COMPUTATIONAL ARTICLE

ISAC CYTOMETRY Journal of Quantitative Cell Science PART A

MIFlowCyt

Classification of Human White Blood Cells Using Machine Learning for Stain-Free Imaging Flow Cytometry

Maxim Lippeveld,[1,2] Carly Knill,[3] Emma Ladlow,[3,4] Andrew Fuller,[3] Louise J Michaelis,[5,6] Yvan Saeys,[1,2] Andrew Filby,[3*†] Daniel Peralta[1,2*†]

[1]Data Mining and Modelling for Biomedicine, VIB Center for Inflammation Research, Ghent, Belgium

[2]Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Belgium

[3]Institute of Cellular Medicine, Newcastle University, Newcastle upon Tyne, UK

[4]Newcastle Upon Tyne Hospitals NHS Foundation Trust, Newcastle upon Tyne, UK

[5]Great North Children's Hospital, Newcastle Upon Tyne Hospitals NHS Foundation Trust, Newcastle upon Tyne, UK

[6]Institute of Health and Society, University of Newcastle, Newcastle upon Tyne, UK

Received 21 May 2019; Revised 10 September 2019; Accepted 2 October 2019

Grant sponsor: Fonds Wetenschappelijk Onderzoek, Grant number1SB9419; Grant sponsor: Nvidia; Grant sponsor: Institutional Strategic Support Fund

Additional Supporting Information may be found in the online version of this article.

*Correspondence to: Daniel Peralta Data Mining and Modelling for Biomedicine, VIB Center for Inflammation Research, Ghent, Belgium. Email: daniel.peralta@irc.vib-ugent.be
Andrew Filby Institute of Cellular Medicine, Newcastle University, Newcastle upon Tyne, UK. Email: Andrew.Filby@newcastle.ac.uk

Abstract

Imaging flow cytometry (IFC) produces up to 12 spectrally distinct, information-rich images of single cells at a throughput of 5,000 cells per second. Yet often, cell populations are still studied using manual gating, a technique that has several drawbacks, hence it would be advantageous to replace manual gating with an automated process. Ideally, this automated process would be based on stain-free measurements, as the currently used staining techniques are expensive and potentially confounding. These stain-free measurements originate from the brightfield and darkfield image channels, which capture transmitted and scattered light, respectively. To realize this automated, stain-free approach, advanced machine learning (ML) methods are required. Previous works have successfully tested this approach on cell cycle phase classification with both a classical ML approach based on manually engineered features, and a deep learning (DL) approach. In this work, we compare both approaches extensively on the problem of white blood cell classification. Four human whole blood samples were assayed on an ImageStream-X MK II imaging flow cytometer. Two samples were stained for the identification of eight white blood cell types, while two other sample sets were stained for the identification of resting and active eosinophils. For both data sets, four ML classifiers were evaluated on stain-free imagery with stratified 5-fold cross-validation. On the white blood cell data set, the best obtained results were 0.778 and 0.703 balanced accuracy for classical ML and DL, respectively. On the eosinophil data set, this was 0.871 and 0.856 balanced accuracy. We conclude that classifying cell types based on only stain-free images is possible with all four classifiers. Noteworthy, we also find that the DL approaches tested in this work do not outperform the approaches based on manually engineered features.

Key terms

Imaging flow cytometry; label-free, stain-free, deep learning, machine learning, classification, white blood cells, leukocytes, eosinophils.

Imaging flow cytometry (IFC) produces up to 12 spectrally distinct, information-rich images of single cells at a throughput of up to 5,000 cells per second with a resolution of 0.25 μm per pixel (60× magnification) (1). This includes at least two stain-free image channels capturing transmitted (bright-field) and scattered light (dark-field), and up to 10 images capturing fluorescence emitted by targeted fluorescent stains. These characteristics make IFC an ideal candidate for in-depth analyses of cell populations as an approach to unlock the inherent heterogeneity contained within all biological systems. For example, IFC has been used to detect rare circulating endothelial cells, which have been correlated with various disease states when present in elevated levels (2). Furthermore, IFC allowed for more automation and informative visualizations in the in vitro micronucleus assay, used to study geno and

†Equally contributed.

Published online 5 November 2019 in Wiley Online Library (wileyonlinelibrary.com)

DOI: 10.1002/cyto.a.23920

cytotoxicity (3). IFC has also been used to study the partitioning of molecules across the plane of cell division in a statistically robust manner (4–6).

In this and most other IFC research, cell populations are studied with manual gating on features extracted from the IFC data by specialized software. With manual gating, cells are hierarchically divided into sub-populations by setting boundaries, or *gates*, on 2D scatter-plots of cell measurements. These measurements are usually a combination of fluorescence intensities derived from a stain targeting a cell of interest, and morphological characteristics derived from the stain-free cell images. Although this approach has led to numerous insights into cell population heterogeneity (7), it has some serious drawbacks, mainly:

i. manual gating is hard to reproduce,
ii. manual gating is subjective and biased, and
iii. manual gating is time-consuming for large experiments (8).

Manual gating is an expert-driven process, which introduces two main sources of operator bias. First, gates set on the scatter plots are highly subjective and can therefore differ significantly between operators. The second source is specific to IFC: The choice of which features to compute, and on which area of interest in the image (referred to as *mask*; Fig. 1) to compute them, greatly influence downstream analysis of the data. The operator skill is again an important factor of variability (9).

Analyzing IFC data with manual gating limits the potential of the information-rich, spatially registered data it provides. This is because gating is a bivariate hierarchical analysis done on 2D scatter plots, which allow only two features to be combined at once, whereas a multivariate approach combining a multitude of features can reveal much more intricate patterns in the same data.

Fluorescent stains have drawbacks as well. Firstly, the staining process is expensive and has potential confounding effects on the cells under study, influencing achieved results (10,11). Secondly, usually several stains are required to precisely identify a cell (12), making the experimental workflow labor intensive and slower. Because of these reasons *stain-free* experiments have become of particular interest in the bioimaging field over the last decades (13–15).

A potential solution to overcome these drawbacks is to automate the gating process with machine learning (ML) and do this with only stain-free measurements. This approach (1) combines all available features by using complex, nonlinear ML models, (2) limits operator bias through automation, and (3) potentially obsoletes fluorescent staining by only using stain-free measurements capturing inherent cell morphology.

In previous work, several approaches have been explored to apply ML for the classification of IFC data: Hennig et al. (16) developed an open-source solution, which uses the software package CellProfiler (17) to extract image features from stain-free cell images, and classical supervised ML to classify the cells in subpopulations. They were able to classify Jurkat cells into five phases of the cell cycle.

Another example is the work by Eulenberg et al. (18), who developed a deep learning (DL) model, termed *DeepFlow*. It is able to reconstruct the cell cycle of Jurkat cells, as well as to study the disease progression of diabetic retinopathy. DeepFlow is a convolutional neural network (CNN), which autonomously extracts relevant features from input images to perform a classification, eliminating the requirement for specialized tools to extract features. DL is currently widely used in image classification and is increasingly being adopted in image cytometry.

In this work, we contribute to the problem of stain-free cell classification by extensively comparing both classical ML and DL, testing out two models per approach. This comparison aims at giving clear insight into the value of novel DL techniques in IFC analysis (Fig. 2). This is an important assessment to make given the significant expertise and computational power required to use DL classifiers. Additionally, we validate the work by Eulenberg et al. on DeepFlow and suggest the use of data augmentation to improve performance.

Our classification experiments are run on two high-quality white blood cell data sets from healthy human whole blood samples, acquired on the ImageStream^X MK-II platform. Unlike the work mentioned above, these data sets do not focus on the cell cycle of Jurkat cells, but on the identification of various types of white blood cells (WBC). In addition, the first data set contains specific cell subtypes (for example, CD4+ and CD8 + T-cells). In the second data set, active and resting eosinophils (EOS) are identified. Activation state is of importance as elevated eosinophil activation is linked with allergic disease and intrinsic asthma for example (19,20). The stain-free classification of these more subtle cell types has not been attempted in previous work and is challenging as differences in their stain-free measurements are expected to be less pronounced.

In short, we use two high-quality IFC data sets to assess whether a ML-aided approach can exploit morphological patterns in bright- and darkfield measurements to enable stain-free classification of various cell types.

We systematically compare both a classical and a DL-based approach.

Materials and Methods

Data sets

The data sets used in this work are both acquired from human blood samples. In each data set, a different fluorescent staining panel was used to phenotype immune cells. Fluorescence

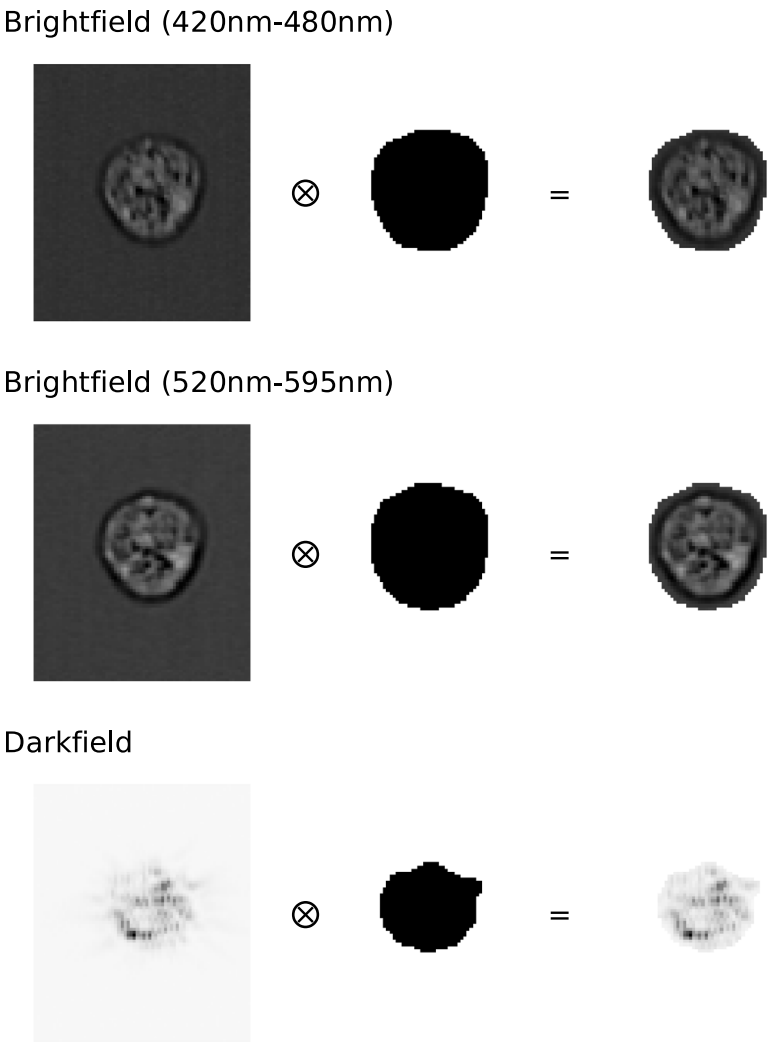

Figure 1. The stain-free brightfield and darkfield images used in this work acquired with the Amnis ImageStream-X MKII for a random cell, images for each channel and accompanying masks are shown, respectively, in the first and second column. The binary masks, which are computed by IDEAS, are combined with the image, setting all background pixels to 0. The images are center-cropped to 90 × 90 pixels to form the final input image for the convolutional neural networks, as seen in the last column. [Color figure can be viewed at wileyonlinelibrary.com]

information was analyzed by expert manual gating to identify the ground truth label for each cell in a sample, akin to phenotyping by conventional flow cytometry. These ground truth labels were used to train ML classifiers, as described later.

Ethical approval to obtain blood from healthy volunteers was granted by the County Durham and Tees Valley Research Ethics Committee (12/NE/0121). All data sets in this work are available upon request.

WBC: This data set contains the measurement results from WBC from two whole blood samples collected into citrate buffer. For phenotyping experiments, 500 µl of whole blood was placed in a 15 ml falcon tube, so that approximately 2×10^6 WBCs were stained with the following antibody cocktail: CD15 FITC (BD, cat no: 332778, clone MMA, 5 µl per test), Siglec8 Pe (Biolegend, cat no: 347104, clone 7C9), CD14 PeCF594 (BD, cat no: 562334, clone Mφ9, 5 µl per test), CD19 PerCP-CY5.5 (BD, cat no 340951, clone SJ25C1, 20 µl per test), CD3 BV421 (BD, cat no: 562426, clone UCHT1, 5 µl per test), CD45 V500 (BD, cat no: 647450, clone 2D1, 5 µl per test), CD4 BV605 (BD, cat no: 562658, clone RPA-T4, 5 µl per test), CD56 APC (BD, cat no: 341025, clone NCAM16.2, 5 µl per test), and CD8 APC-CY7 (BD, cat no: 557834, clone SK1, 5 µl test). Whole blood was incubated with the staining cocktail for 1 h on ice after which red blood cell (RBC) lysis was performed by the addition of 4.5 ml of 1× BD FACS lysis solution (cat no: 349202) prepared from a 10× stock in reagent grade water (SIGMA, cat no: W4502). Lysis was carried out for 10 min at room temperature in the dark. Samples were then spun down at 500 g for 5 min and washed twice in 50 ml of wash buffer (PBS + 2% FBS). Samples were resuspended in a final volume of 60 µl of wash buffer and transferred to 1.5 ml Eppendorf tubes for acquisition.

Eight different white blood cell types were phenotyped with this panel: CD4+/CD8+ T-cells, neutrophils, monocytes,

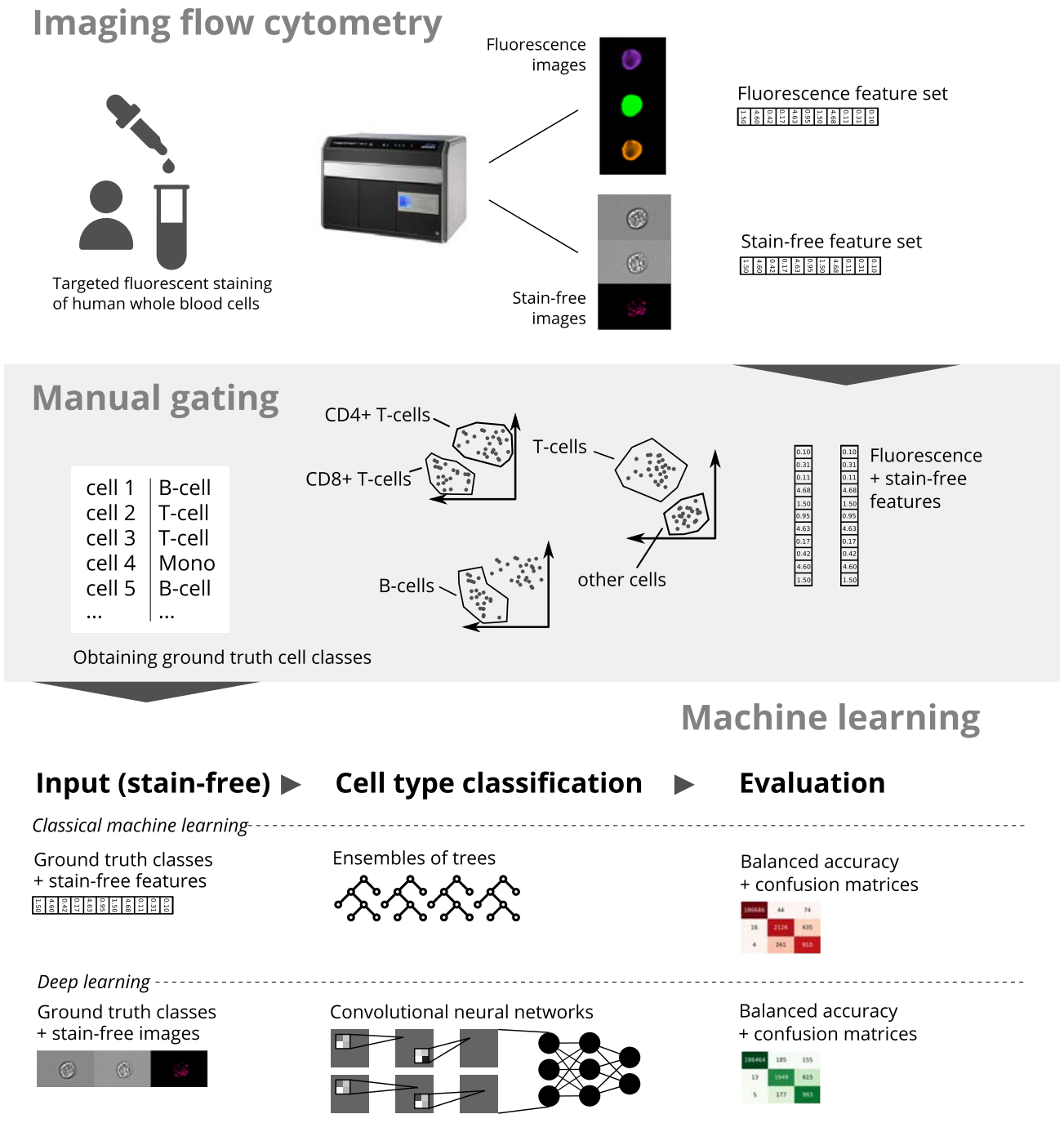

Figure 2. Machine learning enables cell classification based on stain-free imaging flow cytometry imagery. White blood cells from healthy humans are imaged by an imaging flow cytometer, in our case the ImageStream-X MK-II. Features are extracted from stain-free and fluorescence imagery. These features are used in a manual gating procedure to obtain ground truth data. This ground truth and accompanying stain-free images and features are used to train classical machine learning and deep learning models to perform cell classification. [Color figure can be viewed at wileyonlinelibrary.com]

B-cells, CD56+ NKT-cells, other NKT-cells and EOS. See Figure 3 for an overview of the gating process.

The data set is imbalanced: it contains 17,358 CD4+ T-cells, 8,022 CD8+ T-cells, 59,034 CD15+ neutrophils, 2,655 monocytes, 4,256 CD19+ B-cells, 2,214 CD56+ NKT-cells, 1,318 other NKT-cells, and 3,156 EOS.

EOS: This data set contains the measurement results from WBC from 2 whole blood samples collected into Heparin buffer. For eosinophil activation experiments, 1 ml of whole blood was transferred to 15 ml Falcon tubes, one for each of the following conditions (1) Ex-vivo control that was kept on ice for the duration of stimulation, (2) 20 min

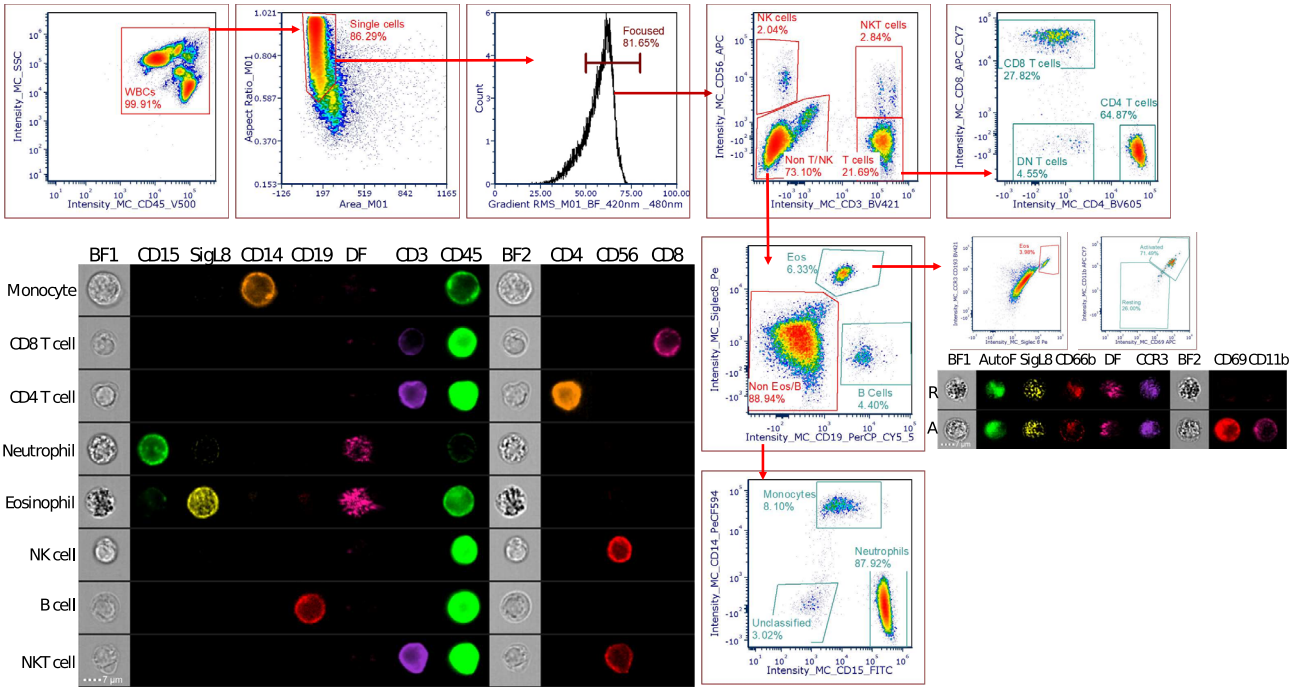

Figure 3. The "ground truth" gating strategy based on fluorescence antibody information. Briefly, WBCs were gated based on CD45 V500 fluorescence and darkfield (DF). Single cells were identified based on the Area and Aspect ratio of the bright-field image in channel (BF1). Focused cells were identified based on the Gradient RMS feature (>50AU). NK, NKT, and T cells were then identified based on CD3 and CD56 fluorescence with the T-cell population further subdivided into CD4 and CD8 subsets. The remaining cells were identified as either B cells (CD19 positivity), Eosinophils (Siglec 8 positivity), Monocytes (CD14 positivity), or Neutrophils (CD15 positivity). Example multispectral, compensated images are shown for each class of immune cell identified at 60× magnification. It should also be noted that for ML/DL, cells were always compensated and samples were preprocessed to the stage of CD45 positive, single in focus cells, as shown above. [Color figure can be viewed at wileyonlinelibrary.com]

stimulation, (3) 40 min stimulation, (4) 60 min stimulation, (5) unstimulated control, incubated for the duration of stimulation. In the first instance, stimulations were performed using PMA/Ionomycin (eBiosciences Cell Stimulation Cocktail, cat no: 00–4,970-03) at a 1× working concentration. In order to ensure all incubations ended at the same time, the 60 min stimulation was started first, then the 40 min stimulation, 20 min later, and finally the 20 min stimulation a further 20 min later. At the end of the stimulation period, the samples were divided into two 15 ml Falcon tubes (500 µl in each). One sample set was stained with the following antibody cocktail: Siglec8 Pe (Biolegend, 5 µl per test), CCR3/CD192 BV421 (Biolegend, cat no: 310714, clone 5E8, 5 µl per test), CD69 APC (BD, cat no: 555533, clone FN50, 20 µl per test), and CD11b (Biolegend, cat no: 101226, clone M1/70, 5 µl per test).

The other set of samples were left unstained to control for the effects of antibody labelling. Samples were incubated for 1 h on ice after which RBC lysis was performed as described for the WBC panel. Again samples were washed 2 times in wash buffer and finally resupsended in 60 µl of the same for transfer into 1.5 ml Eppendorf tubes prior to acquisition.

Active and resting EOS were phenotyped with this panel. See Figure 3 for an overview of the gating process.

The data set is imbalanced: it contains 1,291 active and 2,595 resting EOS, and 186,671 non-EOS.

Acquisition Details for ImageStream^X MKII Imaging Flow Cytometer

IFC was performed using an ImageStream^x MKII (Luminex Corporation, Seattle, WA) system. The system was fully calibrated with ASSIST (Automated Suite of System-wide ImageStream^X Tests) using the INSPIRE software. ASSIST performs specific calibrations and tests, measuring, evaluating and storing thousands of values to ensure all subsystems are operating within normal limits. It is run daily to ensure optimal performance of the ImageStream^X instrument (21).

The following acquisition configuration was used: 100 mW 488 nm blue laser, 120 mW 450 nm violet laser, and a 642 nm 150 mW red laser. In all cases, maximum excitation laser powers were employed in order to achieve best signal to noise without any pixel saturation (raw max pixel values below 4,096). Bright-field imagery was collected using an LED array with wavelengths of 420 nm to 480 nm in channel 1 (brightfield 1) and 570 nm to 595 nm in channel 9 (brightfield 2). Side scatter was collected in channel 6 using a dedicated 758 nm laser, again set to maximize signal and avoid saturation. FITC emission was collected in CH2, Pe in CH3, PeCF594 in CH4, PerCP-CY5.5 in CH5, BV421 in CH7, V500 in CH8, BV605 in CH10, APC in CH11, and APC-CY7 in CH12. The images were acquired in the highest sensitivity mode and using the 60× magnification.

Each of the stained and unstained samples were acquired with excitation lasers on and off to control for any potential residual fluorescence spill-over into label-free channels after spectral correction. Default spectral correction was performed using the built-in wizard in the IDEAS analysis software package. This includes a correction based on the flat-field and dark-current calibration values obtained from the daily ASSIST tests. There is also the application of the compensation matrix derived from the single stained bead controls: Antibody capture beads (ABC total capture beads, Thermo scientific, cat no: A10513) were used to prepare single stained controls by adding 1 drop of positive, 1 drop of negative, and then 1 test amount of each individual antibody per tube. These controls were acquired with the bright-field and side scatter illumination turned off in order to generate the spill-over matrix. This matrix was then applied to fully stained samples.

Classification Models

Models in this work were divided into two categories: (1) models which take precomputed manually engineered features as input, and (2) models which take images as input. The latter type of models automatically learns and extracts required features from the input images. These models, usually CNNs, have been applied successfully in many image classification tasks (22–24). Using models with automatically engineered features further reduces the influence of expert knowledge on the gating process, at the cost of an increase in the computational effort required to train the model.

In total four classification models were tested, two per model category. For the first model category, referred to as classical ML, we tested a random forest (RF) (25) and gradient boosting (GB) classifier (26). Both models are ensembles of weak decision trees, which are widely used and successful in classification settings (27,28). The number of used trees was set to 500 for random forest and 100 for gradient boosting. All other hyper-parameters were kept at default values provided by the scikit-learn library (29).

For the second model type, we tested two DL CNN architectures: ResNet18 (RN) (23) and DeepFlow (DF) (30).

ResNet is a state-of-the-art architecture in image classification. It eases optimization of the network's weights by reformulating convolutional layers as learning residual functions, with reference to the layer inputs. ResNet has obtained the first place on the ImageNet Large-Scale Visual Recognition Challenge 2015, an important competition in the field of computer vision (23). In this work, a variant of 18 layers deep is used.

DeepFlow is an adaptation of the Inception architecture (31), optimized for classification of IFC data. It was previously applied to reconstruct the cell cycle of Jurkat cells, as well as to study the disease progression of diabetic retinopathy, using stain-free IFC imagery.

The DL models were trained for 100 epochs with the Adam optimization algorithm (32). Adam is an adaptive learning rate optimizer. This type of optimizer risks getting stuck in a local optimum in the initial optimization phase. To avoid this, we used a warm-up phase (33,34): for 6 epochs the learning rate was set to 10^{-5}. Afterward, the learning rate was increased to 10^{-3} and 10^{-4} for white blood cell and eosinophil classification, respectively. To allow the optimization to converge, the learning rate was reduced with a factor of 0.8 each time a monitored validation metric stopped improving for 5 epochs. The resulting learning rate schedules are shown in Supporting Information Figures S3 and S2. L2-regularization was also applied with recommended weights of 10^{-4} for ResNet18, and 5×10^{-4} for DeepFlow. For efficiency reasons, early stopping was implemented, as well: if a monitored metric did not improve for 20 epochs training was interrupted. The early stopping and learning rate reduction metric was computed on a held-out set of samples.

Both DL models were implemented in the Keras-Tensorflow Python library (version 1.13.1) (35). DeepFlow implementation was based on code from the DeepFlow Github Repository.[1] ResNet18 implementation was taken from the Keras-ResNet Github Repository.[2] All code used to generate results in this work is publicly available on Github.[3]

Data Preparation

To run the experiments, we needed all stain-free imagery and accompanying masks, features computed on stain-free imagery, and ground truth cell type labels.

After spectral compensation (see Supporting Information Table S3 for compensation matrix), the IDEAS software produces one compensated image file (CIF) per biological sample, containing compensated imagery for all channels, as well as the accompanying masks. These masks indicate the area of the image containing only the pixels of interest in a certain channel. In the case of brightfield images, the mask typically encloses the entire cell. By masking the images, we avoid influence of background noise or irrelevant information surrounding the cell (36) (see Fig. 1).

In order to work efficiently with these images and masks, they were read and decoded from the CIF format using the Python Bio-Formats library (37) and saved to an HDF5 data set. As the CNNs require all images to have uniform dimensions, each channel and corresponding mask for each image was also center-cropped to 90 by 90 pixels. To perform these steps a custom command line tool was written in Python. By decoding the images once and saving them in decoded form, we can significantly speed up the training of the neural network, as decoding images is a costly operation. The code for this tool is made publicly available on GitHub.[4]

After preparation of the images the final input dimension to the CNNs was $90 \times 90 \times 3$. The three input channels were brightfield 1, brightfield 2, and darkfield. Before being fed to the CNN, masks were applied to the images to set all background pixels to 0 (see Fig. 1).

[1] https://github.com/theislab/deepflow, last accessed on 7th of May 2019.
[2] https://github.com/raghakot/keras-resnet, last accessed on 7th of May 2019.
[3] https://github.com/saeyslab/DeepLearning_for_ImagingFlowCytometry, last accessed on 7th of May 2019.
[4] http://github.com/saeyslab/cifconvert, last accessed on 7th of May 2019.

The manually engineered feature set was computed in IDEAS. The software computes 76 base features per image channel that are divided into five categories: size, location, shape, texture, and signal strength. Detailed feature definitions can be found in the IDEAS documentation (36). An overview of all features used in this work is given in Supporting Information Table S4.

Data Augmentation

Many classifiers, including the ones used in this work, are sensitive to class imbalance (38). Therefore, we augmented the data sets before training in order to balance the class occurrence frequencies. For the two classical ML algorithms considered in this article, this was done by randomly oversampling minority classes.

An image-based data augmentation approach is implemented for the CNN classifiers. As is commonly done for CNNs, we applied random *label-saving* image transformations to existing training instances, supplementing the data manifold (22,39). We used random horizontal or vertical flips, rotations (with a randomly chosen angle between -180° and 180°), and translations (with a randomly chosen amount of pixels in the x- and y-direction between -6 and 6).

Eulenberg et al. do not apply data augmentation in their work on DeepFlow (30). However, in Supporting Information Figure S1, we show that this leads to poor generalization performance on our data sets. Therefore, we apply data augmentation during training for the remainder of our experiments.

Model Validation

To validate the classification performance of the trained models, a stratified 5-fold cross validation (CV) strategy is used. For each fold, training data were augmented to balance the class occurrence frequencies and used to train a model. The model was then validated using non-augmented instances from the validation set. The predictions made on the validation data were summarized in a confusion matrix per fold. The obtained matrices were summed together, giving one representative confusion matrix per CV experiment.

Together with the confusion matrix, the *balanced accuracy* was reported. The balanced accuracy is the arithmetic mean of class-specific accuracies. It can be computed from the confusion matrix and is formalized as follows:

$$\frac{1}{n}\sum_{i=1}^{n}\theta_i \qquad (1)$$

where θ_i is the class-specific accuracy, and n is the number of classes. The balanced accuracy is suited for imbalanced validation sets, as it does not suffer from the accuracy paradox. This means it will not favor a classifier that exploits class imbalance by biasing toward the majority class (40).

Visualizing Feature Spaces

Dimensionality reduction techniques can be used to give an insight into high-dimensional spaces, by projecting them onto a low-dimensional space. In this work, we applied Uniform Manifold Approximation and Projection (UMAP) (41) on the high-dimensional, manually engineered feature space exported from IDEAS, and the feature space automatically learned by the DeepFlow CNN.

UMAP provides scalable dimensionality reduction, which preserves global and local structures of the high-dimensional input space. It does so by converting high- and low-dimensional representations of the input to topological representations, and then minimizing the cross-entropy between them. We choose this method over others such as t-SNE, due to its scalability and wide-spread application in bioinformatics (41).

The feature space learned by a CNN is encoded by the intermediate activation pattern following the last convolutional layer of the network, referred to as the *code*. The code is the representation of an input image, which is fed to the fully-connected layers of the CNN that perform the actual classification. The codes are extracted from the network by forward-propagating images through it, and recording their corresponding codes (Fig. 4).

All ML experiments were run on a 12-core machine, with an Intel Xeon CPU (model E5-1650 v2) running at 3.50GHz. The machine has 64 GB of RAM. DL experiments were run on an NVIDIA Titan X GPU with 12 GB of VRAM. Code used for extracting imagery and masks from the CIFs, and for training and validation of the models is made public on GitHub at https://github.com/saeyslab/DeepLearning_for_ImagingFlowCytometry.

Results

We started by setting a baseline classification performance on the EOS and WBC data sets, using well-established models, trained on expert driven, manually engineered features. We then applied DL to the same classification tasks and found that they achieved baseline performance for the EOS data set, but not for the WBC data set.

Classifying Cell Types with Manually Engineered Features

Both classical ML classifiers were able to classify cell types based on manually engineered features (see Figs. 5 and 6). For the WBC data set, especially neutrophils, monocytes, and EOS were accurately classified by both classifiers with recalls, respectively, higher than 0.981, 0.955, and 0.965 for all classifiers. For the EOS data set, separation between non-EOS and EOS was very accurate, with recalls respectively higher than 0.998 and 0.990 (computed by treating active and resting EOS as one class).

Both classifiers struggled to reliably subtype cell types. Recalls for the subtypes are consistently lower than for other classes (see Supporting Information Tables S1 and S2. For the WBC data set, confusion was present between CD4+ and CD8+ T-cells for example, as well as between T-cells and NKT-cells (see Fig. 5). Also for the EOS data set, confusion was present between active and resting EOS (see Fig. 6).

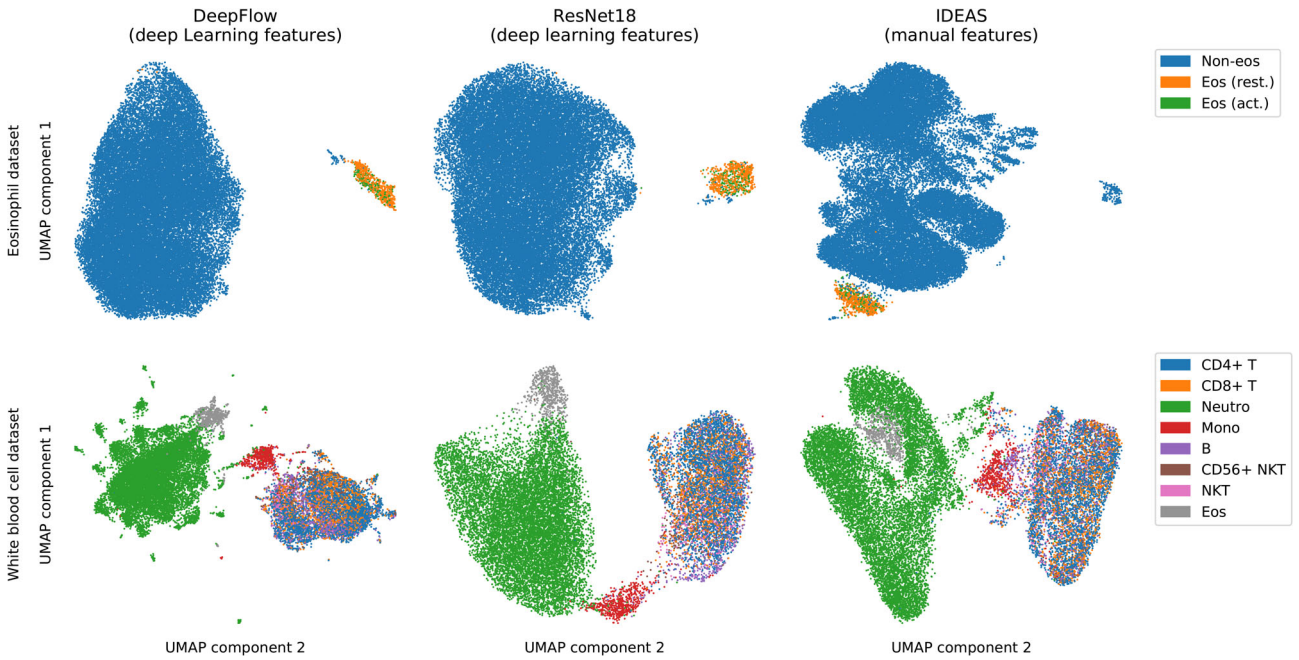

Figure 4. Dimensionality reduction of manually and automatically engineered feature spaces confirmed confusion in cell type classification. High-dimensional feature spaces were projected to a 2D space using uniform manifold approximation and projection. Data points were plotted in the 2D space and colored according to cell type. This revealed that cells of the same cell type cluster together. Clusters of cells types that overlapped were also found challenging to distinguish in the classification experiments. For example, CD4+ and CD8+ T-cells overlapped significantly in the white blood cell data set and also show high confusion in the classification. The same is seen for the active and resting eosinophils in the eosinophil data set.

We found that the GB and RF classifier behaved similarly, with a more pronounced advantage for the GB classifier on the EOS data set. Their respective balanced accuracies were 0.774 versus 0.778 for the WBC data set, and 0.825 versus 0.871 for the EOS data set (see Figs. 5 and 6). The main advantage of the GB over the RF classifier was the better subtyping performance, which is clearly seen in the classification of active versus resting EOS, and CD4+ versus CD8+ T-cells.

Automating Feature Extraction with Deep Learning

DL classifiers are able to autonomously extract relevant information from stain-free imagery to classify WBCs and EOS (see Figs. 5 and 6). Their performance differed between both data sets: for the EOS data set, DeepFlow improved slightly upon the baseline performance set by both classical methods (see Fig. 6). However, for the WBC data set none of the DL classifiers reached baseline performance (DF: 0.703, versus GB: 0.778 balanced accuracy) (see Fig. 5).

As with the classical models, the DL models did not reach satisfactory performance for cell subtyping. For the WBC data set, recalls for CD4+ and CD8+ T-cells did not exceed 0.476 and 0.304, respectively. NKT-cell subtyping suffered less of a drop in recall compared to classical methods, with recall values reaching 0.683 and 0.599 for CD56+ and other NKT-cells, respectively.

Overall, the DF architecture outperformed the RN architecture. The difference was most pronounced for the WBC data set (RN: 0.649, versus DF: 0.703 balanced accuracy). Recalls for all cell types were higher for DF. The biggest improvement over RN occurred in CD4+ T-cell classification (RN: 0.333, versus DF: 0.476 recall) and CD56+ NKT-cell classification (RN: 0.554, versus DF: 0.684 recall). For the EOS data set, the improvement from DF over RN was smaller (RN: 0.831, versus DF: 0.871 balanced accuracy).

Comparing Feature Spaces with Uniform Manifold Approximation and Projection

Visualizing the feature spaces on which the classifiers are trained, provided a visual validation of classification confusion occurring between certain cell types (see Fig. 4). UMAP clustered cells of similar cell types together in the low-dimensional representation. We found that clusters of cell types overlapped for the types with which classifiers struggled. For example, for the WBC data set, confusion occurred between CD4+ and CD8+ T-cells (see Fig. 5). This is clearly reflected in the UMAP visualization by the overlapping clusters of CD4+ and CD8+ T-cells, both for automatically and manually engineered features. On the other hand, accurately classified cell types, such has the EOS, were also well separated in the low-dimensional space. The same occurred for the EOS data set for the confusion between active and resting EOS (see Fig. 4).

Discussion

Manual gating in its current form has three main drawbacks: manual gating is hard to reproduce, it is subjective and

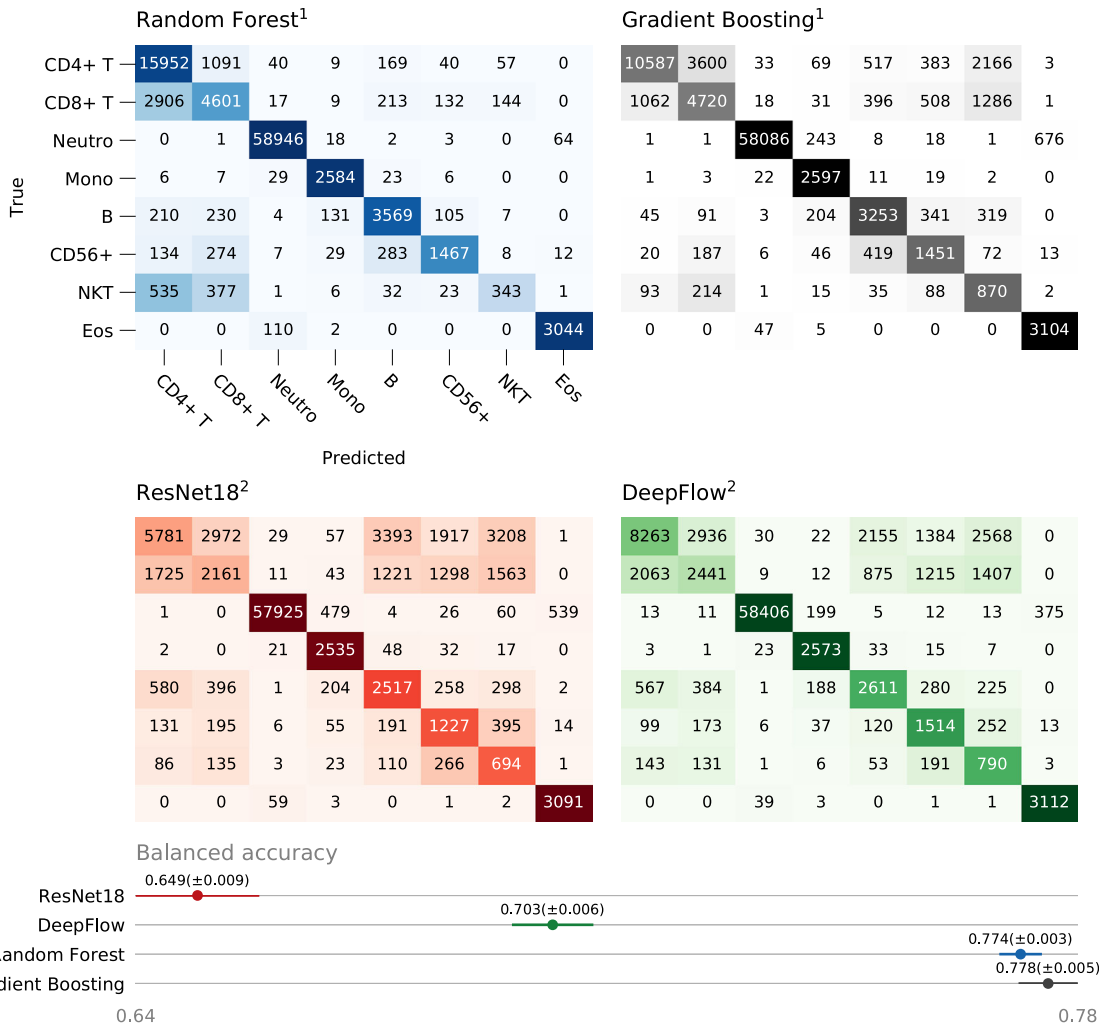

Figure 5. Confusion matrices and mean balanced accuracy scores with standard deviation (indicated by bars) obtained from cross-validation experiments on the white blood cell data set. As seen in the confusion matrices, neutrophils, monocytes, and eosinophils are consistently classified correctly by all classifiers. All classifiers confuse T-cells and NKT-cells. In terms of balanced accuracy, the classical machine learning classifiers (1) behave similarly and outperform the deep learning classifiers (2) by a relatively large margin. In the confusion matrices, we can see that this is mainly due to better T-cell classification by the classical approaches. [Color figure can be viewed at wileyonlinelibrary.com]

biased, and it is time-consuming for large experiments. To overcome these drawbacks, we extensively studied and compared several ML approaches, which only make use of stain-free information to perform automated cell classification.

In the experimental setting of this work, stains do not seem vital for classifying cell types such as monocytes or neutrophils. However, to reliably subtype cells, stains are still required. This is shown in Figure 5 and Supporting Information Table S1. The best performing model, a gradient boosting classifier, has an overall T-cell recall of 0.787, but for the CD4+ and CD8+ T-cell subtypes, recall rates drop to 0.609 and 0.588, respectively. This means that visually, based on the stain-free images, no distinction can be made between CD4+ and CD8+ T-cells. The same pattern can be observed in Figure 6 and Supporting Information Table S2 for active and resting EOS. Other works have concluded also that classifying distinct cell types using stain-free information is possible (42), as well as classifying cell cycle phases (43). To the best of our knowledge, this contribution is the first attempt to classifying cell subtypes using purely stain-free information.

The unreliable cell subtyping might be attributed to two data-related issues: class imbalance and low image resolution. Firstly, all data sets in this work suffer from class imbalance. For example, the EOS data set contains about 187,000 non-EOS, and only about 3,900 EOS, with a 30 versus 70% ratio of active and resting EOS. Especially in DL settings, large and balanced data sets are important, as overfitting occurs regularly (44). In this work, we employ basic data augmentation techniques to counter this problem, which seem to have improved performance to a certain degree. Since acquiring more data is not always an option, developing or employing more advanced data augmentation techniques could improve performance.

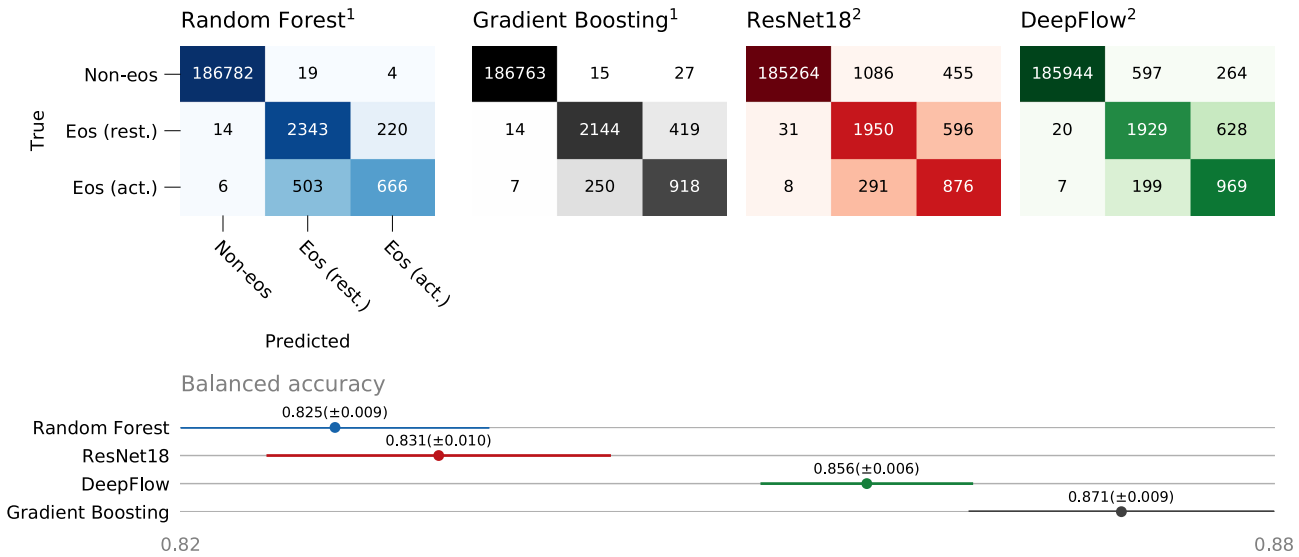

Figure 6. Confusion matrices and mean balanced accuracy scores with standard deviation (indicated by bars) obtained from cross-validation experiments on the eosinophil data set. As seen in the confusion matrices, separation between eosinophils and non-eosinophils is consistently done correctly by all classifiers. Confusion is present in all classifiers when separating between active and resting eosinophils; especially the random forest struggled to make a good distinction. Unlike results of the white blood cell data set, classical classifiers (1) do not outperform deep learning classifiers (2) unanimously. [Color figure can be viewed at wileyonlinelibrary.com]

Secondly, because of the relatively low-image resolution of current imaging flow cytometers, necessary data for cell subtyping is potentially not captured in the stain-free imagery. A solution, which does not eliminate but reduces the necessary staining, might be to design hybrid experiments. These would rely on stain-free information to identify cell types and use a limited amount of stains to further subtype cells. For example, in the WBC data set lymphocytes (NKT-, T-, and B-cells), granulocytes (neutrophils and EOS), and monocytes can be accurately separated using only stain-free information with the models trained in this work. The fluorescence channels that are normally reserved for identifying these larger populations can then be employed for reliably subtyping cells using targeted staining, therefore increasing the possible number of cell populations that can be identified with the same instrumentation.

DL approaches are able to autonomously extract relevant information from stain-free imagery, a conclusion that is supported by previous work (30). However, they do not outperform the classical approaches, which achieve the best results for both data sets. For the WBC dataset, the improvement over DL approaches was fairly large (GB: 0.778 vs DF: 0.703 balanced accuracy); for the EOS data set, it was smaller (GB: 0.871 vs. DF: 0.856). We could state that in this regard the tested DL approaches cannot improve classification performance on the tasks tackled in this work.

The lack of improvement of DL approaches over classical ones could be attributed to a diverse set of reasons. It could be due to purely technical reasons: insufficient or too imbalanced data, a use of too simple or complex architectures, choice of optimizer, and so on. Or, it could be that the expert knowledge embedded in the manual feature engineering is truly vital for stain-free classification. Therefore, if possible, we advise to not only rely on DL approaches when tackling classification tasks like the ones discussed in this work.

However, DL remains an interesting tool that makes cell classification a less manual, and expert-driven process, as no manual feature engineering is required. We must note that optimizing neural networks is not straight forward. It requires the user to overcome obstacles such as overfitting, hyperparameter tuning, handling big data, and dealing with a shortage of, or imbalance in labeled data (18,45). Different methods to deal with these issues, such as transfer learning or data augmentation, need to be made accessible and easy to use. Therefore, we have made our code and trained models publicly available on Github.[5] In order to increase accessibility of this solution, we consider developing an extension for the CellProfiler software, which has gained considerable popularity in the bio-imaging field.

An interesting difference between automatically and manually engineered feature spaces is shown by the UMAP dimensionality reduction. It shows that the manually engineered feature set is better suited for exploring the heterogeneity of cells within a data set. This is because the automatically engineered feature spaces from the DeepFlow and ResNet18 CNNs are only optimized to distinguish between the ground truth cell types in the data set. On the other hand, the manually engineered feature set from IDEAS is more general. This is demonstrated by the UMAP reductions for the EOS data set in Figure 4: the non-EOS are clustered in one homogeneous cluster in the DL feature space, whereas several clusters can be distinguished within the non-EOS in the

[5] https://github.com/saeyslab/DeepLearning_for_ImagingFlowCytometry, last accessed on 7th of May 2019.

IDEAS feature space. The IDEAS features therefore seem to capture more of the heterogeneity within this population.

File formats produced by the Amnis ImageStream platform are closed source, and therefore unsuitable for data science applications. In previous work, an approach was proposed, which requires the user to create many image montages from the images in the original CIF, using a custom script (16). These montages can then be processed by image analysis software, such as CellProfiler. This is a cumbersome and non-user-friendly process. In this work, we have accommodated for this inconvenience by writing a script that decodes images and masks from the original CIF and stores them in one HDF5 data set to be used during further processing. This way we have significantly reduced processing time of the CIFs, opening the possibility to train and test ML models on hundreds of samples. The script is publicly available on Github.[6]

In conclusion, we have found that IFC lends itself well to ML applications due to its information rich data and high-throughput nature. We have shown that besides cell cycle phase classification, white blood cell type classification is also feasible for certain immune cell types, creating the potential to apply this approach in immunodeficiency diagnosis, for example. For the data sets and classification methods studied here we conclude that the limit of this classification approach currently lies at the level of the cell type, and that subtyping remains a challenge for future work.

Acknowledgments

M. Lippeveld is a Predoctoral Fellow of the Fund for Scientific Research FWO Flanders. D. Peralta is a Postdoctoral Fellow of the Research Foundation of Flanders. Y. Saeys is an ISAC Marylou Ingram Scholar. This work is supported by the Wellcome Trust Institutional Strategic Support Fund, and the NVIDIA GPU Grant Program

Conflicts of Interest

The authors have no conflict of interest to declare.

Literature Cited

1. Luminex. Amnis Imaging Flow Cytometers. Austin, TX: Luminex, 2019.
2. Samsel L, Philip McCoy J. Detection and characterization of rare circulating endothelial cells by imaging flow cytometry. In: Barteneva NS, Vorobjev IA, editors. Imaging Flow Cytometry: Methods and Protocols, Methods in Molecular Biology. New York, NY: Springer New York, 2016; p. 249–264.
3. Rodrigues MA. Automation of the in vitro micronucleus assay using the Imagestream R imaging flow cytometer. Cytometry A 2018;93(7):706–726.
4. Thaunat O, Granja AG, Barral P, Filby A, Montaner B, Collinson L, MartinezMartin N, Harwood NE, Bruckbauer A, Batista FD. Asymmetric segregation of polarized antigen on B cell division shapes presentation capacity. Science 2012;335(6067):475–479.
5. Filby A, Perucha E, Summers H, Rees P, Chana P, Heck S, Lord GM, Davies D. An imaging flow cytometric method for measuring cell division history and molecular symmetry during mitosis. Cytometry A 2011;79A(7):496–506.
6. Hawkins ED, Oliaro J, Axel Kallies GT, Belz AF, Hogan T, Haynes N, Ramsbottom KM, Van Ham V, Kinwell T, Seddon B, et al. Regulation of asymmetric cell division and polarity by scribble is not required for humoral immunity. Nat Commun 2013;4(1801).
7. Doan M, Vorobjev I, Rees P, Filby A, Wolkenhauer O, Goldfeld AE, Lieberman J, Barteneva N, Carpenter AE, Hennig H. Diagnostic potential of imaging flow cytometry. Trends Biotechnol 2018;36(7):649–652.
8. Saeys Y, Van Gassen S, Lambrecht BN. Computational flow cytometry: Helping to make sense of high-dimensional immunology data. Nat Rev Immunol 2016;16(7):449–462.
9. Filby A, Davies D. Reporting imaging flow cytometry data for publication: Why mask the detail? Cytometry A 2012;81A(8):637–642.
10. Wojcik K, WDobrucki J. Interaction of a DNA intercalator DRAQ5, and a minor groove binder SYTO17, with chromatin in live cells–influence on chromatin organization and histone-DNA interactions. Cytometry 2008;73(6):555–562.
11. Chen AY, Yu C, Gatto B, Liu LF. DNA minor groove-binding ligands: A different class of mammalian DNA topoisomerase I inhibitors. Proc Natl Acad Sci 1993;90 (17):8131–8135.
12. Miltenburger HG, Sachse G, Schliermann M. S-phase cell detection with a monoclonal antibody. Dev Biol Stand 1987;66:91–99.
13. Freudiger CW, Min W, Saar BG, Lu S, Holtom GR, He C, Tsai JC, Kang JX, Xie XS. Label-free biomedical imaging with high sensitivity by stimulated Raman scattering microscopy. Science 2008;322(5909):1857–1861.
14. Wang S, Shan X, Patel U, Huang X, Lu J, Li J, Tao N. Label-free imaging, detection, and mass measurement of single viruses by surface plasmon resonance. Proc Natl Acad Sci 2010;107(37):16028–16032.
15. de Wit G, Danial JSH, Kukura P, Wallace MI. Dynamic label-free imaging of lipid nanodomains. Proc Natl Acad Sci 2015;112(40):12299–12303.
16. Hennig H, Rees P, Blasi T, Kamentsky L, Hung J, Dao D, Carpenter AE, Filby A. An open-source solution for advanced imaging flow cytometry data analysis using machine learning. Methods (San Diego, Calif) 2017;112:201–210.
17. Kamentsky L, Jones TR, Fraser A, Bray M-A, Logan DJ, Madden KL, Ljosa V, Rueden C, Eliceiri KW, Carpenter AE. Improved structure, function and compatibility for CellProfiler: Modular high-throughput image analysis software. Bioinformatics 2011;27(8):1179–1180.
18. Gupta A, Harrison PJ, Wieslander H, Pielawski N, Kartasalo K, Partel G, Solorzano L, Suveer A, Klemm AH, Spjuth O, et al. Deep learning in image cytometry: A review. Cytometry A 2018;95:366–380.
19. Frew AJ, Kay AB. The relationship between infiltrating CD4+ lymphocytes, activated eosinophils, and the magnitude of the allergen-induced late phase cutaneous reaction in man. J Immunol 1988;141(12):4158–4164.
20. Bentley AM, Menz G, Storz C, Robinson DS, Bradley B, Jeffery PK, Durham SR, Kay AB. Identification of T lymphocytes, macrophages, and activated eosinophils in the bronchial mucosa in intrinsic asthma: Relationship to symptoms and bronchial responsiveness. Am Rev Respir Dis 1992;146(2):500–506.
21. Amnis. INSPIRE - ImageStream-X MKII User Manual. Seattle, WA: Amnis, 2017.
22. Krizhevsky A, Sutskever I. And Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ, editors. Advances in Neural Information Processing Systems. Volume 25. Red Hook: Curran Associates, Inc, 2012; p. 1097–1105.
23. He K, Zhang X, Ren S, and Sun J. Deep residual learning for image recognition. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE: Las Vegas, NV. 2016, pp 770–778.
24. Rezvantalab A, Safigholi H, and Karimijeshni S. Dermatologist level dermoscopy skin cancer classification using different deep learning convolutional neural networks algorithms. *arXiv: 1810.10348 [cs]*, Stat], October 2018.
25. Breiman L. Random forests. Mach Learn 2001;45(1):5–32.
26. Friedman JH. Stochastic gradient boosting. Comput Stat Data Anal 2002;38(4): 367–378.
27. Díaz-Uriarte R, de Andrés SA. Gene selection and classification of microarray data using random forest. BMC Bioinformatics 2006;7(1):3.
28. Pal M. Random forest classifier for remote sensing classification. Int J Remote Sens 2005;26(1):217–222.
29. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al. Scikit-learn: Machine learning in Python. J Mach Learn Res 2011;12(Oct):2825–2830.
30. Eulenberg P, Köhler N, Blasi T, Filby A, Carpenter AE, Rees P, Theis FJ, Wolf FA. Reconstructing cell cycle and disease progression using deep learning. Nat Commun 2017;8(1).
31. Szegedy C, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. Going deeper with convolutions. In 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015pp 1–9.
32. Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv:*1412.6980 [cs]*, December 2014.
33. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L. Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R, editors. Advances in Neural Information Processing Systems. Volume 30. Red Hook: Curran Associates, Inc, 2017; p. 5998–6008.
34. Popel M, Bojar O. Training tips for the transformer model. Prague Bull Math Linguist 2018;110(1):43–70.
35. Martın Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. Technical Report, 2015. Software available from tensorflow.org.
36. Amnis. IDEAS—Image Data Exploration and Analysis Software, November 2015.
37. Kamentsky L. Python Bio-Formats (version 1.5.2). Python. Broad Institute, 2019.
38. Japkowicz N, Stephen S. The class imbalance problem: A systematic study. Intell Data Anal 2002;6(5):429.
39. Pereira S, Pinto A, Alves V, Silva CA. Brain tumor segmentation using convolutional neural networks in MRI images. IEEE Trans Med Imaging 2016;35(5): 1240–1251.

[6] https://github.com/saeyslab/cifconvert, last accessed on 7th of May 2019.

40. K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann. The balanced accuracy and its posterior distribution. In 2010 20th International Conference on Pattern Recognition, Pages 3121–3124, August 2010.

41. McInnes L, John H. UMAP: Uniform manifold approximation and projection for dimension reduction. arXiv:180203426 [cs, stat], February 2018.

42. Meng N, Lam E, Tsia KKM, So HK. Large-scale multi-class image-based cell classification with deep learning. IEEE J Biomed Health Inform 2019:2091–2098.

43. Blasi T, Hennig H, Summers HD, Theis FJ, Cerveira J, Patterson JO, Davies D, Filby A, Carpenter AE, Rees P. Label-free cell cycle analysis for high-throughput imaging flow cytometry. Nat Commun 2016;7:10256.

44. Angermueller C, Pärnamaa T, Parts L, Stegle O. Deep learning for computational biology. Mol Syst Biol 2016;12(7):878.

45. Ching T, Himmelstein DS, Beaulieu-Jones BK, Kalinin AA, Do BT, Way GP, Ferrero E, Agapow P-M, Zietz M, Hoffman MM, et al. Opportunities and obstacles for deep learning in biology and medicine. J R Soc Interface 2018; 15(141).