Position Papers of the 18th Conference on Computer Science and Intelligence Systems pp. 59–66 DOI: 10.15439/2023F8069 ISSN 2300-5963 ACSIS, Vol. 36

Comparison of Deep Learning Architectures for three different Multispectral Imaging Flow Cytometry Datasets

Philippe Krajsic
0000-0002-1722-8752
Leipzig University
Information Systems Institute
Grimmaische Straße 12, 04109 Leipzig, Germany
krajsic@wifa.uni-leipzig.de

Thomas Hornick, Susanne Dunker
0000-0003-0280-9260, 0000-0001-7276-776X
German Centre for Integrative Biodiversity Research
Department of Physiological Diversity
Puschstraße 4, 04103 Leipzig, Germany
thomas.hornick@idiv.de, susanne.dunker@idiv.de

***Abstract*—Multispectral imaging flow cytometry (MIFC) is capable of capturing thousands of microscopic multispectral cell images per second. Deep Learning Algorithms in combination with MIFC are currently applied in different areas such as classifying blood cell morphologies, phytoplankton cells of water samples or pollen from air samples or pollinators. The goal of this work is to train classifiers for automatic and fast processing of new samples to avoid labor-intensive and error-prone manual gating and analyses and to ensure rigor of the results. In this study we compare state of the art Deep Learning architectures for the use case of multispectral image classification on datasets from three different domains to determine whether there is a suitable architecture for all applications or if a domain-specific architecture is required. Experiments have shown that there are multiple Convolutional Neural Network (CNN) architectures that show comparable results with regard to the evaluation criteria accuracy and computational effort. A single architecture that outperforms other architectures in all three domains could not be found.**

***Index terms*—Computer Vision, Classification, Deep Learning, Multispectral Imaging Flow Cytometry**

I. Introduction

MULTISPECTRAL imaging flow cytometry (MIFC) has been recently shown to be useful for environmental monitoring of plant-pollinator interactions and assessment of food and water quality [1], [2], [3]. This measuring technique, originally designed for immunological analyses as blood cell analysis, allows to separate single cells from a fluid suspension by hydro-dynamically focusing of a sample stream in a narrowing flow cell, surrounded by a sheath stream. The resulting acceleration of the cells, in conjunction with the hydrodynamic forces acting on them, separates the cells in the liquid flow. Once the cells occur as single cells in the sample stream they will be imaged by CCD (charge-coupled device) cameras. Two cameras take independent images from LED illumination of the cells (brightfield images), images from different laser excitation and respective light emission of the cells (fluorescence images) as well as images from red light illumination (scatter images). The instrument is capable of capturing 12 images per particle (two brightfield, nine fluorescence and one scatter image) with a sample throughput of up to 2000 particles/s [1], [2], [4]. The high sample throughput allows for an unprecedentedly high measuring efficiency in comparison to the traditional benchmark of manual microscopy. The images produced by the imaging flow cytometer instrument come with some specifics in comparison to other images used in common image recognition tasks. As the microscopic images are recorded in high throughput, they have a relatively low resolution (a pixel size of 0.5 × 0.5 µm) but at the same time, a high number of images per sample can be collected (500-5000 particles à 12 images/particle). In addition, MIFC already provides images where the object of interest is depicted in the image center and the images have a uniform background. We could already demonstrate the successful application of the different architectures ResNet V2 [4] and Inception V3 [2] for this kind of data to distinguish 27 classes of phytoplankton (relevant for water quality assessment) and 35 classes of pollen (relevant for plant-pollinator studies, food and air quality), respectively. [5] used a ResNet50 architecture to discriminate seven classes of different blood cell morphologies. These examples already show that for different MIFC datasets, different architectures have been applied. But no systematic evaluation of different architectures has taken place so far. The different mentioned application examples have different numbers of classes which need to be differentiated. For immunological applications, a limited set of cell types needs to be distinguished, while in environmental monitoring, potentially several thousands of classes need to be differentiated eventually [3].

Most models used to classify MIFC datasets are based on architectures that were assessed on large image datasets such as ImageNet [1], [2]. The images of the pre-trained dataset differ from MIFC images in the way that they depict

We thank the German Research Foundation for support via the iDiv Flexpool Funding projects Grant Number: Grant Number: 34600830-13, (‘PolDiv’-Project) and Grant Number: RA-373/20 (iCyt - Support Unit). Furthermore, we thank Prof. Dr. Patrick Maeder and David Boho from University Ilmenau for their valuable thoughts during the preparation of the study and all student helpers and interns for help during image annotation.

RGB images, have more labels and higher resolution. To acknowledge the specificity of the images and the multiple channels available in MIFC datasets, we want to enhance the existing studies and aim to answer the following research questions:

- What is the best architecture for different kinds of datasets with respect to accuracy?
- What is the most suitable degree of CNN architecture complexity for the individual datasets?
- From the best performing architectures which is the most sustainable one with respect to computational effort and resource consumption?
- Is there a general architecture for all different MIFC datasets with best performance on accuracy and computational effort?

The remainder of the paper is structured as follows: Section II presents the applied methods to answer the research questions. Section III presents the obtained classification results. A discussion of the results obtained with reference to the initial research questions, takes place in section IV. Possible limitations of the work are addressed in Section V. The study concludes with a summary of all findings and an outlook on future work in section VI.

II. Methods

The methodology used, from the selection of CNN architectures, data acquisition, data preparation, MIFC measurements to training and validation strategy used, is described below.

Selection of CNN architectures

For the comparison of different CNN architectures we considered seven different architectures: DenseNet [6], Inception V3 [7], Inception-ResNet V2 [8], MobileNetV2 [9], ResNet V2 [10], VGG [11] and Xception [12]. DenseNet (121, 169, 201), ResNet (50, 101, 152) and VGG (16, 19) have different depth configurations that were also considered. These architectures represent frequently used architectures in science and practice and thus were selected for this architecture comparison. Based on this architecture selection, it is also possible to assess whether more complex architectures are more suitable than shallower architectures for classification, and what correlation there is between the complexity of an architecture and resource consumption during model training. Table I provides an overview of the selected architectures with the specific network depths and input size.

TABLE I: Overview of CNN architectures

Model	Depth	Input Size
DenseNet-121	121	224x224x3
DenseNet-169	169	224x224x3
DenseNet-201	201	224x224x3
Inception V3	48	229x229x3
Inception-ResNet V2	164	229x229x3
MobileNet V2	53	224x224x3
ResNet-50	50	224x224x3
ResNet-101	101	224x224x3
ResNet-152	152	224x224x3
VGG-16	16	224x224x3
VGG-19	19	224x224x3
Xception	71	229x229x3

Dataset acquisition and annotation

The availability of high-quality datasets is crucial for the performance of a supervised learning approach. To ensure the cross-domain applicability of our study, we used three datasets from different fields of application. The stored red blood cells (RBC) dataset consists of 63,000 samples in seven clinically relevant blood cell morphologies associated with storage lesions [5].

The phytoplankton dataset consists of six naturally co-occurring species, three cyanobacteria (*Chroococcus minutus* - species "C" SAG 41.79, *Microcystis aeruginosa* - species "M" SAG 1450-1, *Synechocystis sp.* - species "S" PCC 6803, and three chlorophytes *Acutodesmus obliquus* - species "A" SAG 276-3a, *Desmodesmus armatus* - species "D" SAG 276-4d, and *Oocystis marssonii* - species "O" SAG 257-1), grown exponentially under controlled laboratory conditions (60 µmol photons/ m-2/s-1; 14/10 h light/dark cycle; WC-Medium) [13]. Measurements were performed in subsequent biologically independent mono-culture experiments (rep 0, rep 1). The total dataset contains 12,000 samples of species A,C,D,M,O and S.

The pollen dataset samples were collected in the botanical garden of Leipzig, Germany and natural fields in the surrounding area of Leipzig during peak flowering time from 2018 to 2020. It consists of 4,800 samples, which are randomly stratified, i.e. each class has the same number of samples. There are twelve species in seven genera, which are mostly wind-distributed and comprise morphological similar allergenic species. Similarity in pollen features within a genus restricts classical light microscopic discrimination of the selected species to the genus level. For that reason we tested classification on different taxonomic levels (genus and species) to see if CNN were even capable of classifying on species level. To get a representative dataset, we used 20% high-quality images (pollen in focus, non-cropped, without other debris particles on image) and 80% low-quality images (pollen either out of focus, partially cropped or pollen images with additional debris particles on image).

MIFC measurements

All samples were measured with an Amnis® Image Stream®X MK II imaging flow cytometer (Amnis part of Cytek, Amsterdam, Netherlands). For the phytoplankton dataset, measurements were performed according to [13], for the blood cell dataset according to [5] and for the pollen dataset according to [2] as shown in Table II.

Data preparation and augmentation

All images were channel-wise standardized (i. e. rescaled to have a mean of 0 and unit variance) and normalized (i. e. rescaled to a value range of 0 and 1) utilizing the pixel values from the respective training dataset.

Most CNN architectures are translation invariant but not invariant with regard to scale, rotation and different perturbations. To address the problem of overfitting we artificially

TABLE II: Datasets

Characteristics	Phytoplankton	Pollen	Blood cell
Classes	6	12	7
No. particles	12000	4800	63000
Channels	12 (1-12)	7 (1-6,9)	3 (1,9,12)
Lasers	3 (488/561/785)	3 (488/561/785)	5
Magnification	40x	40x	60x
Sheath fluid	D-PBS	D-PBS	PBS
References	[13]	unpublished	[5]

increased the dataset size to alleviate scarcity issues. Several random data augmentations that yield credible images are introduced to the training datasets to help the models to generalize better and to be more robust with regard to random perturbations and noise [14].

Brightness and contrast are randomly adjusted channel-wise in [-0.3, 0.3] and [0.5, 2.0] intervals to achieve a robustness of the classifier for different MIFC calibrations, fluorescence and random background noise. Further random geometric transformations as rotation, horizontal flipping and central cropping (interval [0.8, 1.0]) are introduced to make the classifier robust against different cell orientations and cell sizes that may occur across different measurements. As all images have varying aspect ratios they were resized to 116 by 116 pixels, padded with zeros.

Training and validation strategy

For each dataset a k-fold stratified cross-validation [15] was performed to find an optimal hyperparameter combination that is less biased or optimistic compared to a simple train/validation/test split. For that purpose the datasets were split into k=5 equally-sized subsets that have the same class distribution as the original datasets. For each subset the set is used as a test dataset on which the model performance is evaluated and the remaining sets are used to train the model. Over all runs the performance metrics accuracy, macro averaged precision, recall and F_1 score were averaged to get an estimate how good the final model performs and how robust it is with regard to data variability.

All models were trained on all available multispectral channels (1,9,12 for RBC, 1-12 for phytoplankton and 1-6 and 9 for wind pollen).

A grid search was used for hyperparameter optimization. Considered hyperparameters were optimizer function (RMSProp [16], Adam [17]), batch size (16, 32, 64) and learning rate (1e-4, 1e-5, 1e-6). For all architectures categorical cross entropy was chosen as a loss function. In total, 18 hyperparameter combinations were evaluated per architecture.

The learning rate was reduced by a factor of 10 if the validation loss had not decreased within 20 epochs and the training was stopped when the validation loss had not decreased within 30 consecutive epochs.

The best hyperparameter combination for each architecture was assessed on a fixed holdout dataset eventually. For each model the number of trainable parameters (weights and biases) was calculated as a measure of model complexity. As the number of available image channels varies between the datasets, the number of parameters for the same architecture differs. Additionally we measured the floating point operations (FLOPs) for a single forward pass to quantify inference performance.

III. Results

We wanted to find a CNN architecture that shows the best metrics for MIFC datasets from different domains and determine whether complex architectures outperform simpler ones.

Three datasets recorded with an imaging flow cytometer containing samples from three different application domains, i.e., wind pollen, phytoplankton and blood cells, were used to evaluate the different CNN architectures captured in Table I. These CNN architectures were trained on the different datasets. The task of each model was to recognize patterns and structures in the images in order to achieve the highest possible accuracy in assigning the images to their respective classes.

Classification Results

Table III and Table IV show the results for each architecture on the species or on the genus level. The balanced wind pollen dataset shows an increase of performance metrics from the training on the species level to the training on the same dataset on genus level. This is not surprising, as the number of classes are reduced from twelve to seven and the number of samples per class is increased which gives the classifier during training more exposure to training samples and thus the potential to generalize better. We observed that the Inception-ResNet is the best performing architecture (96.88% accuracy on species level, 98.96% on genus level) in both tests, closely followed by Inception, DenseNet and Xception. An impact of the size and complexity of the model on the classification accuracy cannot be determined on the basis of the results obtained.

The classification results for the phytoplankton dataset are shown in Table V (train on rep-0, test in rep-1) and Table VI (train on rep-1, test in rep-0). The phytoplankton dataset was trained on one independent replicate of the measurement, evaluated on another independent replicate of the measurement and vice versa. In both tests the VGG-16 and VGG-19 architecture showed the best classification accuracy of approximately 92%, closely followed by Inception and DenseNet architectures with comparable accuracies. Again the size of the models regarding number of parameters seems not to have a real impact on the accuracy of the model.

Table VII (train on Canadian, test on Swiss) and Table VIII (train on Swiss, test on Canadian) show the results for the blood quality dataset. We observed that there is no single prevailing architecture for the blood cell dataset that was trained on samples originating from Switzerland, evaluated on samples originating from Canada and vice versa. DenseNet-121 and Xception are both the best performing architectures. Here we observe F_1 scores of 75.68% (train on Canadian, test on Swiss) and 87.90% (train on Swiss, test on Canadian).

TABLE III: Wind pollen (species level)

Architecture	Top-1 Acc.	Precision	Recall	F_1 score
Inception-ResNet	96.88%	96.89%	96.88%	96.87%
Inception	96.25%	96.29%	96.56%	96.24%
DenseNet-121	96.25%	96.33%	96.46%	96.24%
Xception	96.04%	96.09%	96.35%	96.04%
DenseNet-169	96.04%	96.19%	96.29%	96.01%
ResNet-152	95.83%	95.95%	96.22%	95.81%
ResNet-50	95.63%	95.72%	96.13%	95.60%
DenseNet-201	95.42%	95.47%	96.04%	95.43%
ResNet-101	95.42%	95.80%	95.97%	95.43%
VGG16	94.38%	94.57%	95.66%	94.39%
VGG19	94.38%	94.42%	95.56%	94.37%
MobileNet	89.38%	89.52%	94.84%	89.34%

TABLE IV: Wind pollen (genus level)

Architecture	Top-1 Acc.	Precision	Recall	F_1 score
Inception-ResNet	98.96%	98.95%	98.81%	98.87%
Inception	98.75%	98.82%	98.69%	98.69%
DenseNet-201	98.75%	98.42%	98.59%	98.64%
DenseNet-169	98.75%	98.88%	98.66%	98.62%
Xception	98.54%	98.77%	98.58%	98.51%
VGG16	98.54%	98.71%	98.53%	98.48%
DenseNet-121	98.33%	98.28%	98.44%	98.08%
ResNet-152	98.33%	98.05%	98.40%	98.07%
ResNet-101	97.71%	97.54%	98.34%	97.69%
VGG19	97.71%	97.36%	98.30%	97.63%
ResNet-50	97.50%	97.40%	98.06%	97.23%
MobileNet	96.46%	96.32%	97.67%	96.14%

TABLE V: Phytoplankton (train: rep-0, test: rep-1)

Architecture	Top-1 Acc.	Precision	Recall	F_1 score
VGG16	92.62%	92.72%	92.62%	92.54%
VGG19	92.50%	92.54%	92.56%	92.46%
Inception	92.27%	92.23%	92.46%	92.20%
Inception-ResNet	92.03%	92.03%	92.35%	91.97%
DenseNet-121	91.52%	91.49%	92.19%	91.45%
DenseNet-169	91.47%	91.45%	92.07%	91.41%
ResNet-101	91.22%	91.19%	91.95%	91.16%
DenseNet-201	91.20%	91.15%	91.85%	91.15%
Xception	91.18%	91.14%	91.78%	91.12%
ResNet-152	90.53%	90.47%	91.65%	90.45%
ResNet-50	90.37%	90.27%	91.45%	90.26%
MobileNet	87.32%	87.12%	90.69%	87.16%

TABLE VI: Phytoplankton (train: rep-1, test: rep-0)

Architecture	Top-1 Acc.	Precision	Recall	F_1 score
VGG16	91.88%	92.01%	91.88%	91.83%
VGG19	91.87%	92.10%	91.87%	91.82%
DenseNet-121	91.87%	91.97%	91.87%	91.82%
DenseNet-201	91.38%	91.64%	91.75%	91.35%
Inception	91.30%	91.52%	91.66%	91.27%
ResNet-152	91.28%	91.43%	91.60%	91.23%
DenseNet-169	91.17%	91.32%	91.54%	91.11%
ResNet-101	90.92%	91.03%	91.46%	90.87%
Xception	90.42%	90.61%	91.34%	90.38%
ResNet-50	90.28%	90.45%	91.24%	90.20%
Inception-ResNet	90.10%	90.46%	91.13%	90.12%
MobileNet	87.45%	87.86%	90.39%	87.27%

TABLE VII: Blood quality (train: Canadian, test: Swiss

Architecture	Top-1 Acc.	Precision	Recall	F_1 score
DenseNet-121	87.09%	75.07%	81.47%	75.68%
DenseNet-201	86.75%	74.63%	80.53%	75.25%
Inception-ResNet	86.48%	76.29%	79.58%	76.42%
DenseNet-169	86.40%	74.62%	81.02%	75.28%
Inception	86.27%	75.54%	80.14%	76.04%
ResNet-152	85.96%	74.98%	80.58%	75.14%
VGG16	85.93%	73.76%	79.23%	74.23%
ResNet-50	85.69%	72.34%	80.29%	73.47%
ResNet-101	85.35%	72.65%	79.97%	73.11%
VGG19	85.33%	73.85%	81.06%	75.09%
Xception	84.85%	76.07%	79.73%	75.65%
MobileNet	83.89%	71.99%	78.56%	72.91%

TABLE VIII: Blood quality (train: Swiss, test: Canadian)

Architecture	Top-1 Acc.	Precision	Recall	F_1 score
Xception	87.91%	88.78%	87.71%	87.90%
Inception	87.90%	88.68%	87.67%	87.83%
ResNet-152	87.84%	88.79%	87.91%	87.95%
Inception-ResNet	87.79%	88.54%	87.84%	87.79%
DenseNet-121	87.71%	88.67%	87.82%	87.86%
VGG19	87.68%	88.48%	87.99%	87.75%
DenseNet-201	87.55%	88.68%	87.62%	87.64%
ResNet-50	87.26%	88.48%	87.31%	87.41%
VGG16	87.12%	88.22%	87.35%	87.29%
DenseNet-169	87.04%	88.25%	87.25%	87.23%
ResNet-101	86.85%	88.13%	86.91%	86.97%
MobileNet	86.63%	87.86%	86.41%	86.60%

There is a noticeable difference in the F1 scores obtained between train on Canadian and test on Swiss and vice versa.

The ability of the models to distinguish between classes of each dataset is illustrated in Fig. 1 using the associated confusion matrix. In (a), the accuracies of the mappings for wind pollen at species level are given. The Inception-ResNet achieves optimal assignments for the vast majority of pollen classes and shows only slight weaknesses in distinguishing pollen from the same genus level (e.g., *Corylus avellana* and *Corylus colurna*). These pollen species from the same genus level show many similarities in their appearance, so that misclassifications can occur here. For the genus-level class assignment in (b), it can be seen that the accuracy of the Inception-ResNet could be increased compared to the species-level assignment. Compared to (a), no discrimination on species level is necessary, allowing the model to more easily identify differences in classes at the genus level. Confusion matrix (c) shows the accuracy of the VGG16 network class assignments for phytoplankton (train on rep-0, test on rep-1), (d) for phytoplankton (train on rep-1, test on rep-0). No significant differences can be detected between the two variants in terms of accuracy in class assignment. It is noticeable that the model shows similar weaknesses in correct assignment for classes 'C' (*Chroococcus minutus*) and 'O' (*Oocystis marssonii*) for both variants. The shape and size of both species is quite similar which might explain a low discriminatory power. Since both species belong to different taxonomic groups, the discrimination could be enabled with flow cytometric taxonomic separation [18]. The accuracy of class assignment for the Blood Quality dataset is shown in confusion matrix (e) (train on Canadian, test on Swiss) and (f) (train on Swiss, test on Canadian). Again, a similar level of assignment accuracy to classes can be observed for both variants. In (e), however, the class CrenatedSpheroid with an accuracy of only 52.3% shows a conspicuously high number of false assignments, which cannot be observed to the same extent in (d). In general, the blood cell group assignment is more complicated than species identification in the two other datasets and for a distinction there is less a discrete difference but rather a continuously diverging morphology.

FLOPs - Floating Point Operations

In addition to identifying a suitable architecture for the highest possible classification accuracy, the computational effort and associated resource consumption of the architectures was considered in terms of FLOPs, number of floating point operations. Therefore, top-1 accuracy was set in relation to the number of FLOPs (in billions) and the number of model parameters (in millions). Fig. 2 (a) and (b) illustrates this comparison across the model architectures for the wind pollen dataset. With regard to the classification of wind pollen at the species level, it was observed that the Top-1 Accuracy decreased with increasing number of FLOPs. For classification on genus level this tendency is not observed. On both levels the best Accuracy-Flops ratio, and thus most resource efficient architecture, is provided by the Inception-ResNet architecture. Fig. 2 (c) and (d) illustrates this comparison for the phytoplankton dataset. For (c) Phytoplankton (train on rep-0, test on rep-1) the best Accuracy/FLOP ratio is achieved by the Inception network. Here, a relativly high accuracy can be achieved on a small number of FLOPs. That means that less operations are required to run a single instance of the Inception model compared to VGG16 or VGG19 to achieve similar accuracy. The same can be stated for (d) Phytoplankton (train on rep-1, test on rep-0). Here, the DenseNet-121 architecture achieves the best Accuracy/FLOP ratio and allows a resource-efficient use of the model. For the blood quality dataset. shown in Fig. 2 (e) and (f), DenseNet-169 achieves the best ratio for (e) RBC (train on Canadian, test on Swiss) and the Inception networks was found to be the most ressource-efficient one for (f) RBC (train on Swiss, test on Canadian).

IV. Discussion

In the present work, different architectures of artificial neural networks are analyzed with respect to their performance for the classification of different datasets (wind pollen, phytoplankton, blood cells) generated by MIFC. With reference to the research questions raised, the following findings could be obtained based on this study:

1) What is the best architecture for different kinds of datasets with respect to accuracy? Our findings differ from applied techniques used in previous literature, where different network architectures were used to classify pollen, phytoplankton and blood cells. In [4], the authors achieved an average accuracy of 99% for combined images of phytoplankton at species level using a ResNet v2 with 50 convolutional layers, which was not surpassed by the VGG16 as the best performing model used in our analysis. In this context, it should be noted that the models used here are not precisely tuned, since the focus of this work is on the comparison of different architectures than on the optimization of a single architecture.

For the classification of wind pollen, the authors in [2] achieve an accuracy of max. 96% with an Inception V3 network with 48 convolutional layer. These results can be confirmed within the scope of this study, so that similar results could be achieved with the compared architectures (e.g., Inception-ResNet (164 layers) = 96.88%, Inveption V3 (48 layers) = 96.25%). This suggests, that deeper networks are not necessarily superior to shallower networks for pollen classification.

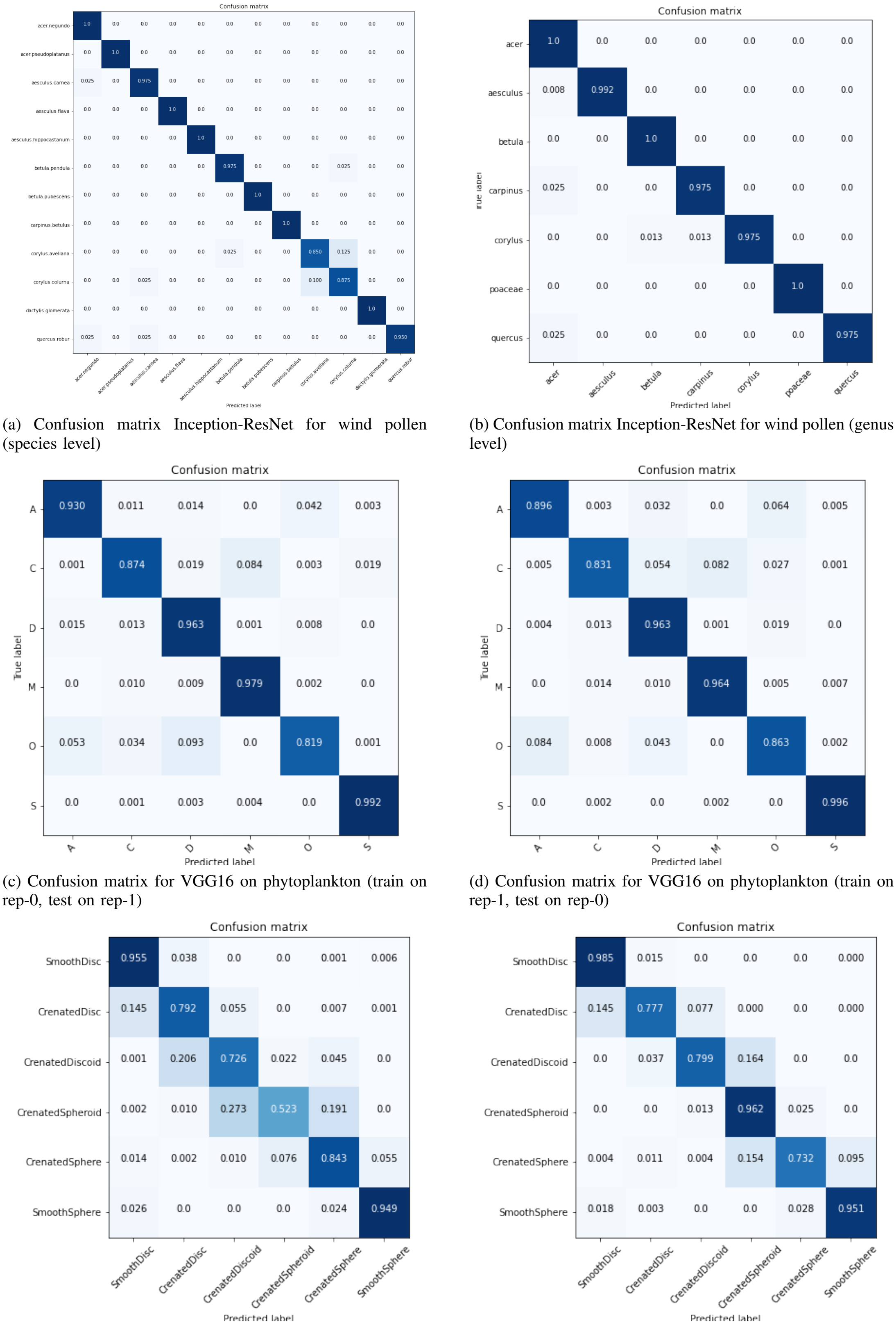
With respect to the classification of the blood cell data set, it can be highlighted that both similarities and differences can be identified in recent studies. The authors in [5] used a ResNet-50 for a classification of red blood cells. The results show that a comparably good classification accuracy could be achieved and that similar difficulties exist in the correct assignment of certain classes (e.g. crenated spheroid). The ResNet-50 in the comparative study achieves an average accuracy of 80%. The best performing architectures in this study are DenseNet-121 (87.09% for train on Canadian, test on Swiss) and XCeption (87.91% for train on Swiss, test on Canadian).

2) What is the most suitable degree of CNN architecture complexity for the individual datasets? We could show that deeper neural networks do not necessarily perform better than shallow networks. Instead, an accurate classification may be achieved with comparably shallow networks, such as VGG-16, VGG-19 or Inception (48 layers). This fact leads to the conclusion that the use of such, shallower networks would be advantageous, especially in the case of limited hardware resources.

3) From the best performing architectures which is the most sustainable one with respect to computational effort and resource consumption? The advantage of using CNN with fewer layers is that they have lower hardware requirements and shorter training times compared to their deeper counterparts. Shorter training times allow testing of more hyperparameters and simplify the overall training process. This is particularly useful in environments with limited resources or where a resource-efficient use deep learning techniques is aspired.

Additionally, shorter training times can facilitate the integration of improvement methods into the training data, such as the implementation of "human in the loop" annotations. Human in the loop means that the training of a network is monitored by a human expert who can intervene at critical steps and correct the network. For example, the expert can check misclassifications, effectively reducing annotation noise. With shorter training times, such feedback loops can be executed more quickly.

4) Is there a general architecture for all different MIFC datasets with best performance on accuracy and computational effort? Overall, it can be highlighted that no single best architecture could be identified for the respective datasets, as they are often very close in terms of accuracy (deviations in many cases under 1%). It can be emphasized that there is no best-performing architecture from a generally valid point of view with regard to the accuracy-resource ratio.



(a) Confusion matrix Inception-ResNet for wind pollen (species level)

(b) Confusion matrix Inception-ResNet for wind pollen (genus level)

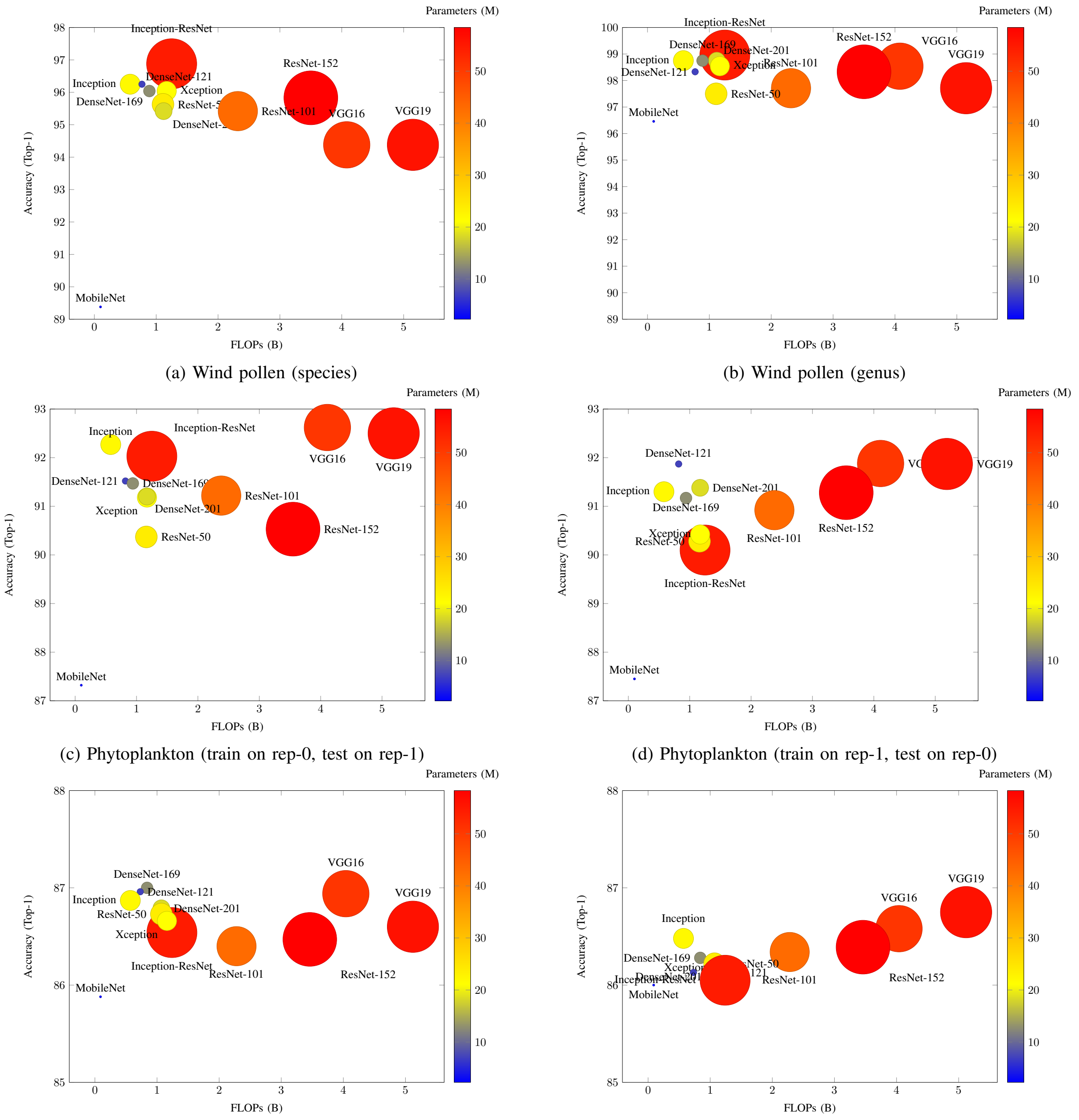
(c) Confusion matrix for VGG16 on phytoplankton (train on rep-0, test on rep-1)

(d) Confusion matrix for VGG16 on phytoplankton (train on rep-1, test on rep-0)

(e) Confusion matrix for DenseNet-169 on blood quality (train on Canadian, test on Swiss)

(f) Confusion matrix for VGG19 on blood quality (train on Swiss, test on Canadian)

Fig. 1: Confusion matrices of the best performing models on the respective datasets



(a) Wind pollen (species)

(b) Wind pollen (genus)

(c) Phytoplankton (train on rep-0, test on rep-1)

(d) Phytoplankton (train on rep-1, test on rep-0)

(e) RBC (train on Canadian, test on Swiss)

(f) RBC (train on Swiss, test on Canadian)

Fig. 2: FLOPs (Billions) compared to Accuracy (Top-1)

Depending on the data set, different architectures achieve an optimal accuracy-resource ratio.

V. Limitations

A limitation of the present work is, that different, mostly larger architectures, may tend to overfit on less complex datasets, like the blood cell dataset, that is less complex than the wind pollen dataset. As a consequence, an overfitting with a lower generalizability of some results cannot be excluded and should be considered when interpreting our results. Furthermore, the CNN architectures used in this study not have fully optimized hyperparameters, which could impact the performance of the models. Those architectures may not performed as well as they could have following hyperparameter optimization.

VI. Conclusions

The aim of the present work was to compare different CNN architectures for the classification of MIFC datasets. In this context, seven different architectures with different complexity and depth were trained and tested on three datasets (wind pollen, phytoplankton, red blood cells). The evaluation results demonstrate, that complex architectures with a large number of trainable parameters or increasing depth are not always required to achieve state of the art results. A DenseNet architecture with 121 layers reaches comparable results to more complex architectures such as Inception-ResNet and VGG that consume significantly more computing resources during training. The ecological footprint during model training and inference can be reduced by using simpler architectures without sacrificing accuracy.

A CNN architecture that is most qualified for all datasets under consideration could not be determined, as most considered architectures show comparable results with regard to the evaluation criteria.

Future research could include training and testing on even larger datasets with more classes and higher variability. In addition, hyperparameter optimizations can be performed on individual architectures to identify a universally best architecture for the classification of the investigated datasets. Furthermore, it has to be evaluated whether the use of a single architecture is reasonable at all and can be complemented by the implementation of ensemble methods.

Acknowledgment

Computations for this work were done in part using resources of the Leipzig University Computing Center.

Availability of data and materials

Data and materials used in the study are available upon reasonable request.

References

[1] S. Dunker, "Hidden Secrets Behind Dots: Improved Phytoplankton Taxonomic Resolution Using High-Throughput Imaging Flow Cytometry," *Cytometry Part A*, vol. 95, no. 8, pp. 854–868, 2019, doi: 10.1002/cyto.a.23870.

[2] S. Dunker, E. Motivans, D. Rakosy, D. Boho, P. Mäder, T. Hornick, and T. M. Knight, "Pollen analysis using multispectral imaging flow cytometry and deep learning," *New Phytologist*, vol. 229, no. 1, pp. 593–606, 2021. doi: 10.1111/nph.16882

[3] S. Dunker, M. Boyd, W. Durka, S. Erler, W. S. Harpole, S. Henning, U. Herzschuh, T. Hornick, T. Knight, S. Lips *et al.*, "The potential of multispectral imaging flow cytometry for environmental monitoring," *Cytometry Part A*, vol. 101, no. 9, pp. 782–799, 2022. doi: 10.1002/cyto.a.24658

[4] S. Dunker, D. Boho, J. Wäldchen, and P. Mäder, "Combining high-throughput imaging flow cytometry and deep learning for efficient species and life-cycle stage identification of phytoplankton," *BMC Ecology*, vol. 18, no. 1, pp. 1–15, 2018. doi: 10.1186/s12898-018-0209-5

[5] M. Doan, J. A. Sebastian, J. C. Caicedo, S. Siegert, A. Roch, T. R. Turner, O. Mykhailova, R. N. Pinto, C. McQuin, A. Goodman, M. J. Parsons, O. Wolkenhauer, H. Hennig, S. Singh, A. Wilson, J. P. Acker, P. Rees, M. C. Kolios, A. E. Carpenter, and D. Geman, "Objective assessment of stored blood quality by deep learning," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 117, no. 35, pp. 21 381–21 390, sep 2020. doi: 10.1073/pnas.2001227117

[6] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, vol. 2017-January, pp. 2261–2269, 2017. doi: 10.1109/CVPR.2017.243

[7] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-Decem, pp. 2818–2826, 2016. doi: 10.1109/CVPR.2016.308

[8] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-ResNet and the impact of residual connections on learning," in *31st AAAI Conference on Artificial Intelligence, AAAI 2017*. AAAI press, feb 2017. doi: 10.48550/arXiv.1602.07261 pp. 4278–4284.

[9] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 4510–4520, 2018. doi: 10.1109/CVPR.2018.00474

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Identity Mappings in Deep Residual Networks," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9908 LNCS, pp. 630–645, mar 2016. doi: 10.1007/978-3-319-46493-038

[11] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," sep 2015. doi: 10.48550/arXiv.1409.1556

[12] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, vol. 2017-January. Institute of Electrical and Electronics Engineers Inc., nov 2017. doi: 10.1109/CVPR.2017.195. ISBN 9781538604571 pp. 1800–1807.

[13] P. Hofmann, A. Chatzinotas, W. S. Harpole, and S. Dunker, "Temperature and stoichiometric dependence of phytoplankton traits," *Ecology*, vol. 100, no. 12, p. e02875, dec 2019. doi: 10.1002/ecy.2875

[14] C. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," *Journal of Big Data*, vol. 6, no. 1, 2019. doi: 10.1186/s40537-019-0197-0

[15] X. Zeng, T. R. Martinez, X. Inchuan Zeng, and T. N. R M A Rtinez, "Distribution-balanced stratified cross-validation for accuracy estimation," *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 12, no. 1, pp. 1–12, 2000. doi: 10.1080/095281300146272

[16] S. Ruder, "An overview of gradient descent optimization algorithms," 2017. doi: 10.48550/arXiv.1609.04747

[17] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017. doi: 10.48550/arXiv.1412.6980

[18] S. Dunker, "Hidden secrets behind dots: Improved phytoplankton taxonomic resolution using high-throughput imaging flow cytometry," *Cytometry Part A*, vol. 95, no. 8, pp. 854–868, 2019. doi: 10.1002/cyto.a.23870