact, due to its closed-set nature, a standard classification model will link an unknown cell (THP-1) to the class with the maximum score given by the SoftMax function, leading to a misclassification, which clearly demonstrates the need of an open-set recognition approach.

Biomedical Optics EXPRESS

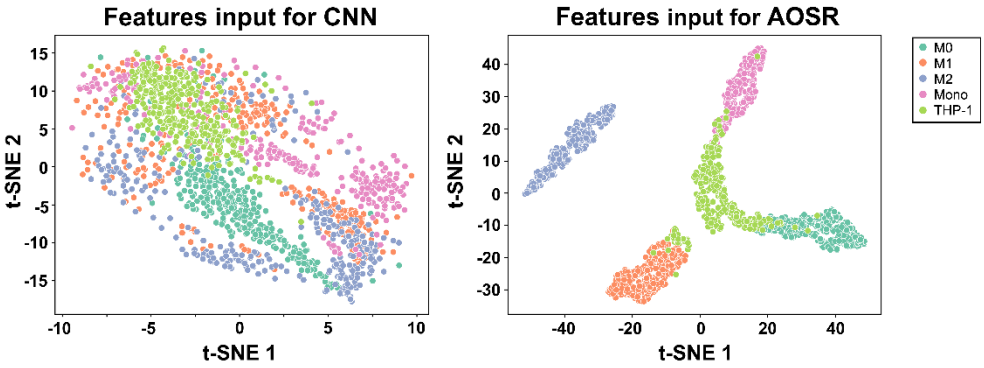

Fig. 7. The t-SNE visualisation (Python: sklearn.manifold.TSNE applying a perplexity value of 100) of snapshot features of the closed-set CNN and feature input for AOSR extracted from the closed-set CNN. The presented AOSR features are extracted from the penultimate layer of the closed-set CNN - 2 architecture using an epoch of 15.

3.3. Open-set prediction

For open-set cell class recognition, during the training phase we do not know the number and feature space of all classes to predict, so precedent modelling of an unknown class is challenging. Therefore, almost all existing open-set approaches include standard neural networks architectures, which are first trained in a closed- set environment and afterwards adapted to detect unknown sample classes.

In this work, we followed this concept and utilised the best performing closed-set architecture, modified for open-set recognition, applying the AOSR approach. Such unknown cell class detection architecture was first tested for different combinations of epoch numbers of closed- and open-set architecture. For this purpose, the decoder training of AOSR, needs a combination of two loss functions, governed by two different epoch values, where epoch 2 was randomly chosen to be 10 times epoch 1. In more detail, the AOSR algorithm was characterised by the first loss function, which is a classical sparse categorical cross entropy CNN function. This function is needed to fit the new classifier and considering the presence of a new class - the unknown one - for which the closed-set CNN was not trained. Subsequently, the second loss function, defined as the "auxiliary risk" loss function, manages the auxiliary domain and its risk to detect unknown cell class. In fact, the latter loss function uses a significantly higher epoch number to perform the proper AOSR training and learning, while the first one fits the classifier model structure with a comparatively small epoch number. An epoch number of 10, 15 and 20 was tested for the closed-set architecture, while a combination of epoch number 2, 3, 4 and 5, respectively 20, 30, 40 and 50 were tested for the AOSR model (see ESI Table S2). All possible epoch number combinations over a wide range of β (0.007-1.5) were trained and tested with scattering snapshots -including also experimental no-cell content of 10% and 20%- to obtain the best performing training accuracy and unknown cell class detection (see ESI Table S2). Results indicate an epoch number of 15 for the closed-set architecture combined with an AOSR epoch 1 = 4 and epoch 2 = 40 for the open-set architecture as the best performing parameter combination. Next, we investigated training accuracy as well as unknown cell class detection for THP-1 and test data (known cell classes) *versus* β using the best performing epoch number combination (15, 4 and 40). We tested unknown cell class detection with a dataset of 410 THP-1, 221, 182, 179 and 65 M0-, M1-, M2- macrophages and monocytes respectively. Note that β is an hyperparameter contributing to the definition of an auxiliary domain distribution and therefore defines the importance of the correctly classified unknown sample domain (see Fig. 8). In fact, for higher β, the open-set architecture accuracy significantly drops down, while unknown

Biomedical Optics EXPRESS

cell detection increases. Therefore, a compromise between CNN accuracy and unknown cell detection precision must be established by selecting the best performing β. We selected a β-value of 0.1 (dashed grey line in Fig. 8(a)) as the best combination of cell prediction as well as unknown cell detection accuracy (see MCC outcome in ESI Fig. S6 and best performing epoch number combination in ESI Table S2). In more detail, results (see THP-1 plot in Fig. 8(a)) show a significantly higher unknown cell detection for training data without addition of no-cell content (96% at $\beta = 0.1$) compared to data with different level of no-cell content (78% for 20% no cell content and 80% 10% no cell content at $\beta = 0.1$), which underlines the need of a scattering snapshot pre-processing for precise unknown cell detection. In fact, a higher training accuracy results in a higher THP-1 cell detection (see THP-1 plot in Fig. 8(a)), but also in a higher unknown cell detection of known sample data (see Test plot in Fig. 8(a)). On the contrary lower network accuracy leads to higher unknown detection, with also higher misclassification of known cells as unknown ones.

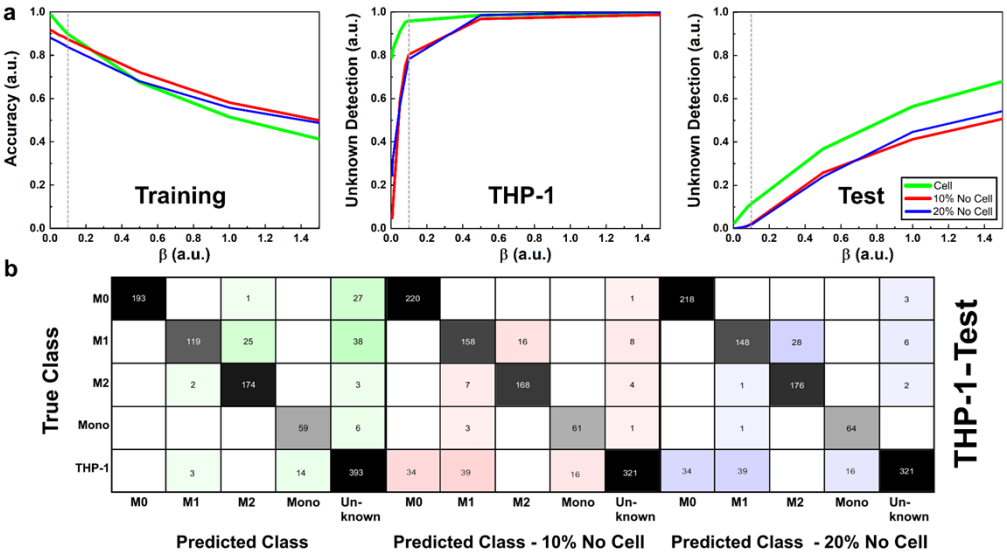

Fig. 8. Open-set outcome for CNN - 2 using an epoch of 15 combined with AOSR using AOSR epoch 1 = 4 and AOSR epoch 2 = 40. (a) Cell class prediction accuracy and unknown (THP-1) cell class detection rate are presented for alternating β and different levels of experimental sample error. The best performing β value for CNN accuracy and unknown cell detection is indicated with a grey dashed line. (b) The confusion matrices of the closed CNN - 2 are shown for different levels of experimental sample error using a snapshot dataset of 221, 182, 179, 65 and 410 cells for M0-, M1-, M2- macrophages, monocytes, and THP-1 respectively.

However, the open-set misclassification with different amounts of no-cell content showed a similar trend to the closed-set architecture, where M1-macrophages are more likely classified as M2-macrophages. A minor misclassification between M1- and M2-macrophages was expected due to the cell differentiation process from peripheral blood monocytes. Furthermore, M1 were more likely classified as unknown cells, followed by M0, while M2 were less represented in unknown cell predictions. This finding is in good agreement with t-SNE investigations (see Fig. 7) and scattering snapshot representations illustrated in Fig. 2.

Biomedical Optics EXPRESS

4. Conclusion

The aim of this work is to demonstrate the high potential of pure light scattering snapshots in CNN-based classification models. Additionally, we report the impact of uncertainty in single cell distinction, when predicting unknown cell classes, which is a relatively unexplored field in life science applications using DL-based classification models. For instance, standard metrics, which evaluate DL models measure the overall model performance on a limited dataset, provide no indication of the model confidence in the correctness of individual predictions of unknown cells, which were not seen during the training phase. Furthermore, models cannot be easily updated after initial model testing. In fact, resulting in prediction failures, when a model is faced with out-of-distribution data.

Therefore, we developed different CNN architectures to predict unknown cell classes from scattering snapshots in concomitance of experimental measurement noise. In addition, in comparison with standard flow cytometric approaches, which are known to have high instrumentation and service costs, the presented measurement method is straightforward and cost-effective, permitting a classification of cell classes without large numbers of training data and resource-intensive cell labelling. More importantly, measurements are realised using a lab-on-a-chip approach permitting the measurement of living cells in suspension, which are also collectable and re-usable for other diagnostic investigations or therapeutic approaches. In fact, the measurement procedure for scattering snapshots was designed to be easily automated and versatile.

We investigated scattering snapshots of four monoblast cell classes (known classes) and an acute monocytic leukaemia cell line (unknown class). First, we trained a classifier model, through which we pre-processed snapshots to separate non-cell (debris) from cell content. This initial image processing step allowed us to subsequently substitute cell snapshots with a specific number of experimental sample noise in the labelled training dataset and analyse its effect on the prediction accuracy. Results show for known classes (seen during the training phase) a high cell class prediction accuracy, using a CNN closed-set architecture. Because unknown cell classes would be mis-classified with a closed-set architecture, we modified the CNN-model and implemented the open-set CNN based AOSR architecture. Such network modification allowed us to detect unknown from known cell classes distribution without significant reduction of known cell class prediction accuracy. The interplay of closed-set accuracy and out-of-distribution recognition was optimised for scattering snapshots, showing high detection accuracy for all investigated cell classes.

In conclusion, we presented a procedure able to label-free predict out-of-distribution cells from scattering snapshots, using an open-set CNN model, which significantly broadens the scope of application for the presented cell signature classification method. Our approach can be easily adapted to accommodate other single cell image data inputs. We firmly believe that the conduction of uncertainty studies at single cell level will revolutionise the trust in cell classification and facilitate circulating tumour cell detection in microfluidics, where no training data is often unavailable.

Acknowledgments. We thank Raffaele Mennella for his proofreading of the manuscript.

Disclosures. The authors declare no conflicts of interest related to this article.

Data availability. Data underlying the results presented in this paper are not publicly available at this time but may be obtained from the authors upon reasonable request.

Supplemental document. See Supplement 1 for supporting content.

References

1. H. T. Maecker and J. P. McCoy Jr, "A model for harmonizing flow cytometry in clinical trials," Nat. Immunol. **11**(11), 975–978 (2010).
2. S. N. Lone, S. Nisar, T. Masoodi, M. Singh, A. Rizwan, S. Hashem, W. El-Rifai, D. Bedognetti, S. K. Batra, M. Haris, A. A. Bhat, and M. A. Machaet, "Liquid biopsy: A step closer to transform diagnosis, prognosis and future of cancer treatments," Mol. Cancer **21**(1), 79 (2022).

Biomedical Optics EXPRESS

3. J. Guck and E. R. Chilvers, "Mechanics meets medicine," Sci. Transl. Med. **5**(212), 212fs41 (2013).
4. 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 Adv. **8**(12), 120701 (2018).
5. P. Rosendahl, K. Plak, A. Jacobi, M. Kraeter, N. Toepfner, O. Otto, C. Herold, M. Winzi, M. Herbig, Y. Ge, S. Girardo, K. Wagner, B. Baum, and J. Guck, "Real-time fluorescence and deformability cytometry," Nat. Methods **15**(5), 355–358 (2018).
6. O. Otto, P. Rosendahl, A. Mietke, S. Golfier, C. Herold, D. Klaue, S. Girardo, S. Pagliara, A. Ekpenyong, A. Jacobi, M. Wobus, N. Töpfner, U. F. Keyser, J. Mansfeld, E. Fischer-Friedrich, and J. Guck, "Real-time deformability cytometry: on-the-fly cell mechanical phenotyping," Nat. Methods **12**(3), 199–202 (2015).
7. D. R. Gossett, H. T. K. Tse, S. A. Lee, Y. Ying, A. G. Lindgren, O. O. Yang, J. Rao, A. T. Clark, and D. Di Carlo, "Hydrodynamic stretching of single cells for large population mechanical phenotyping," Proc. Natl. Acad. Sci. **109**(20), 7630–7635 (2012).
8. M. Masaeli, D. Gupta, S. O'Byrne, H. T. K. Tse, D. R. Gossett, P. Tseng, A. S. Utada, H. J. Jung, S. Young, A. T. Clark, and D. Di Carlo, "Multiparameter mechanical and morphometric screening of cells," Sci. Rep. **6**(1), 37863 (2016).
9. 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).
10. D. Rossi, D. Dannhauser, B. M. Nastri, A. Ballini, A. Fiorelli, M. Santini, P. A. Netti, S. Scacco, M. M. Marino, F. Causa, M. Boccellino, and M. Di Domenico, "New trends in precision medicine: a pilot study of pure light scattering analysis as a useful tool for non-small cell lung cancer (NSCLC) Diagnosis," J. Pers. Med. **11**(10), 1023 (2021).
11. L. Ziegler-Heitbrock, "Monocyte subsets in man and other species," Cell. Immunol. **289**(1-2), 135–139 (2014).
12. D. Min, B. Brooks, J. Wong, R. Salomon, W. Bao, B. Harrisberg, S. M. Twigg, D. K. Yue, and S. V. McLennan, "Alterations in monocyte CD16 in association with diabetes complications," Mediators Inflamm. **2012**, 649083 (2012).
13. J. Sun, L. Wang, Q. Liu, A. Tárnok, and X. Su, "Deep learning-based light scattering microfluidic cytometry for label-free acute lymphocytic leukemia classification," Biomed. Opt. Express **11**(11), 6674–6686 (2020).
14. M. Shifat-E-Rabbi, X. Yin, C. E. Fitzgerald, and G. K. Rohde, "Cell image classification: a comparative overview," Cytometry Part A **97**(4), 347–362 (2020).
15. D. Arifler, C. MacAulay, M. Follen, and M. Guillaud, "Numerical investigation of two-dimensional light scattering patterns of cervical cell nuclei to map dysplastic changes at different epithelial depths," Biomed. Opt. Express **5**(2), 485–498 (2014).
16. X. Su, C. Capjack, W. Rozmus, and C. Backhouse, "2D light scattering patterns of mitochondria in single cells," Opt. Express **15**(17), 10562–10575 (2007).
17. S. K. Yarmoska, S. Kim, T. E. Matthews, and A. Wax, "A scattering phantom for observing long range order with two-dimensional angle-resolved Low-Coherence Interferometry," Biomed. Opt. Express **4**(9), 1742–1748 (2013).
18. A. Merino, L. Puigví, L. Boldú, S. Alférez, and J. Rodellar, "Optimizing morphology through blood cell image analysis," Int. J. Lab. Hematol. **40**, 54–61 (2018).
19. N. Tatsumi and R. V. Pierre, "Automated image processing: past, present, and future of blood cell morphology identification," Clin. Lab. Med. **22**(1), 299–315 (2002).
20. I. Kviatkovsky, A. Zeidan, D. Yeheskely-Hayon, E. L. Shabad, E. J. Dann, and D. Yelin, "Measuring sickle cell morphology during blood flow," Biomed. Opt. Express **8**(3), 1996–2003 (2017).
21. D. Watson, N. Hagen, J. Diver, P. Marchand, and M. Chachisvilis, "Elastic light scattering from single cells: orientational dynamics in optical trap," Biophys. J. **87**(2), 1298–1306 (2004).
22. J. J. Wang, L. Han, Y. P. Han, G. Gouesbet, X. Wu, and Y. Wu, "Shaped beam scattering from a single lymphocyte cell by generalized Lorenz–Mie theory," J. Quant. Spectrosc. Radiat. Transf. **133**, 72–80 (2014).
23. A. Mahdavi and M. Carvalho, "A survey on open set recognition," In *2021 IEEE Fourth International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*, IEEE 37–44 (2021).
24. D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mane, "Concrete problems in AI safety," ´ arXiv, arXiv:1606.06565 (2016).
25. N. A. Smuha, "The EU approach to ethics guidelines for trustworthy artificial intelligence," Computer Law Review International **20**, 97 (2019)..
26. B. Shneiderman, "Bridging the gap between ethics and practice: Guidelines for reliable, safe, and trustworthy human-centered ai systems," ACM Transactions on Interactive Intelligent Systems **10**, 26 (2020).
27. S. Mohseni, H. Wang, Z. Yu, C. Xiao, Z. Wang, and J. Yadawa, "Practical machine learning safety: A survey and primer," arXiv, arXiv,2106.04823, 4 (2021).
28. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," Communications of the ACM **60**, 6 (2012).
29. K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers:Surpassing human-level performance on imagenet classification," *2015 IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, 2015, pp. 1026-1034.
30. W. J. Scheirer, A. de Rezende Rocha, A. Sapkota, and T. E. Boult, "Toward open set recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence **35**, 757 (2013).

Biomedical Optics EXPRESS

31. A. Bendale and T. Boult, "Towards open world recognition," in *IEEE Conference on Computer Vision and Pattern Tecognition*, 2015.
32. D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," arXiv, arXiv:1610.02136 (2017)..
33. S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," Machine Learning **79**, 151 (2010).
34. D. Li, Y. Yang, Y.-Z. Song, and T. M. Hospedales, "Deeper, broader and artier domain generalization," in *Proceedings of the IEEE International Conference on Computer Vision*, (2017).
35. M. Wang and W. Deng, "Deep visual domain adaptation: A survey," Neurocomputing **312**, 135 (2018).
36. 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," in Proceedings of the AAAI conference on artificial intelligence, Vol. 33, No. 01 (2019).
37. C. Geng, S. J. Huang, and S. Chen, "Recent advances in open set recognition: A survey," IEEE Transactions on Pattern Analysis and Machine Intelligence **33**, 3614 (2020).
38. A. Mahdavi and M. Carvalho, "A survey on open set recognition," arXiv, arXiv:2109.00893 (2021).
39. C. C. Aggarwal and P. S. Yu, "Outlier detection for high dimensional data," in *Proceedings of the 2001 ACM SIGMOD international conference on Management of Data*, (2001).
40. V. Hodge and J. Austin, "A survey of outlier detection methodologies," Artificial intelligence review **22**, 85 2004.
41. I. Ben-Gal, "Outlier detection," in *Data Mining and Knowledge Discovery Handbook* (Springer, 2005.)
42. H. Wang, M. J. Bah, and M. Hammad, "Progress in outlier detection techniques: A survey," IEEE Access **7**, 107964 (2019).
43. G. Pang, C. Shen, L. Cao, and A. V. D. Hengel, "Deep learning for anomaly detection: A review," arXiv, arXiv:2007.02500 (2020).
44. S. Bulusu, B. Kailkhura, B. Li, P. K. Varshney, and D. Song, "Anomalous example detection in deep learning: A survey," IEEE Access **8**, 132330 2020.
45. R. Chalapathy and S. Chawla, "Deep learning for anomaly detection: A survey," arXiv, arXiv:1901.03407 (2019).
46. M. A. Pimentel, D. A. Clifton, L. Clifton, and L. Tarassenko, "A review of novelty detection," Signal Processing **99**, 215 (2014).
47. D. Miljkovic, "Review of novelty detection methods," *The 33rd International Convention MIPRO, IEEE* (2010).
48. M. Markou and S. Singh, "Novelty detection: a review—part 1: statistical approaches," Signal Processing **83**, 2481 (2003).
49. M. Markou and S. Singh, "Novelty detection: a review—part 2: neural network based approaches," Signal Processing **83**, 107964 (2003).
50. L. P. Jain, W. J. Scheirer, and T. E. Boult, "Multi-class open set recognition using probability of inclusion," In *Computer Vision–ECCV 2014: 13th European Conference, Proceedings*, Part III 13 (Springer International Publishing, 2014), pp. 393–409 .
51. Z. Ge, S. Demyanov, Z. Chen, and R. Garnavi, "Generative openmax for multi-class open set classification," arXiv, arXiv:1707.07418 (2017).
52. 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).
53. 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), 12666 (2017).
54. D. Dannhauser, D. Rossi, F. Causa, P. Memmolo, A. Finizio, T. Wriedt, J. Hellmers, Y. Eremin, P. Ferraro, and P. A. Netti, "Optical signature of erythrocytes by light scattering in microfluidic flows," Lab Chip **15**(16), 3278–3285 (2015).
55. D. Dannhauser, D. Rossi, A. T. Palatucci, V. Rubino, F. Carriero, G. Ruggiero, M. Ripaldi, M. Toriello, G. Maisto, P. A. Netti, G. Terrazzano, and F. Causa, "Non-invasive and label-free identification of human natural killer cell subclasses by biophysical single-cell features in microfluidic flow," Lab Chip **21**(21), 4144–4154 (2021).
56. D. Dannhauser, D. Rossi, P. Memmolo, A. Finizio, P. Ferraro, P. A. Netti, and F. Causa, "Biophysical investigation of living monocytes in flow by collaborative coherent imaging techniques," Biomed. Opt. Express **9**(11), 5194–5204 (2018).
57. 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," R. Soc. Open Sci. **9**(9), 220270 (2022).
58. D. Dannhauser, F. Causa, E. Battista, A. M. Cusano, D. Rossi, and P. A. Netti, "In-flow real-time detection of spectrally encoded microgels for miRNA absolute quantification," Biomicrofluidics **10**(6), 064114 (2016).
59. M. I. Maremonti, V. Panzetta, D. Dannhauser, P. A. Netti, and F. Causa, "Wide-range viscoelastic compression forces in microfluidics to probe cell-dependent nuclear structural and mechanobiological responses," J. R. Soc. Interface **19**(189), 20210880 (2022).
60. D. Dannhauser, G. Romeo, F. Causa, I. De Santo, and P. A. Netti, "Multiplex single particle analysis in microfluidics," Analyst **139**(20), 5239–5246 (2014).
61. N. Otsu, "A threshold selection method from gray-level histograms," IEEE Trans. Syst., Man, Cybern. **9**(1), 62–66 (1979).
62. J. Šťastný and M. Minařík, "A Brief Introduction to Image Pre-Processing for Object Recognition," (2007).